ukais

UK ACADEMY FOR INFORMATION SYSTEMS

Proceedings of the 30th UK Academy for Information Systems (UKAIS) International Conference

Newcastle University

23rd-24th April 2025

ISBN: 978-1-7390875-2-4

Preface

Proceedings of the 30th UK Academy for Information Systems (UKAIS) International Conference: 2025

ISBN: 978-1-7390875-2-4

On behalf of the UKAIS and its committee, we welcome you to the conference proceedings for UKAIS 2025. This volume contains the papers presented at UKAIS2025: UK Academy for Information Systems, Annual International Conference held on April 23-24, 2025. The conference has 13 tracks and includes 3 Keynote presentations − Professor Michelle Carter is Professor of Information Systems and Director of Equality, Diversity, and Inclusion at the Alliance Manchester Business School, University of Manchester; Professor Edgar Whitley is Professor of Information Systems at The London School of Economics and Political Science (LSE); Klaus-Michael Vogelberg is the Chief Architect and Technology Advisor with SAGE Plc.

The UKAIS conference is the premier academic event in the Information Systems calendar within the UK and attracts leading scholars from the UK and overseas. It is a charity, whose aims are to enhance the recognition and knowledge of IS within the UK, and to provide a forum for discussing issues in IS teaching and research. UKAIS recognises the importance of including practitioners in its work.

The UK Academy for Information Systems was established in 1994 to foster a better understanding of the Information Systems field within the UK. We provide a forum for discussing issues in IS teaching and research and lobby professional/policy bodies on behalf of our field, such as the Office for Students, UKRI/Research England, UK business and UK Government. There is a conference every year, which is usually preceded by a PhD consortium. UKAIS Aims:

- To promote a better knowledge and understanding of information systems within the United Kingdom.
- To improve the practice of information systems teaching and research.
- To enable successful knowledge transfer of IS research into teaching and practice in order to provide a positive economic and societal impact.

Many thanks to all those that have given of their time so freely to review papers for the academy, it is much appreciated.

Thanks to everyone that has made this happen, the UKAIS Board and all the Track Chairs and Reviewers, we thank you all

April, 2025
Newcastle

Laurence Brooks
Honglei Li
Savvas Papagiannidis
The UKAIS 2025 conference chairs

Organising Team

Conference Chairs:

Laurence Brooks (University of Sheffield), Honglei Li (Northumbria University), Savvas Papagiannidis (Newcastle University)

Programme Committee

Roba Abbas	IEEE SSIT Technical Activities
Tahir Abbas Syed	University of Manchester
Fulya Acikgoz	University of Sussex
Zeeshan Ahmed Bhatti	University of Portsmouth
Greg Austin	Social Cyber Institute (SCI)
Laurence Brooks	University of Sheffield
Abdelsalam Busalim	Technological University Dublin
Dinara Davlembayeva	Newcastle University Business School
Mohammad Delgosha	University of Birmingham
Emma Forsgren	University of Leeds
Emma Gritt	University of Leeds
Najmeh Hafezieh	Royal Holloway, University of London
Nastaran Hajiheydari	Queen Mary, University of London
Mahsa Honary	Lancaster University Management School
Mengyun Hu	Newcastle University Business School
Oliver George Kayas	Liverpool John Moores University
Maria Kutar	University of Salford
Honglei Li	Northumbria University
Yuzhu Li	University of Massachusetts Dartmouth
Haiyan Lu	Newcastle University Business School
Davit Marikyan	University of Bristol
Katina Michael	Arizona State University (ASU)
Patrick Mikalef	Norwegian University of Science and Technology
Jian Mou	Pusan National University
Hajar Mozaffar	University of Edinburgh Business School
Matti Mäntymäki	University of Turku
Yu-Chun Pan	Northeastern University London
Savvas Papagiannidis	Newcastle University Business School
Ilias Pappas	Norwegian University of Science and Technology
Gamila Shoib	University of Bath
Chekfoung Tan	University College London
Eleni Tzouramani	University of the West of Scotland
Polyxeni Vasilakopoulou	University of Agder
Arturo Vega Pinedo	Newcastle University Business School
Glenn Withers	Australian National University (ANU)
Hina Yasin	University of Portsmouth
Efpraxia Zamani	Durham Business School
Guoqing Zhao	School of Management, Swansea University

Table of Contents

"AI lacks the ability to inspire students like real teachers can." - Exploring UK University Students Views on Lecturer Generative AI Usage

Dr David Grundy

Newcastle University Business School, Newcastle University

Abstract

This study examines UK university students' perceptions regarding lecturers' integration of generative artificial intelligence (Gen-AI) in teaching. By analysing qualitative survey responses from 205 business students, the research uncovers diverse opinions about Gen-AI's role in higher education. Five key themes emerged: scepticism and concern about Gen-AI; value of human interaction and the personal touch; potential benefits with caution; negative impact on teaching quality and creativity; uncertainty. Students expressed apprehension that Gen-AI could erode educational authenticity and undermine essential human engagement, such as personalised mentorship, tailored feedback, and inspirational instruction. While acknowledging the efficiency and support Gen-AI may offer, respondents advocated for a balanced approach where technology complements rather than replaces human educators. The paper calls for cautious adoption of Gen-AI, considering its limitations and negative implications for creativity and student engagement. It highlights the need for dialogue and research, including longitudinal, comparative, and experimental studies, to track evolving student perspectives.

Keywords: Generative Artificial Intelligence (Gen-AI), Student Perceptions, Higher Education, Human Interaction, Teaching Quality

1. Introduction & Literature Review

As an educator looking to continuously improve the learning experiences of my students, I[1] have focused much of my time lately on the possible applications of Artificial Intelligence in an educational framework. While generative AI's (Gen-AI) certainly a change, with revolutionary potential to challenge pedagogic traditionalism with its emergence, the reception amongst university students is under-explored terrain. In this paper, I will examine UK university student perceptions of Gen-AI as a pedagogical tool—a piece of work which has felt timely and relevant against the backdrop of an academic environment moving at breathtaking pace since the emergence of Chat-GPT into the public consciousness in November 2022. As while many papers have been published recently on the how's of Gen-AI

[1] Just to note that, in this journal article, as a piece of teaching action research intended to inform and change the practice of the author in their use of AI in the classroom, the use of first-person singular throughout is intentional. "I" refers throughout to the corresponding author.

in a classroom, exploring the ‘could we’ side of puzzle, very few have asked the question to the ultimate recipient of learning - ‘should we?’.

The applicability of Gen-AI in education varies from providing students with individualised learning processes and experiences to automated grading systems. The adoption of the technology, however, does not come without problems, especially understanding how students view and interact with Gen-AI-driven educational tools (Sevnarayan, 2024). Since studies investigating students' opinions about Gen-AI in education by their lecturers are at a very nascent stage, the approach adopted in this research paper is exploratory. It thus seeks to fill this knowledge gap by making sense of extant literature on how students reacted toward other educational innovations, to create a foundation on which Gen-AI studies in education may be anchored (Southworth et. al, 2023; Opesemowo and Ndlovu, 2024). Historically, student reactions to any educational innovation are marked by ambiguity and moderated by a variety of factors, such as perceived usefulness, perceived ease of use (He, Chen, and Kitkuakul, 2018), and effects on learning outcomes. Research into educational innovations within the context of flipped classrooms, online learning platforms, and interactive technologies does provide however a framework under which one may situate probable student reactions to Gen-AI in education (Ayanwale and Molefi, 2024).

For example, a number of research studies have been focused on the flipped classroom model (Akçayır and Akçayır, 2018) in which traditional pedagogies are reversed in that instructional content is delivered through an online environment and classroom time is used for engaging activities. Students who enjoy this type of model respond with praise, citing flexibility and increased engagement, while others have a very hard time with the self-directed learning portion of the instructional model (Bishop & Verleger, 2013; Gilboy, Heinerichs, & Pazzaglia, 2015). According to the results of qualitative surveys regarding flipped classrooms, explained by Abeysekera and Dawson (2015), the acceptance relates to the possibility of the students to adapt to new styles of learning and manage their time properly. Students also reacted differently to the dramatic move to online learning platforms, even more in the case of the COVID-19 pandemic (Yilmaz and Şahin, 2021). While most of the discussion centres on the ease and access associated with online learning, concerns about the loss of face-to-face interaction and the/its difficult motivation have been often reported. Preliminary qualitative studies that investigated these dimensions have indicated that students' prior experience with online tools, and their self-regulation skills, turned out to be strong predictors of the perceptions and levels of engagement they had in the end (Dericks et al., 2022). These include newer interactive technologies that have been developed to increase learner engagement (Kay & LeSage, 2009; Caldwell, 2007), such as clickers and educational apps. Students usually respond well to the technologies if they feel a clear association between their use and better learning and higher levels of engagement; however, this novelty typically wears off, and positive perceptions are only maintained through consistent integration into the learning process (Draper & Brown, 2004).

Early identification of student perceptions and concerns in a teaching innovation is key for lecturers to ensure they can evaluate teaching effectiveness and impact of what they are doing. Studies on several topics, including flipped classrooms, online learning platforms, and interactive technologies, identify that through the introduction phase of the new teaching innovation key dimensions within educational innovations can be noted and used to recalibrate student perceptions. In this phase, a teacher gets preliminary ideas about students' reactions to innovations, challenges, and perceived benefits accruable from them, which is important in understanding how these innovations intersect with students' learning experiences. For instance, Abeysekera and Dawson's (2015) study on the flipped classroom isolated flexibility and increased engagement as key dimensions that impact student acceptance. Means et al. (2014) did so on the bases of dimensions such as accessibility and motivation to students' responses towards the setting up of online learning platforms. It is by such understanding of the dimensions that educators and researchers can design future investigations or addressing issues via specific teaching interventions, making alterations which are most impactful to students. Furthermore, the exploration of students views at this early stage enables the uncovering of underlying attitudes and concerns towards new teaching innovations that students might have. More simply put, qualitative responses from open-ended questions can often reveal nuanced perspectives that bring a deeper understanding of the student experience, giving us data with which to improve the development of educational strategies and technologies – ultimately leading to greater effectiveness and impact. For example, according to Kay and LeSage (2009), interactive technologies are positively associated with improved learning outcomes and engagement in students, and this is innately linked to a positive perception by the students themselves that this is aiding them in their studies. Thus, student perception of that a new teaching innovation or approach is helping them learn is firmly linked to evidence of effectiveness of that innovation. Student acceptance of teaching innovations is therefore key. By understanding these student perceptions (and the level of acceptance), lecturers researching their practise can iteratively build on firm ground regarding new teaching innovations, this iterative process helps ensure that new teaching approaches are anchored in real-world experiences and perceptions from students and has a greater likelihood of leading to being more effective and accepted educational innovations.

Drawing from these studies, perceived usefulness to them personally, effects on their personal learning outcomes, and adaptability to new learning Gen-AI methods were pre-survey the expected likely major factors that may affect students' decisions to either accept or reject a lecturer's use of Gen-AI in their education. In consideration of this being an exploratory study, qualitative approaches were seen to help to shed more detailed light on students' perceptions about Gen-AI. In other words, understanding students' reactions to educational innovations may help me/us begin to frame some preliminary notions of how Gen-AI might be received within university settings. Although the exact views of students about usage of Gen-AI in education by their lecturers are largely unresearched, insights from existing studies into other educational innovations provide a useful foundation[2]. This paper contributes to this under-researched area by investigating student perceptions of Gen-AI by their lecturers and thus providing information toward future practice in Gen-AI-driven education.

2. Data Collection and Analysis

In this study, I employed a qualitative design to investigate the learning experiences and outcomes of students across several university Business programs. Data collection in this research came from a survey of perceptions and experiences of students regarding potential Gen-AI uses in their modules. I opted for the survey method for the fact that it is ideal for eliciting wide responses from a large student population within a relatively short period of time. In sum, my study involved 205 respondents drawn from mixed cohorts of students studying various programs and, therefore, representing the interdisciplinary nature of the modules being studied. More specifically, the students represented programs in business, accounting, mathematics, and combined business honours degree programs. The data collection was in April 2024 using Microsoft Forms as the online survey tool since it facilitates fast collection and management of data. There were eight preceding Likert-scale questions that asked students their views on a variety of possible Gen-AI uses in education before they came to an open text box on their overall views. The focus question under analysis in this paper and used in the survey was this open-ended to enable the capturing of

[2] From a view to this paper's contribution to the literature – this "what's in it for me?" highly personalised viewpoint of academic innovation in particular does seem to carry over to student viewpoints of the use of AI by their lecturers and does seem to underpin later analysis in this paper.

richer, qualitative insights that would be given by elaborating on experiences and giving detailed feedback. I distributed the survey to students by sharing a link in the lectures, with an appropriate invitation to participate and a link to the Microsoft Forms survey. Participation was on a voluntary basis, and no identifying information was collected ensuring confidentiality and anonymity of their responses.

For the analysis of the data the development of the themes from these open-ended responses followed a structured process informed by established methodologies in thematic analysis. Such a process can be broken down into several key steps. The first step would be reading the entire dataset several times to become deeply familiar with its content. This is an important step, as it immerses a researcher into the data, thus helping in the identification of preliminary patterns and subtleties (Braun & Clarke, 2006). The first coding process that I applied is where I systematically highlighted phrases, words, and sentences that seemed relevant to the research question. These codes are just labels that are attached to capture the essence of the data fragments. For example, sentences like *"Content created by Gen-AI lacks authenticity and depth"* and *"I am concerned that AI in teaching means a decline in the quality of education"* were coded for concerns about the quality and authenticity. After the data had been coded, I started to group these codes into possible themes. This involved looking for patterns wherever certain codes could be aggregated under themes. For instance, codes on concerns about quality, authenticity, and engagement were grouped under a theme of "Scepticism and Concern about Gen-AI in Education" (Braun & Clarke, 2006). The preliminary themes were then reviewed against the data to ensure that they accurately captured it. This entailed going back to the original data to check that the themes were representative and coherent. Some themes were merged, refined, or even discarded at this phase. For instance, themes such as "Quality Concerns" and "Engagement Issues" developed into the higher theme "Scepticism and Concern about Gen-AI in Education" because they reflected the same kind of fears. As soon as I had found all the themes I identified and named them. This involved explaining precisely what each theme covered, and that the names were representative of the data they represented. For example, "Value of Human Interaction and Personal Touch" was chosen to ensure the capture of all data points related to this very wide-ranging category focused on the need to maintain human elements in the educational process. I then collated these into a properly structured narrative, outlining the themes identified in detail. This includes verbatim quotations to show that the identified themes were directly related to the actual words used by the respondents, thus ensuring the analysis was grounded in the words of the participants themselves.

3. Thematic Analysis

Five clear themes were apparent within the data, which are shown below, along with common keywords, in Table 1 followed by a data visualisation in a Heat Map by theme in Table 2. Responses indicate quite clearly a scepticism about Gen-AI use in education by students, and emphasis on the view from participants that there needs to be a maintenance of human interaction and the personal touch which teachers give to students. Many respondents acknowledge potential benefits but also underline the need for cautious implementation so that the enhancement of teaching quality, creativity, and personalised learning is not affected. The general tone that comes out of most responses is one of ambivalence, where the respondents balance the pros and cons of Gen-AI in education.

Theme	Keywords and Word Counts	Total Occurrences
Scepticism and Concern about Gen-AI in Education	Quality (11), Concerned (10), Worry (7), Sceptical (6), Inferior (4), Decline (3), Bias (3)	44
Value of Human Interaction and Personal Touch	Human (51), Personal (23), Interaction (11), Email (19), Personalised (7), Mentorship (3), Guidance (3)	117
Potential Benefits with Caution	Potential (15), Benefits (14), Help (14), Efficient (6), Support (5), Tailor (4), Customise (3)	61
Negative Impact on Teaching Quality and Creativity	Creativity (7), Standardised (6), Generic (3), Inspiring (3), Mechanical (3), Cookie-cutter (2)	24
Uncertainty	Mixed (13), Uncertain (10), Undecided (7), Neutral (7), Cautious (5), Pros and Cons (4)	46

Table 1. Summary of Themes, Keywords and Occurrences to Students thoughts of Lecturer use of Gen-AI in areas of Teaching and Instruction

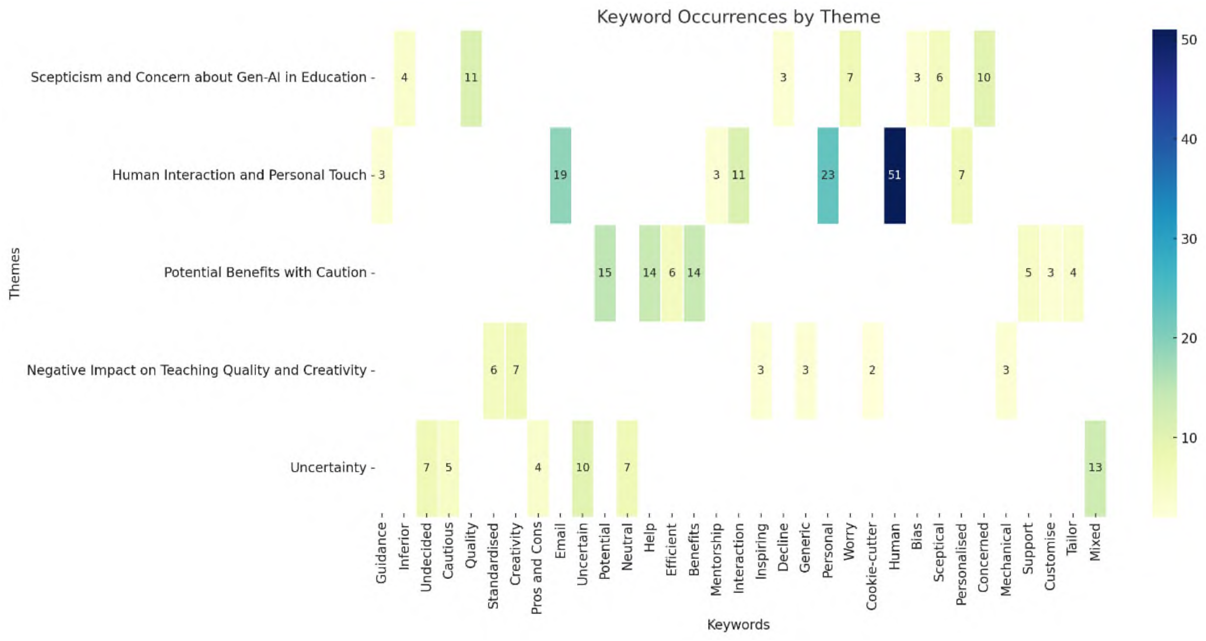


Table 2. Heatmap of keyword occurrences by theme

Theme 1: Scepticism and Concern about Gen-AI in Education

Many of the respondents are concerned about the impact that Gen-AI can have on education quality, authenticity, student engagement, and the value of the teaching profession. The theme of scepticism and concern about Gen-AI in education refers to the different types of apprehensions expressed by the respondents about the effect Gen-AI would have on the quality, authenticity, and overall effectiveness of educational experiences. Some of the keywords that most clearly support this theme include quality, concerned, worry, sceptical, inferior, decline, bias.

Numerous responses indicate deep concerns regarding the potential decline in the quality of educational materials if Gen-AI is integrated. One respondent replied, *"I fear that AI content my within modules will lead to a lowering of the standard of my course,"* reflecting a fear that the introduction of Gen-AI may degrade the standard of learning. Another way it is put is, *"AI content lacks the genuineness and dimension brought by instructors,"* indicating that the quality of the education may be eroded to what might be regarded as superficial Gen-AI-generated materials. The word *"concerned*" and *"worry"* comes up many times, pointing at a general unease about Gen-AI in education. For example, one of the respondent's comments, *"I am concerned about the accuracy of AI-generated content,"* showing concerns regarding the reliability and correctness of information generated by Gen-AI. This is a critical aspect associated within this theme is scepticism towards the role of Gen-AI in education with the respondents’ raised questions about the credibility and effectiveness of Gen-AI-generated content. As one of them says, *"I'm sceptical about the reliability of AI-generated teaching aids,"* which shows a clear doubt on the reliability of Gen-AI tools. Another expresses scepticism more broadly: *"Any lecturer or course creator relying on AI to create a curriculum clearly doesn't care enough about their subject to spend the time creating the right courses,"* highlighting that reliance upon Gen-AI might from a students’ perspective reflect a lack of commitment or effort from educators and their content. There is a general idea here expressed by students that Gen-AI-generated content is poor compared to human-made content. Another responds simply, *"AI-generated materials are subpar and should not form part of any curriculum,"* bringing out the inability of Gen-AI to produce quality learning materials in their view. Another generalised concern is voiced by this respondent: *"AI content in teaching is a disaster waiting to happen; it should be banned,"* showing the view from the respondent that Gen-AI content can seriously harm the practise of education itself.

The last point is bias in the content created by Gen-AI. The respondents are concerned that Gen-AI may further the present prejudices. One shared, *"I'm worried about the potential for AI to perpetuate biases in educational materials,"* thus fearing that Gen-AI may propagate the already existing societal biases into schools and hence affect its fairness and inclusiveness. Another response supports this view, *"AI-generated content may not consider the cultural context of the students," the* student expressing their fear that that Gen-AI might lack nuanced understanding necessary to address diverse student needs effectively.

Theme 2: Value of Human Interaction and Personal Touch

It is important, many of the respondents pointed out, that human interaction and personalised communication with regards to students' emails and feedback are continued in education. Students want more personal interactions with the lecturers, not less; even when that engagement is an email response. This includes concerns about the loss of direct human engagement, the irreplaceability of personalised interaction, and the unique value brought in by mentorship and guidance from the human educator. Numerous responses underscored human input in teaching. The recurrent use of the term *"human"* underlines the fact that there are certain qualities that cannot be replaced in education. For example, one respondent indicated, *"AI can't provide the same level of mentorship and guidance as human instructors."*. This generally points to a fear that Gen-AI is bereft of the empathetic and intuitive abilities required in teaching, which human teachers possess. Indeed, there is a strong belief that something in the human teacher enables him or her to turn the student on to a subject with his or her passion and personal engagement, something Gen-AI is perceived to lack. This personalisation was also a key theme in many of the comments. Many times, the word *"personal"* was mentioned when referring to keeping students personal regarding their relationship with instructors. For example, one responded commented, *"AI can generate helpful content, but e-mail inboxes need to be personal."* The point is, as much as can be, there needs to be personalised communication, especially in aspects relating to relationship development and personalised feedback. Another response, *"I am all for AI in education, but student emails need a human touch,"* again goes on to show that, from the student respondents perspective, while Gen-AI can be used in certain spheres, direct and personal communication must be the preserve of human educators if the quality of the educational experience is to be preserved; even when that 'personal communication' to them is just an email. Another respondent expresses this anxiety on a larger scale, *"I am concerned that AI in education will mean the loss of so much critical human interaction,"* outlining a fear of losing the most vital human element in teaching.

Another key point repeatedly emphasised by the respondents is interaction in the whole process of education. *"AI in teaching may reduce interactive learning"* stated one more respondent. This statement expresses the very profound feeling that Gen-AI is going to reduce the interactive aspect, which underlines the entire process of learning. *"AI in instruction feels impersonal and detached"* is an added insight, insinuating for that student that the lack of human interaction in Gen-AI-driven education could mean detachment and impersonality that they view will affect their learning experience. Email interactions with lecturers were a major area where most of the respondents wanted to see human involvement. *"AI is fine for instructional materials, but emails to students need a human touch,"* said one respondent, capturing the typical sentiment that whereas Gen-AI can assist in the construction of educational content, personal email interactions should be handled by humans. Email itself would seem to be regarded by students as an important person link with lecturers with one of the respondents responded, *"I don't like the idea of email answers by AI, but individual learning experiences sound interesting,"* suggesting that the respondent didn't want Gen-AI to write responses for email; however, at the same time, showed interest in the possibilities Gen-AI had within other educational aspects.

Paradoxically, personalisation in learning, often perceived (or espoused by EdTech companies trying to sell it) as a strong possibility for the application of Gen-AI to create personal learning journeys and materials, was seen by student respondents as something AI couldn't address given diverse needs among students. *"AI-generated content might not be adapted to different learning styles,"* one respondent replied, meaning that Gen-AI may not be able to provide the sort of individual approach required for effective learning among different profiles of students. Another noted, *"The use of AI in teaching could foster a lack of personalised learning,"* thereby putting forward the belief that Gen-AI may not be able to handle the needs of students individually well enough, which is something that human educators are better equipped to do. The two critical elements of personalised learning that Gen-AI cannot replicate, according to the respondents, are mentorship and guidance. According to one of the respondents, *"AI can't provide similar mentorship and guidance like human instructors,"* thus showing a feeling that the mentorship role of the teachers, personal interaction, support, and encouragement, is beyond what Gen-AI can do. Another example is *"The use of Gen-AI in teaching is an insult to the profession of education,"* which again shows a far greater concern for Gen-AI encroaching on two of the most integral parts of the educational process: the mentorship and guidance provided by human teachers.

Theme 3: Potential Benefits with Caution

The Potential Benefits with Caution theme is a nuanced and balanced one by respondents that recognises Gen-AI's potential to improve certain aspects of education while bringing to the fore careful and limited implementation so as not to cause harmful effects. In a nutshell, though the respondents see Gen-AI as a promising tool that can offer diverse advantages, they are very wary of its limitations and urge balance in using it under human oversight. Responses often yielded a place for the potential of Gen-AI to improve educational experience. Typical comments include, *"AI might give benefits but has also the potential issues,"* a balanced view of the role of Gen-AI in education. Another wrote, *"I see the potential of Gen-AI in teaching, but have reservations,"* openness to Gen-AI's potential tempered with caution. Respondents saw the benefits of Gen-AI as striking, especially in those spheres where it can complement the work of a human. For example, *"AI could be useful for repetitive tasks, freeing up teachers to engage more deeply with students,"* shows how students seem well-aware that Gen-AI can take care of mundane tasks to let educators focus on more meaningful interactions. The statement *"AI might be useful in diagnosis—gaps in knowledge—but should not be the entire determinant of how one learns"* speaks much about the quite nuanced perception of students of the utility of AI in diagnosis rather than being the primary driver of education.

Another important dimension in this theme is the supporting role of Gen-AI. Most respondents felt that Gen-AI could prove to be useful support, but not at a cost to the human educator. For example, one participant commented, *"AI could help provide extra resources and support but it must be accurate and reliable,"* thereby bringing out the potential value of Gen-AI as ancillary support and not necessarily as a replacement. The idea of the Gen-AI helping the educator is also captured in comments like, *"AI might be helpful, but it also has significant limitations."* This view respects potential help from Gen-AI but retains an element of its constraining factors, hence for the student that Gen-AI should supplement, not replace, human effort. Efficiency is one recurring benefit attributed to Gen-AI, as a few of the respondents show appreciation for the capacity to make some educational processes more efficient. The comment *“AI could help make for more efficient learning but I’m not sure that’ll work for me"* sums it up: a recognition though Gen-AI is purportedly able to make learning processes more efficient, a personal concern out the applicability of the benefits.

Customisation of educational experiences, often hailed as one of the positive abilities of Gen-AI, also is viewed through this subtle lens by respondents *"AI can help tailor the learning experience to individual needs but it might miss the nuances of teaching that work for me"* shows an understanding of the customisation ability of Gen-AI. Similarly, another respondent wrote, *"AI might help in personalising the learning experience but it may not capture the full context of our needs,"* thereby elaborating that while Gen-AI can facilitate personalised learning, in the respondents view it misses the depth and understanding that teachers possess. So, while, on the one hand, there is an appreciation for customisation by Gen-AI, on the other hand, there is a distinct call for cautious implementation. *"AI has potential but I'm not quite sure about its overall impacts on learning"* succinctly reflects the ambivalence towards the set of customisation abilities of Gen-AI. Similarly, *"Gen-AI could be a very useful tool but it also raises concerns"* indicates cautious optimism regarding the role of Gen-AI in education.

Theme 4: Negative Impact on Teaching Quality and Creativity

There is a strong feeling that Gen-AI could lead to a much more standardised, less imaginative, and less exciting learning experience, and hence not serve to develop learning that is deep and nuanced. Opinions of the respondents answer that compared to human teachers, Gen-AI in pedagogy is less creative. The creativity in teaching was regarded by respondents as one of the main factors that does not let students get bored or disturbed. One of the participants responded, *"AI lacks creativity in developing innovative modes of* *teaching."* This picked up the fear that Gen-AI-generated content might be too rigid and formulaic, lacking the very innovative approaches which human teachers naturally incorporate into lessons. Many respondents often fear that Gen-AI will lead to a more homogenised, less personalised way of learning. One comment said, *"AI in instruction could lead to a cookie-cutter approach to education,"* thus meaning that Gen-AI would turn out the same content and not take into consideration students' various needs. Another concern voiced in this regard was, *"AI in education might lead to a one-size-fits-all approach that doesn't work for everyone."* The fear is that Gen-AI will deliver generic content that lacks the specifics and context that make learning materials relevant and engaging, *"Using AI for creating teaching materials might make the learning experience too generic."* With students concerned that the effectiveness of materials generated through Gen-AI will be lost because they are not geared toward the special interests and backgrounds of students. One respondent responded, *"I worry that AI could standardise education way too much reducing the richness of diverse teaching methods,"* reflecting a concern about Gen-AI homogenising teaching approaches and, therefore, doing a disservice to educational diversity. Furthermore, it is argued that the standard and quality of educational experience will decrease. One respondent claimed, *"The implementation of AI in teaching could reduce the quality of our education,"* which considers using Gen-AI to have a direct impact on reducing standards.

Students perceive lecturer inspiration is a very important component of good teaching, and most of the respondents are of the view that Gen-AI cannot play an inspirational role that a teacher does. *"AI doesn't have the ability to inspire students like real teachers can,"* one of them wrote. This strongly brings out the thinking that motivational and inspirational functions of teaching are essentially humane and therefore impossible for Gen-AI to replicate. Another concern is the mechanical nature that Gen-AI-generated content can take, with some of the respondents concerned that Gen-AI will make education too robotic and less personal. One of the respondents replied, *"AI may make learning less interactive and more mechanical."* This shows that there is a fear that the dynamic and interactive nature of human teaching could get lost with the increased use of Gen-AI in education. The strongest concerns are related to the narrative about the possible negative impact of Gen-AI on teaching quality and creativity. Here, what the keywords underline is one clear concern: the perception that Gen-AI just cannot replicate the nuanced, creative, and inspirational possibilities of human educators. This theme underlines the belief from respondents that while Gen-AI might offer efficiencies, this comes at the cost of richness and effectiveness of the educational experience.

Theme 5: Uncertainty

Several of the respondent's express ambivalence, recognising both the potential upsides and great risks associated with Gen-AI in education, calling for a balanced approach. Many respondents showed mix positive and negative sentiments towards Gen-AI in education. For example, one respondent replied, *"I can see both the benefits and drawbacks of using AI in teaching,"* which generally reflects an understanding that Gen-AI can bring improvements in education but also some realisation about its limitation. Another reflection of this ambivalence came from the following comment: *"I'm on the fence about AI; it has pros and cons*." This statement underlines the inner conflict of the person who values Gen-AI but fears its implications. The uncertainty of the effect could be further developed from the responses, where some doubt is cast on the potential of Gen-AI to have an overall impact on education. One common theme was, *"It is hard to say whether AI will improve or harm my education at this point."* this is because the student feels there is not clear evidence that has been presented to them, or from their own experience, in the long-term effects of Gen-AI in the educational context; that they don't have enough information make a judgement at this time.

Most of the respondents were undecided, neither in agreement nor disagreement with the use of Gen-AI. Comments like *"I'm undecided about AI in teaching; it could be both good and bad"* show this indecisiveness. This need therefore is a call for more intense discussions and reviews on the role of Gen-AI in education, hence helping people to form definite opinions. Neutrality was another common perspective that viewed that many respondents neither strongly supported nor opposed the use of Gen-AI in education. For example, *"AI in education. It could help or hurt."* Again, the respondents accept potentials but are at the same time conscious of the limitations and just how important careful implementation will be. Caution was a recurrent theme. Many respondents clearly advocated careful and measured incorporation of Gen-AI in education. One says, *"AI might be useful in some areas but I am not entirely convinced,"* cautioning optimism that seems to have concerns about overreliance on Gen-AI tempered. One of them noted, *"I am positive about AI in Education. It should, however, be used with caution to avoid over-reliance,"* an argument that a balanced approach should take advantage of the utility Gen-AI would add to education without obstructing human aspects of teaching. Lastly, many respondents mentioned the pros and cons of Gen-AI, which means they have considered all the probable effects of it. A common comment was, *"Gen-AI might make education better but it also presents huge challenges, I just don't know,"* thereby summarising mixed perspectives entailing optimism and concern.

5. Conclusions and an Agenda for Future Study

This study explores the perceptions university students have when it comes to Gen-AI being used by lecturers within education, where preliminary findings have identified a spectrum of views from the pessimism to optimism. Analysis of the results identifies five key themes: scepticism and concern about Gen-AI, the value of human interaction and personal touch, potential benefits with caution, negative impact on teaching quality and creativity, and uncertainty. The paper is led and titled with a student comment *"AI lacks the ability to inspire students like real teachers can."* which reflects what many students surveyed for this paper would seem to genuinely fear; that the use of Gen-AI by their lecturers will degrade the quality of education, that Gen-AI-generated content is less valid, and that they will also lose out on important human interactions with the teachers who inspire their learning. Their concerns are that Gen-AI might standardise and depersonalise the quality of education, reducing richness and diversity in learning experiences. Areas for future study could include ways in which the concerns of the students are addressed, or lecturers could strive to integrate Gen-AI in a manner that would complement and enhance their work while not giving a feeling of deterioration of education quality amongst their students. This could mean gathering much more feedback from students on using Gen-AI tools for tasks like grading and providing additional resources or feedback while, at the same time, aiming to maintain personal interactions and tailored feedback—both of which students have said are valuable.

Students overwhelmingly stress the importance of keeping personal communication and human mentorship in education. They look for the special qualities of a human teacher to be irreplaceable, which are one of empathy and inspiration, necessary for learning to be effective. Future research may wish to structurally explore ways that Gen-AI can be used to augment—never replace—human elements of teaching, or to deconstruct this 'irreplaceable' quality to see whether it fades over time as Gen-AI becomes more of a societal norm. Whether personal email responses, face-to-face mentorship, and tailored guidance should remain central to the educational experience may in the future depend on the views of students and their willingness to accept or not these alterations. While acknowledging the possibilities of Gen-AI in efficiency and support, students, in return, seem to push for balance and proof of them benefiting, and it's evident that at this early stage of educational Gen-AI usage, lecturers have very minimal empirical work to show students that Gen-AI could be able to afford them real benefits. While Gen-AI is being promoted by many Edtech companies to lecturers and academics process the mundane and enable personalisation, students remain ambiguously enthusiastic, suggesting even small-scale change may be worth investigating in terms of how students perceive the role of technology to supplement rather than replace human educators and, therefore, detract or benefit their educational experience. Therefore, continuously seeking student perspectives on lecturers' strategies for integration of Gen-AI may be useful in terms addressing proactively student's concerns and help inform future steps.

There is a view from the students in this research that Gen-AI could result in an education that is less creative and more standardised. The fear is that the very innovative and engaging qualities developed by human teachers might get missed out in the Gen-AI-generated content, leading to a mechanical learning environment where inspiration to learn is lacking. It may be that lecturers need to consider providing students with training in how best to engage with Gen-AI tools as a support for their learning, ensuring that they understand what the tools can and cannot do. Or, from the sceptical perspective, it is possibly worth researching whether knowing the Gen-AI tools (for example by investing in Gen-AI literacy amongst students) makes any difference at all. Student ambivalence is a function of an appreciation of the potential value of Gen-AI alongside profound concerns. This mixed perspective demands further research and careful thought into the role of Gen-AI in education: to make sure that it enhances the learning experience of students without undermining the essential human elements that they perceive as valuable or irreplaceable.

Since this study has an exploratory nature, several ways forward are open for future research to further shed light on students' opinions about Gen-AI usage in education. This may involve long-term studies that would trace the change in perception over time by means of carefully controlled observations of how students' initial perceptions of Gen-AI in education have evolved as they become more experienced users of Gen-AI tools and determine the long-term impact of the integration of Gen-AI on learning outcomes and students' satisfaction. This study is from student responses in April 2024. Will over time students become more accepting of Gen-AI? Would the same set of students in three years' time be less sceptical, or indeed, more? Such research could be expanded to a wide range of educational settings and disciplines to reveal possible variations in views on Gen-AI across different academic contexts. This will add insight into how subject-specific needs and pedagogies impact perception of Gen-AI. Comparative studies across different countries and cultural contexts may further indicate how cultural attitudes toward technology and education influence the acceptance and effectiveness of Gen-AI tools. Experimental studies that contrast traditional pedagogical approaches with Gen-AI-supported methods can provide empirical evidence on the actual impact of Gen-AI on learning outcomes and student engagement. In-depth qualitative studies, such as focus groups or individual interviews, may further elaborate on the nuances behind students' views and uncover concerns and suggestions not evident in survey-based research.

The more research pursued in these directions, the better lecturers are equipped to have a robust understanding of how Gen-AI can be effectively integrated into their classrooms to ensure it serves to enhance, rather than detract, from a learning experience where the (current at least) students perceive their human interaction with them as teachers to be inspiring and irreplaceable.

References

Abeysekera, L., & Dawson, P. (2015). Motivation and cognitive load in the flipped classroom: Definition, rationale and a call for research. *Higher Education Research & Development, 34*(1), 1-14. https://doi.org/10.1080/07294360.2014.934336

Akçayır, G., & Akçayır, M. (2018). The flipped classroom: A review of its advantages and challenges. Computers & Education, 126, 334-345. https://doi.org/10.1016/j.compedu.2018.07.021

Ayanwale, M. A., & Molefi, R. R. (2024). Exploring intention of undergraduate students to embrace chatbots: From the vantage point of Lesotho. International Journal of Educational Technology in Higher Education, 21(20). https://doi.org/10.1186/s41239-023-00355-0

Bishop, J. L., & Verleger, M. A. (2013). The flipped classroom: A survey of the research. *Proceedings of the ASEE National Conference, Atlanta, GA, 30*(9), 1-18.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE Life Sciences Education, 6*(1), 9-20. https://doi.org/10.1187/cbe.06-12-0205

Cohen, L., Manion, L., & Morrison, K. (2017). *Research methods in education* (8th ed.). Routledge. https://doi.org/10.4324/9781315456539

Dericks, G., Thompson, E., Roberts, M., & Phua, F. (2022). Online learning during the Covid-19 pandemic: How university students' perceptions, engagement, and performance are related to their personal characteristics. *Current Psychology*, 41, 1-15. https://doi.org/10.1007/s12144-021-01450-8

Draper, S. W., & Brown, M. I. (2004). Increasing interactivity in lectures using an electronic voting system. *Journal of Computer Assisted Learning, 20*(2), 81-94. https://doi.org/10.1111/j.1365-2729.2004.00074.x

Gilboy, M. B., Heinerichs, S., & Pazzaglia, G. (2015). Enhancing student engagement using the flipped classroom. *Journal of Nutrition Education and Behavior, 47*(1), 109-114. https://doi.org/10.1016/j.jneb.2014.08.008

Hart, C. (2012). Factors associated with student persistence in an online program of study: A review of the literature. *Journal of Interactive Online Learning, 11*(1), 19-42.

He, Y., Chen, Q., & Kitkuakul, S. (2018). Regulatory focus and technology acceptance: Perceived ease of use and usefulness as efficacy. Cogent Business & Management, 5(1), 1459006. https://doi.org/10.1080/23311975.2018.1459006

Kay, R. H., & LeSage, A. (2009). Examining the benefits and challenges of using audience response systems: A review of the literature. *Computers & Education, 53*(3), 819-827. https://doi.org/10.1016/j.compedu.2009.05.001

Kebritchi, M., Lipschuetz, A., & Santiague, L. (2017). Issues and challenges for teaching successful online courses in higher education: A literature review. *Journal of Educational Technology Systems, 46*(1), 4-29. https://doi.org/10.1177/0047239516661713

Krueger, R. A., & Casey, M. A. (2014). *Focus groups: A practical guide for applied research* (5th ed.). Sage Publications.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2014). The effectiveness of online and blended learning: A meta-analysis of the empirical literature. *Teachers College Record, 115*(3), 1-47. https://doi.org/10.1177/016146811311500307

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1), 1-13. _https://doi.org/10.1177/1609406917733847

Opesemowo, O. A., & Ndlovu, M. (2024). Artificial intelligence in mathematics education: The good, the bad, and the ugly. Journal of Pedagogical Research, 0 (0), 1-14. https://doi.org/10.33902/JPR.202426428

Saldana, J. (2009). *The Coding Manual for Qualitative Researchers*. Sage Publications.

Sevnarayan, K. (2024). Exploring the dynamics of ChatGPT: Students and lecturers' perspectives at an open distance e-learning university. *Journal of Pedagogical Research*, 8(2), 212-226. https://doi.org/10.33902/JPR.202426525

Southworth, J., Migliaccio, K., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). Developing a model for Gen-AI Across the curriculum: Transforming the higher education landscape via innovation in Gen-AI literacy. *Computers and Education: Artificial Intelligence*, 4, 100127. https://doi.org/10.1016/j.caeai.2023.100127

Yilmaz, R. M., & Şahin, M. C. (2021). Students academic and social concerns during COVID-19 pandemic. Education and Information Technologies, 26, 6897-6917. https://doi.org/10.1007/s10639-020-10315-2

Barring Heaven's Gates: AI Secured Clouds as a Mechanism for Enhanced Research Security

Brendan Walker-Munro[1], Ravi Nayyar[2]

[1] Faculty of Business, Law & the Arts, Southern Cross University, Australia

[2] University of Sydney, Australia

Abstract (around 150 words)

The purpose of this paper is to explore the policy and legal underpinnings of the emerging use of AI in secured cloud environments for the transmission and storage of classified or sensitive research information in higher education institutions (HEIs). Using research security as a conceptual lens, we propose that universities will be increasingly required to adopt such secured exchanges for the protection of research data; not just from a good practice perspective, but also to protect privacy, intellectual property rights and meet any obligations under cybersecurity or critical infrastructure laws. Early adopters of the technology will benefit the most, with those institutions which do not adequately embed AI in the secured cloud offerings they uptake likely to be more vulnerable to internal and external threat actors.

Keywords: research security, secured cloud, higher education, regulation

1.0 Background

The ubiquitousness of information technologies in higher education institutions (HEIs) is difficult to ignore. HEIs are increasingly moving their offerings to online environments, offering time-poor students of all ages and backgrounds the ability to earn a degree from their living room. Concurrent with upheavals in teaching methodologies, the nature of research at HEIs has also mutated. No longer restricted to single laboratories or campuses, the nature of contemporary academic research is part of a globalised network of international scholars, all of them collaborating and competing on an international stage for funding and reputation (Wagner, 2018).

HEIs have also long been places of research for technologies and knowledge with military, security, intelligence gathering or policing utility. This can create wicked challenges for HEI administrators who must balance the needs of the administration and security state on the one hand (often requiring secrecy and non-attribution) with the needs of the academic community on the other (whom demand open publication and science in the name of the greater good). Since antiquity, intelligence and security agencies have fostered such close relationships with HEIs that ‘no one should be surprised that [universities] own complex bureaucracies of knowledge creation, gathering and dissemination might just be of interest to a security and intelligence system looking anywhere and everywhere for knowledge of and the means for protection against threat’ (Gearon, 2019: 11).

In that environment, the pervasiveness of technologization throws up numerous regulatory challenges. Written exams can now be completed in seconds by large language models almost indistinguishably from human effort (Gilson et al. 2022; Katz et al. 2024). Some HEIs have returned to the oral exam or “*viva voce*” to combat cheating (Renzella, Cain, & Schneider, 2022). HEI research has not been spared. For example, the very first time that COVID-19 appeared in Switzerland was in a secured laboratory using genomic templates sent by foreign scientists over email and predating the first human infection by two full weeks (Young, 2023: 192).

Simultaneously, HEIs are traditionally poor environments for security compliance, especially in the domain of cybersecurity (Bongiovanni, 2019; Bongiovanni, Renaud, & Cairns, 2020). HEIs favour principles of openness, transparency, sharing and collaboration, and can react very strongly in response to curbs on those freedoms. In some nations, these principles may be embedded as constitutional freedoms; in others, they may be protected national ordnances or other legislative directions. By favouring an attempt to balance both, universities inevitably compromise both, leading to ‘lack of clarity about the proper scope of universities’ autonomy when managing speech tensions…stakeholder confusion, concerns about institutional integrity, and declines in public trust in universities’ (Gross Methner, 2019: 361).

At this intersection between higher education, technology, and national security, the developing paradigm of “research security” has started to emerge. Recognising that HEIs are ever more involved in the fundamental research underpinning the next generation of military, security and intelligence products, governments around the world have started to ask the question “how do we stop foreign entities stealing our knowledge?” Unfortunately, in many cases the inquiry seems to stop there, with governments in the US, UK, Canada, Australia, New Zealand and the Netherlands (amongst others) enacting laws to restrict or contain information disclosure to foreign actors even in the face of international obligations to the contrary (Walker-Munro, 2024a; Walker-Munro, 2024b; Walker-Munro, 2024c). In other countries – particularly in the European diaspora – the contrary has occurred. Emerging first in Swedish think tank STENT (Gothenberg, 2022) and the works of scholars such as Professor Tommy Shih (Shih, 2023; Shih, 2024), the notion of "responsible internationalisation" has since become embedded in both transnational (Council of the European Union, 2024) and national norms (HRK: German Rectors' Conference, 2020; Ministry of Higher Education and Science, 2022; Universities UK, 2022; British Council, 2024).

Thus, the country selection for this paper is not arbitrary, as Sweden and Australia share unique characteristics relating to HEI security. For example, both Australia and Sweden have relatively small populations but highly active HEI research profiles that rely on international collaboration. Over the past three years, both Australia and Sweden have increasingly collaborated with China (up 5.5% and 7.2% in the last 12 months respectively) whilst collaborations with long-time economic and national allies the US and United Kingdom have decreased (SciVal, 2024). Both Sweden and Australia spend roughly equivalent amounts (which are comparatively massive based on their small populations) on their defence and military: 1.3% and 1.9% of their GDP (World Bank Group, 2024). And both Sweden and Australia have close ties between their military and certain universities, ties which are being repeatedly challenged by both a volatile geopolitical order and growing social sentiment gravitating towards the severing of such ties (Lundin, Stenlås, & Gribbe, 2010; Eldeblad, 2023; Krien, 2024).

At the same time, both Australian and Swedish HEI environments operate under rising governmental scrutiny and influence. In Sweden, the government moved in 2023 to shorten tenure to academic boards from 3.5 years to 17 months (as well as introducing mandatory security screening), a move labelled by many 'a threat to the authority of the higher education institutions' and 'against a democratic spirit' (Myklebust, 2023). Meanwhile in Australia, extensive amendments to national security laws since 2018 to counter "foreign interference" have routinely been criticised by HEIs for regulatory overreach and disproportionate impact on academia (Group of Eight Australia, 2022; Thomson, 2023). Both countries have also undertaken extensive independent evaluations of their rights to academic freedom, and in both cases concluded that recent governmental measures over the last five years had involved limits on academic freedom at best, and a “chilling effect” on controversial or unpopular research at worst (French, 2019; Tovatt et al. 2024).

Thus, we seek here to examine the emergence of secured cloud environments and the influence of artificial intelligence (AI) on securing HEI research. Secured clouds enable researchers to collaborate and share data irrespective of the physical or virtual barriers of their home countries; AI potentially permits them to do it safely at scale and speed. It is not the purpose of this paper to canvas a comprehensive definition of AI, given that the literature is rife with fulsome attempts to do so (Zhang & Aslan, 2021; Mishra and Tyagi 2022; Lin et al. 2022; Polamarasetti, 2024). Instead, we adopt a working definition that AI ‘employ[s] algorithms to examine data, gain knowledge, and determine a forecast or course of action without involving humans’ (Alzoubi et al., 2024).

We focus on Australia and Sweden as case studies towards the adoption of AI-secured clouds for protecting HEI research. Part 2 of this paper will conduct a non-technical examination of secured cloud environments *per se*, and the benefits that arise from their deployment. It will also briefly canvas Australia and Swedish developments in law and policy that benefit the development of secured clouds for securing HEI research. Part 3 will briefly examine the utility of AI as a form of protection for these highly critical secured environments, and the implications for law and policy that such approaches might raise. Part 4 will then make some general recommendations and observations for future research that will be needed to develop secured cloud environments into the future of higher education before concluding.

2.0 Secured Clouds, Australia and Sweden

A universal definition of “the cloud” has been historically difficult to establish, as individual companies may use a mix of proprietary and commercial-off-the-shelf technologies and software to establish, access or maintain their specific cloud environments, and scholars regularly use their own versions for the term. A useful definition is supplied by the US National Institute of Standards and Technology (NIST), who described cloud computing as ‘a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction' (Mell & Grance, 2011: 6). More colloquially, comedian Kitty Flanagan is well-known for her viral video where she described the process of cloud computing as 'when you send all your digital files to the cloud…you're really just sending stuff to someone else's much bigger computer' (ABC, 2018).

NIST defines the essential characteristics of a cloud environment to involve a.) access to computing infrastructure on-demand by a user, b.) consistent access to resources irrespective of the access device used, c.) multiple virtual and physical IT infrastructure, d.) the ability to rapidly scale and descale a cloud environment dependent on need, and e.) automatic metering of resource usage, supply and demand (Mell & Grance, 2011). The three predominant forms of cloud offerings appear to be Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The precise delineations of these offerings have been dealt with elsewhere (Carroll, van der Merwe, & Kotze, 2011; Santos et al. 2024) and do not warrant repeating here; suffice to say, that while these categories use different approaches in infrastructure and software delivery, they still have similar risk factors affecting their uptake and operationalisation (Shahzad, 2014).

As a concept in computing, the cloud has done significant work in equitizing access to information across numerous domains whilst causing 'political disruption' because of the cloud's power to force re-evaluations of both data security and freedom of access (Mittelstadt, 2017; Cohen, 2019). The increased use of cloud computing has also driven both an increased desire for access to information, commensurate with an increased need to secure that information from unauthorised access. In turn, legal and regulatory structures continue to struggle in appropriately reacting to cloud environments, either by permitting increasingly unregulated "edge" cases of data use and access, or by encouraging market evolution or revolution to innovate around restrictions (Amoore & Raley, 2016). So, it is not surprising to scholars and practitioners alike that the security of cloud providers exists as a discrete field within the broader cybersecurity ecosystem, with cloud providers driving an economic need for cloud security services (Niemiec et al., 2022).

2.1 Australian Secured Cloud

In early July 2024, Australia's National Intelligence Community (NIC) – comprising of the ten national security and intelligence agencies such as the Australian Security and Intelligence Organisation (ASIO), Australian Security Intelligence Service (ASIS), the Australian Federal Police and others – announced the development of a classified cloud environment. The Top Secret (TS) Cloud, developed in partnership with Amazon Web Services (AWS), was said to 'provide a state-of-the-art collaborative space for our intelligence and defence community to store and access top secret data' (Department of Defence, 2024). It also 'will be purpose-built for Australia's Defence and National Intelligence Community agencies to securely host our country's most sensitive information', with both the artificial intelligence and machine learning features of Azure said to be active features of the TS Cloud (Department of Defence, 2024).

At roughly the same time that Australia's NIC was developing their secured cloud environment, a collaboration between Curtin University and well-known cyber solutions company CyberCX resulted in the development of "Nebula". According to press releases surrounding the event, Nebula is 'powered by Microsoft's sovereign PROTECTED accredited Azure, M365 and security platforms providing the innovation community with access to high performance computing, data modelling and artificial intelligence capabilities' (CyberCX, 2024). Further, Nebula is intended to 'allow users to develop and store research-related files including sensitive data sets, research papers and other key documents safely separated from their everyday workplace environment' (CyberCX, 2024).

2.2 Sweden Secured Cloud

In July 2017, Swedish authorities were alerted to a massive data breach involving the plain-text dissemination of the country's entire drivers licence database to a foreign marketing firm. The database, provided in full and without protection or encryption, reportedly identified members of Swedish special forces personnel, defects in government and military vehicles, as well as persons in witness protection programs (Saarinen, 2017).

Sweden currently does not appear to possess any publicly disclosed secure cloud environment, either for military/intelligence purposes or secured university collaborations (Ravichandran & Balmuri, 2011; Usman Ali & Ayub, 2012). That said, the country is well-placed to adopt a secured cloud: Swedish governmental uptake of digital technologies is amongst the most mature of the EU Member States (Lilyanova, 2022), and Sweden is also one of the few countries in the world that has 'created an entire [economic] system based on security provision domestically and abroad', such that Sweden operates 'in the unusual company of some of the most powerful states in the world despite its relatively modest material power' (Coetzee & Berndtsson, 2023: 171). Simultaneously, moves in the last year by service providers Tietoevry Connect and Evroc have seen the establishment of sovereign cloud infrastructure in Norrbotten and Stockholm, with more capacity (consisting of a total of eight hyperscale data centres) promised over the coming years (Gotor, 2023; Evroc, 2024).

2.3 Law and policy for secured clouds

Both Australia and Sweden adopt protective legislation designed to shelter computing environments from espionage, sabotage or forms of malicious interference. Under the Commonwealth of Australia's *Criminal Code*, engaging in acts involving the theft of information, disruption or manipulation of computer data, or interference in domestic political or governmental processes are treated as the gravest crimes of national security (Kendall, 2024). Further, Australia also adopts a *Security of Critical Infrastructure Act 2018* (SOCI), requiring operators of designated infrastructure – including cloud service providers – to meet compliance and reporting obligations as well as provide for enhanced resilience and cybersecurity (Mitchell & Samlidis, 2021). Certain provisions of Australia's *Privacy Act 2018* also apply to cloud infrastructure, but only in patchy and unpredictable ways (Adrian, 2013).

In Sweden, both general criminal law and the *Act on Criminal Responsibility for Terrorist Offences* imposes penal sanctions on persons who commit computer offences against government entities (Falk, 2022). Further, regulatory legislation including the *Protective Security Act* (PSA), the *Electronic Communication Act*, and the *Data Protection Act* obligate numerous service providers to secure not only the information they hold from outside interference, but also the integrity and availability of the systems that hold that information. Sweden's PSA applies automatically to any person that engages in ‘security-sensitive activities’, which are broadly defined to meet criteria of being ‘activities that are of significance to Sweden’s security’ and/or ‘covered by an international protective security commitment that is binding on Sweden’ (i.e., NATO regulations or EU directives). Sweden also has the added regulatory complexity of belonging to the European Union (EU) to be considered. Adoption of a secured cloud environment for research collaborations will require measures for anonymization, pseudonymization, and encryption to meet the requirement of the *General Data Protection Regulation* (Issaou, Örtensjö, & Sirajul Islam, 2023).

However, many of these protections exempt HEIs either implicitly (by providing exemptions, such as in the Swedish Higher Education Act’s protection of academic freedoms) or explicitly (i.e., in Australia, the “enhanced” security obligations of the SOCI Act do not apply to HEIs). Whilst the obligations may in fact fall upon the cloud service provider/s themselves, HEIs that are exempted offer a point of entry to such providers for both *bona fide* researchers on the one hand, and potential insider threats on the other. Insider risk is widely misinterpreted and insufficiently examined in HEI settings in the literature (Ulven & Wangen, 2021), but other forms of law and policy can help “bridge the gap”, i.e., Australian HEIs can be required to adopt elements of the Protective Security Polic Framework (PSPF) on an *ad hoc* basis (Department of Home Affairs, 2025), whilst Sweden seems to have a more mature approach in which risk factors for insider threat are more commonly explored during hiring processes (Ulven & Wangen, 2021; Bergstrom, 2023).

3.0 Securing Cloud with AI

Despite the apparent benefits, HEIs appear unwilling to adopt cloud computing environments at scale. Many of the factors which influence these decisions appear related to concerns over the security of these applications, confidentiality and integrity of data uploaded to the cloud (Changchit & Chuchuen, 2018). The distributed nature of cloud computing requires users to sacrifice control over their data – including precise knowledge of where that data physically exists and who has access to it – in exchange for ease of access and sharing (Shei et al., 2016). Indeed, both internal and external threats loom large in the minds of adopters and users of cloud services (El-

Gazzar, Hustad, & Olsen, 2016; Kinuthia & Chung, 2017). At least one empirical assessment has determined that although we may trust cloud computing at a societal level, individual users' perceived risks, social influence and peers in using cloud services greatly influences their formation of trust (Ho & Velazquez, 2015).

At the same time, cloud services provision is a tightly contested market with extremely poor vendor diversity. Three players dominate the market share of cloud provision (Haranas, 2024): Google Cloud (11%), Microsoft Azure (24%) and Amazon Web Services (31%). This is problematic from the viewpoint not only of market diversity, but cybersecurity, where one of the predominant risks for deploying secured clouds comes from the justifiable concern in properly trusting vendors in highly resourced jurisdictions. Even the largest of cloud providers (i.e., Microsoft) has suffered outages (Ghosh et al., 2022), human errors (Department of Homeland Security, 2024) and actual compromise by foreign actors (Franceschi-Bicchierai, 2024). The Cyber Safety Review Board (CSRB, 2023) report into the Summer 2023 Microsoft Exchange Online incident likewise painted a harrowing picture of Microsoft's security culture as 'wholly inadequate'. Such concerns are likely to be writ even larger in the future since President Donald Trump abolished the CSRB in 2025 (Croft, 2025).

Further, cloud environments whilst are "secured" par-for-the-course from outside threats, this does little to protect the data from insider threats, i.e., a person with legitimate credentials and reason for accessing the secured environment, with malicious intent to do so (Kandias, Virvilis & Gritzalis, 2011; Alhanahnah, Jhumka & Alouneh, 2016). In a secured cloud, insider threats are both severe and highly probable: a person who has the appropriate access rights to an information system and misuses his or her privileges poses one of the gravest risks to HEI research security (Cotton, Viñas-Racionero & Scalora, 2024). Identification of such insider threats is also neither straightforward nor predictable: '[m]itigation of this problem is often complicated… an insider can focus on a variety of target systems and orchestrate his attack motivated by a number of reasons' and can also use 'the privilege of time, so as to study the information system and deploy a serious attack, which is very difficult to predict and detect' (Kandias, Virvilis & Gritzalis, 2011: 1613).

From that perspective, how cloud environments are secured and the mechanisms for securing those environments from a legal and policy standpoint forms an important part of considering their utility in HEIs. This is because a standardised cloud environment – in which research data is uploaded to distributed network infrastructure that may be in multiple countries and/or geographic locations – can make foreign interference, manipulation and theft *easier* because the researcher or research group no longer controls the physical location of the infrastructure or who has access to it. This is a particularly salient concern where cloud storage may be hosted in countries with domestic or municipal laws enabling warrantless access to digital information by security or intelligence personnel (Parasol, 2018).

3.1 Securing Clouds with AI

Given that the principal security concerns for cloud environments emerge from third-party hosting of data, complexities of underlying infrastructure and access arrangements to that infrastructure, use of AI technologies and systems offers an attractive mechanism for securing these enterprises (Spanaki et al., 2018; Rakgoale et al. 2024). The nature of risks which face cloud environments for the conduct of HEI research are largely unchanged from those which implicate industry. For example, HEIs will still need to protect cloud-based data from privacy breaches, disclosure or theft of valuable intellectual property, and the degradation of trust in both the participating HEIs and the cloud service providers (Zuo & Hu, 2009; Ali, Nagalingam & Gurd, 2017).

The benefit of AI with respect to cloud security – especially in the research security space, where management of insider threats are critical – can most be realised in three areas: *access control* (limiting who can access a secured cloud, how, and where from), *transactional monitoring* (identifying patterns and flagging abnormalities in access, review, modification or exfiltration of data) and *anomaly management* (AI agents can be used to triage or prioritise incidents). Existing research has already identified the utility of AI incorporating both machine learning and deep learning models to separate regular network activity from that which might be anomalous or harmful, as well as “learn” from previous assessments to further a granular and accurate standard of threat detection (Dasgupta, Akhtar & Sen, 2022; Kasongo 2023; (Alzoubi et al., 2024).

The use of a secured cloud environment (and specifically with AI) thus has a significant appeal to the adoption of research security, because the creation and maintenance of university research collaborations involve the taking of risks, and arbitrary restrictions on who can and cannot collaborate with a HEI can have devastating flow-on and unintended effects on research arrangements. It is perhaps unsurprising then that at the same time as the research security community has been calling for responsible internationalisation, supply chain scholars have established the notion "balanced resilience"; that is, applying the appropriate appetite of risk taken in managing the chain relative to the size and scope of the threat (Pettit, Croxton, & Fiksel, 2013; Gualandris & Kalchschmidt, 2014; Gualandris & Kalchschmidt, 2015). The utility of AI in protecting against insider threats – indeed, *specifically* regulating the access control, transactional monitoring and anomaly management in HEI cloud environments – can potentially conserve resources and offer efficiency to a notoriously underfunded domain of university management (Shih, 2023; Shih, 2024)

3.2 Law and Policy Implications for AI-Secured Clouds

What then are these implications for these emerging AI-secured cloud environments?

Firstly, HEIs that look to adopt secured clouds will need to take into account the level, scale and standard of AI-specific regulation that might then apply to their operations. For example, Australia is still yet to formulate a concise policy position on AI or cloud risk (Horton, 2024), but Sweden – as a member of the EU – is bound to operate in accordance with the terms of the AI Act (Laux, Wachter & Mittelstadt, 2024). However, one of the significant blind spots in these forms of regulation is currently the approach of cloud vendors seeking to add untested or poorly implemented AI to their offerings. By "AI-fying" their storage solutions, such providers potentially obscure the very real data governance issues at play (Zhang & Aslan, 2021; Erdmann & Toro-Dupouy, 2025).

Secondly, HEIs will need to be mindful of the interactions between transnational municipal laws can also present a challenge to offshore cloud data hosting, such as the difficulties posed by conflicts between the United States Clarifying Lawful Overseas Use of Data (CLOUD) Act and the EU GDPR's regulation of cross-border data transfers (Rojszczak, 2020). In Australia, the SOCI applies *inter alia* to any cloud services provider that holds "business critical data", irrespective of whether the provider uses infrastructure placed beyond the territorial reach of Australian law (Department of Home Affairs, 2025).

Thirdly, the criminal law will continue to be adapted in the face of emerging technologies to capture new forms of offending. In doing so, there is always the possibility that lawmakers may over- or under-regulate any given emerging technology or prohibit activities which are otherwise legitimate and innovative. For example, criminals have – like businesses and HEIs – become increasingly attracted to claims of hosting services that are purportedly "bulletproof" (in the criminal context, this means hosting infrastructure resilient to law enforcement interventions such as takedowns or blocking: ASD, 2025). Attempts to prohibit these offerings and/or subject their owners to penal sanctions may end up inadvertently capturing legitimate market offerings of AI-secured clouds, and not just in the HEI segment.

4.0 Conclusions, Reform and Further Research

This paper has surfaced some broad-level concerns with the use of AI technologies and systems in securing HEI research. AI offers a significant benefit for improving research security through the mitigation of potential insider threat and cybersecurity risks in the adoption of secured cloud environments. Appropriately enabling cloud security, especially in HEIs, is therefore a critical concern where researchers might increasingly rely on cloud computing, and the traditional cybersecurity measures experience challenges in protecting cloud environments (Zhang & Aslan, 2021).

One avenue of potential reform could be to extend existing legal protections or obligations applying to cybersecurity (such as the Swedish Higher Education Act or Australia's SOCI Act) to HEIs. In doing so, proposed laws will need to be subject to comprehensive consultation and stakeholder management, lest entirely legitimate and beneficial activities end up proscribed by an overzealous administration. Law and policymakers will need to address the intersection of AI and cloud security, especially in the HEI environment which remains highly unique and characterised by features found in no other field of enterprise (such as academic freedom). This will require a fulsome examination of the HEI research ecosystem, with a view to understanding the diverse ways HEIs engage with and respond to geopolitical risk, requiring them to (Shih, Chubb & Cooney-O'Donoghue, 2023: 15):

> …develop guidelines that consider the increasingly multipolar research landscape amid geopolitical tensions. The research sector's inability to handle matters related to data security, multiple affiliations, or ethics dumping can mean that national political forces are likely to use additional compliance.

Another potential area for governmental intervention with HEIs is through the judicious use of funding mechanisms (Shih, 2024). HEIs are – for the most part – publicly funded institutions with a mandate for research and education in the public good. They are not traditionally meant to operate as business entities (though many of them do). Where HEIs contract with a secured cloud provider there will be a significant cost imposition and transitioning to another provider is likely to be cost prohibitive. Therefore, governments could help defray the costs of securing AI/cloud environments for HEIs as a way of sharing or transferring national security risks implicit in sensitive or controlled research. The Canadian government for example already supplies funding to HEIs to institute research security, where the funds are a *pro rata* amount of their Federal funding (Wilner et al. 2022).

Further, the impact of insider risk in both HEI settings and secured clouds needs to be better understood. As outlined above, policy can always help "bridge the gap" but when HEIs and governments both have responsibilities and obligations in any space, those same responsibilities and obligations will inevitably overlap and cause friction. Broad obligations – such as the Australian PSPF – can help establish a baseline of compliance for affected HEIs, but what is really required in these spaces is guidance around implementation of such policies. One clear example in the research security space was the implementation guidance issued following President Donald Trump's enactment of National Presidential Security Memorandum No. 33 on research security (Corn, 2021).

Lastly, there will need to be some coordination of AI and cloud security across international and transnational boundaries, with improved national and economic security a clear incentive for engaging on these cyber policy issues. Regulatory coordination drives mutual trust (Nayyar, 2023), which will continue to be a critical quality of the HEI research ecosystem. Not only can harmonised regulatory approaches benefit HEIs looking to get the best “bang for buck” from their public funding but also help buttress trust in cloud providers by making procurement decisions – often driven by compliance with regulatory standards – far easier (Nayyar, 2023: 12). Such approaches can only help the HEI sector in protecting itself, both within and without, from the geopolitical hazards with which is more frequently having to grapple.

References

ABC, ‘The Cloud: Kitty Flanagan’ (YouTube, 27 July 2018) Available at https://www.youtube.com/watch?v=CFdZWgiAj8I.

Adrian, A. (2013). How much privacy do clouds provide? An Australian perspective. *Computer Law & Security Review, 29*(1), 48-57.

Alhanahnah, M. J., Jhumka, A., & Alouneh, S. (2016). ‘A multidimension taxonomy of insider threats in cloud computing’. *The Computer Journal, 59*(11), 1612-1622.

Ali, I., Nagalingam, S., & Gurd, B. (2017). ‘Building resilience in SMEs of perishable product supply chains: enablers, barriers and risks’. *Production Planning & Control, 28*(15), 1236-1250.

Alzoubi, Y. I., Mishra, A., & Topcu, A. E. (2024). ‘Research trends in deep learning and machine learning for cloud computing security’. *The Artificial Intelligence Review, 57*(5), 132.

Amoore, L., & Raley, R. (2016). ‘Securing with algorithms: Knowledge, decision, sovereignty’. *Security Dialogue, 48*(1) 3-10, DOI: 10.1177/0967010616680

ASD. (2025). *"Bulletproof" hosting providers*. Australian Signals Directorate. Available at https://www.cyber.gov.au/about-us/view-all-content/publications/bulletproof-hosting-providers

Bergstrom, E. (2023). *To Spy the Lie: Detecting the Insider Threat of Espionage*. Master’s Thesis, Stockholm University. Available at https://www.diva-portal.org/smash/get/diva2:1784465/FULLTEXT01.pdf

Bongiovanni, I. (2019). ‘The least secure places in the universe? A systematic literature review on information security management in higher education’. *Computers & Security 86*, 350-357.

Bongiovanni, I., Karen Renaud, K., & George Cairns, G. (2020). ‘Securing intellectual capital: an exploratory study in Australian universities’. *Journal of Intellectual Capital 21*(3), 481-505.

British Council. (2024). *Managing risk and developing responsible Transnational education (TNE) partnerships*. Available at https://www.britishcouncil.org/sites/default/files/uuki_bc_risk_in_tne_report.pdf.

Carroll, M., van der Merwe, A., & Kotze, P. (2011) ‘Secure Cloud Computing: Benefits, Risks and Controls’. *Proceedings of the Information Security Association South Africa*, DOI: 10.1109/issa.2011.6027519.

Changchit, C., & Chuchuen, C. (2018). ‘Cloud computing: An examination of factors impacting users’ adoption’. *Journal of Computer Information Systems, 58*(1) 1-9, DOI: 10.1080/08874417.2016.1180651

Coetzee, W. S., Berndtsson, S. J. (2023). ‘Understanding Sweden’s security economy’. *Defense & Security Analysis 39*(2), 171-190, DOI: 10.1080/14751798.2023.2182479, 171.

Cohen, J. E. (2019). ‘Turning privacy inside out’. *Theoretical Inquiries in Law, 20*(1), 1-32.

Corn, G. P. (2021). ‘National security decision-making in the age of technology: Delivering outcomes on time and on target’. *Journal of National Security Law & Policy, 12*, 61-70.

Cotton, A. C., Viñas-Racionero, M. R., & Scalora, M. J. (2024). ‘The threat within versus the threat beyond: An examination of the differences between insider and outsider threats to college campuses'. *Journal of Threat Assessment and Management.* https://doi.org/10.1037/tam0000234

Council of the European Union. (2024, May 23). COUNCIL RECOMMENDATION on enhancing research security, OR en. 9097/1/24/REV 1. Available at https://data.consilium.europa.eu/doc/document/ST-9097-2024-REV-1/en/pdf

Croft, D. (2025, January 23). 'Trump axes Cyber Safety Review Board members'. *CyberDaily.EU*. Available at https://www.cyberdaily.au/government/11625-trump-axes-cyber-security-review-board-members

CyberCX (2024, August 1). 'CyberCX partners with Curtin University to launch sovereign cloud platform for sensitive research'. Media release. Available at https://cybercx.com.au/news/nebula-launch/.

Dasgupta, D., Akhtar, Z., & Sen, S. (2022). 'Machine learning in cybersecurity: a comprehensive survey'. *The Journal of Defense Modeling and Simulation, 19*(1), 57-106.

Department of Defence. (2024, July 4). 'Australian Government partners with Amazon Web Services to bolster national defence and security'. Media release. Available at https://www.minister.defence.gov.au/media-releases/2024-07-04/australian-government-partners-amazon-web-services-bolster-national-defence-and-security.

Department of Home Affairs. (2025, March). 'About the PSPF'. Available at https://www.protectivesecurity.gov.au/about

Department of Homeland Security. (2024, April 2). 'Cyber Safety Review Board Releases Report on Microsoft Online Exchange Incident from Summer 2023'. Media release. Available at https://www.dhs.gov/news/2024/04/02/cyber-safety-review-board-releases-report-microsoft-online-exchange-incident-summer.

El-Gazzar, R., Hustad, E., Olsen, D. H. (2016). 'Understanding cloud computing adoption issues: A Delphi study approach'. *Journal of Systems and Software, 118*, 64-84, DOI: 10.1016/j.jss.2016.04.061

Erdmann, A., & Toro-Dupouy, L. (2025). 'The influence of the institutional environment on AI adoption in universities: identifying value drivers and necessary conditions'. *European Journal of Innovation Management*. DOI: 10.1108/EJIM-04-2024-0407

Evroc. (2024). 'Sweden-Based Evroc to Build Europe's First Sovereign Hyperscale Cloud'. Available at https://www.datacenter-forum.com/evroc/sweden-based-evroc-to-build-europes-first-sovereign-hyperscale-cloud.

Falk, C. D. (2022). *Cyber Supply Chain Security and the Swedish Security Protected Procurement with Security Protective Agreement* (Masters' thesis, Stockholm University) Available at https://www.diva-portal.org/smash/get/diva2:1784357/FULLTEXT01.pdf.

Franceschi-Bicchierai, L. (2024, July 10). Microsoft emails that warned customers of Russian hacks criticized for looking like spam and phishing. *TechCrunch*. https://techcrunch.com/2024/07/10/microsoft-emails-that-warned-customers-of-russian-hacks-criticized-for-looking-like-spam-and-phishing/

French, R. (2019). *Report of the Independent Review of Freedom of Speech in Australian Higher Education Providers.* Available at https://www.education.gov.au/higher-education-publications/resources/report-independent-review-freedom-speech-australian-higher-education-providers-march-2019.

Gearon, L. F. (2019). ‘The University-Security-Intelligence Nexus: Four domains’, in Liam Francis Gearon (Ed.), *The Routledge International Handbook of Universities, Security and Intelligence Studies.* Routledge, London, 7-77.

Ghosh, S., Shetty, M., Bansal, C., & Nath, S. (2022, November). How to fight production incidents? An empirical study on a large-scale cloud service. *Proceedings of the 13th Symposium on Cloud Computing,* 126-141.

Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R. A., Chartash, D. (2022). ‘How does ChatGPT perform on the medical licensing exams? The implications of large language models for medical education and knowledge assessment’ *MedRxiv*, DOI: 10.1101/2022.12.23.22283901.

Gothenberg, A. (2022). *Recommendations to higher education institutions on how to work with responsible internationalisation.* STINT, Stockholm. Available at https://www.stint.se/wp-content/uploads/2022/09/STINT_Ansvarsfull-internation_web.pdf.

Gotor, F. R. (2023). ‘Hydro-powered sovereign cloud protects Sweden’s industrial heartland’. Available at https://www.tietoevry.com/en/success-stories/2023/hydro-powered-sovereign-cloud-protects-swedens-industrial-heartland/.

Gross Methner, S.E. (2019). ‘A Catholic University Approach to Campus Speech: Using Constitutional Academic Freedom to Hold the Tension of Free Speech, Inclusive Diversity, and University Identity’, *University of St. Thomas Law Journal 15*(2), 358-418.

Group of Eight Australia. (2022). *Essential decisions for national success: Reducing the regulatory overload on our universities*. Available at https://go8.edu.au/wp-content/uploads/2022/04/Go8-Reducing-the-regulatory-overload.pdf.

Gualandris, J., & Kalchschmidt, M. (2014). ‘Mitigating the effect of risk conditions on supply disruptions: the role of manufacturing postponement enablers’. *Production Planning & Control: Management of Operations, 26*(8), 637-653.

Gualandris, J., & Kalchschmidt, M. (2015). ‘Supply risk management and competitive advantage: a misfit model’. *The International Journal of Logistics Management, 26*(3), 459-478.

Haranas, M. (2024, February 2). Cloud Market-Share Q4 2023 Results: AWS Falls As Microsoft Grows. *CRN.* https://www.crn.com/news/cloud/2024/cloud-market-share-q4-2023-results-aws-falls-as-microsoft-grows?page=1

Ho, S. M., & Velázquez, M. O. (2015). ‘Do you trust the cloud? Modeling cloud technology adoption in organizations’. Paper presented to the Twenty-first Americas Conference on Information Systems, Puerto Rico.

Horton, A. (2024, 28 October). Mitigating Australia’s cloud-computing risks is still work in progress. *The Strategist*. https://www.aspistrategist.org.au/mitigating-australias-cloud-computing-risks-is-still-work-in-progress/

HRK: German Rectors’ Conference. (2020). *Resolution of the Executive Board on 6 April 2020: Guidelines and standards in international university cooperation*. Available at https://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-01-Beschluesse/Beschluss_Leitlinien_und_Standards_HRK_Praesidium_6.4.2020_EN.pdf.

Issaoui, A., Örtensjö, J., & Islam, M. S. (2023). ‘Exploring the General Data Protection Regulation (GDPR) compliance in cloud services: insights from Swedish public organizations on privacy compliance’. *Future Business Journal, 9,* 107, DOI: 10.1186/s43093-023-00285-2.

Kandias, M., Virvilis, N., & Gritzalis, D. (2011). 'The insider threat in cloud computing'. *International Workshop on Critical Information Infrastructures Security* (pp. 93-103). Springer: Heidelberg.

Karran, T. (2007). 'Academic Freedom in Europe: A Preliminary Comparative Analysis'. *Higher Education Policy, 20*, 289-313.

Kasongo, S. M. (2023). 'A deep learning technique for intrusion detection system using a recurrent neural networks based framework'. *Computer Communications, 199*, 113–125

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). 'Gpt-4 passes the bar exam'. *Philosophical Transactions of the Royal Society, A382*(2270), 20230254.

Kendall, S. (2024). 'Espionage law in the UK and Australia: Balancing effectiveness and appropriateness'. *The Cambridge Law Journal, 83*(1), 62-98.

Kinuthia, N., & Chung, S. (2017). 'An empirical study of technological factors affecting cloud enterprise resource planning systems adoption'. *Information Resources Management Journal 30*(2) 1-22, DOI:10.4018/IRMJ.2017040101.

Krien, A. (2024, June 15). 'Universities accused of using national security as "an alibi" over weapons ties'. *The Saturday Paper*. Available at https://www.thesaturdaypaper.com.au/news/education/2024/06/15/how-transparent-are-university-disclosures-weapons-ties.

Laux, J., Wachter, S., & Mittelstadt, B. (2024). 'Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk'. *Regulation & Governance, 18*(1), 3-32.

Lilyanova, V. (2022). 'Digital public services in the National Recovery and Resilience Plans', European Parliament Research Service. PE 739.271. Available at https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/739271/EPRS_BRI(2022)739271_EN.pdf.

Lundin, P., Stenlås, N., & Gribbe, J. (2010). *Science for Welfare and Warfare: Technology and State Initiative in Cold War Sweden*. Science History Publications.

Mell, P., & Timothy Grance, T. (2011). *The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology* (US Department of Commerce, Special Publication 800-145).

Ministry of Higher Education and Science (2022). *Committee on guidelines for international research and innovation cooperation.* Available at https://ufm.dk/publikationer/2022/afrapportering-udvalg-om-retningslinjer-for-internationalt-forsknings-og-innovationssamarbejde.

Mishra, S., & Tyagi, A. K. (2022). 'Emerging trends and techniques in machine learning and Internet of things based cloud applications'. In: Tyagi, A. K., & Sreenath, N. (Eds) *Handbook of research of internet of things and cyber-physical systems*, pp 149-167. Apple Academic Press: CRC Press.

Mitchell, A. D., & Samlidis, T. (2021). 'Cloud services and government digital sovereignty in Australia and beyond'. *International Journal of Law and Information Technology, 29*(4), 364-394.

Mittelstadt, B. (2017). 'From individual to group privacy in big data analytics'. *Philosophy & Technology, 30*(4), 475-494.

Myklebust, J. P. (2023, May 20). 'Academics hit back over interference in university boards', *University World News*. Available at https://www.universityworldnews.com/post.php?story=20230519150323626.

Nayyar, R. (2023). *The Quad: Carved in Code. Shaping (Inter)national Security through Reshaping Economic Incentives*. GIGA Policy Brief. Available at https://www.giga-hamburg.de/assets/pure/44369008/GIGA_DigitalDiplomacyStatecraft_PB_02_Nayyar.pdf

Niemiec, M., Pappalardo, S. M., Bozhilova, M., Stoianov, N., Dziech, A., & Stiller, B. (2022, October). Multi-sector Risk Management Framework for Analysis Cybersecurity Challenges and Opportunities. In Andrzej Dziech, Wim Mees, Marcin Niemiec (Eds.), *International Conference on Multimedia Communications, Services and Security*. Cham: Springer International Publishing, pp. 49-65.

Parasol, M. (2018). 'The impact of China's 2016 Cyber Security Law on foreign technology firms, and on China's big data and Smart City dreams'. *Computer Law & Security Review, 34*(1), 67-98.

Pettit, T. J., Croxton, K. L., & Fiksel, J. (2013). 'Ensuring supply chain resilience: development and implementation of an assessment tool'. *Journal of Business Logistics, 34*(1), 46-76.

Polamarasetti, A. (2024). 'Role of Artificial Intelligence and Machine Learning to Enhancing Cloud Security'. *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC)*, Guntur, India, 2024, 1-6, doi: 10.1109/ICEC59683.2024.10837120.

Rakgoale, D. M., Kobo, H. I., Mapundu, Z. Z., & Khosa, T. N. (2024). 'A Review of AI/ML Algorithms for Security Enhancement in Cloud Computing with Emphasis on Artificial Neural Networks'. *2024 4th International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, Vanderbijlpark, South Africa, 329-336, doi: 10.1109/IMITEC60221.2024.10851076.

Ravichandran, P. K., & Santhosh Keerthi Balmuri, S. K. (2011). *Evaluating Different Cloud Environments and Services related to Swedish Armed Forces* (Masters of Science Thesis, Malmo University). Available at https://www.diva-portal.org/smash/get/diva2:1480432/FULLTEXT01.pdf.

Renzella, J., Cain, A., & Schneider, J.-G. (2022). 'Verifying student identity in oral assessments with deep speaker'. *Computers and Education: Artificial Intelligence*, *3*, 100044, DOI: 10.1016/j.caeai.2021.100044.

Rojszczak, M. (2020). 'CLOUD act agreements from an EU perspective'. *Computer Law & Security Review, 38*, 105442.

Saarinen, J. (2017, July 24). 'Sweden exposed sensitive data on citizens, military personnel', *IT News*. Available at https://www.itnews.com.au/news/sweden-exposed-sensitive-data-on-citizens-military-personnel-469046.

Santos, A., Martins, J., Pestana, P., Gonçalves, R., São Mamede, H., & Branco, F. (2024). 'Factors affecting cloud computing adoption in the education context-Systematic Literature Review'. *IEEE Access, 12*, 71641, DOI: 10.1109/ACCESS.2024.3400862.

SciVal. (2024). *Current collaborators 2021-2024*. Elsevier, available at https://www.scival.com/collaboration/.

Shahzad, F. (2014). 'State-of-the-art Survey on Cloud Computing Security Challenges, Approaches and Solutions'. *Procedia Computer Science, 37*, 357-362.

Shei, S., Christos Kalloniatis, C., Haralambos Mouratidis, H., & Delaney, A. (2016). 'Modelling secure cloud computing systems from a security requirements perspective', in Sokratis Katsikas, Costas Lambrinoudakis, Steven Furnell

(Eds.), *Trust, Privacy and Security in Digital Business: Proceedings of the 13th International Conference of TrustBus, Porto, Portugal, September 7-8, 2016*. Springer: 48-62.

Shih, T. (2023). ‘Research funders play an important role in fostering research integrity and responsible internationalization in a multipolar world’. *Accountability in Research*, DOI: 10.1080/08989621.2023.2165917.

Shih, T. (2024). ‘The role of research funders in providing directions for managing responsible internationalization and research security’. *Technological Forecasting and Social Change, 201*, 123253, DOI: 10.1016/j.techfore.2024.123253.

Shih, T., Chubb, A., & Cooney-O’Donoghue, D. (2023). ‘Scientific collaboration amid geopolitical tensions: a comparison of Sweden and Australia’. *Higher Education, 15*, DOI:10.1007/s10734-023-01066-0.

Spanaki, K., Gürgüç, Z., Mulligan, C., & Lupu, E. (2018). ‘Organizational cloud security and control: a proactive approach’. *Information Technology and People, 32*(3), 516-537.

Thomson, V. (2023, February 14). ‘Research universities play a vital role in protecting Australia’s national security’. Media release. Available at https://go8.edu.au/media-release-research-universities-play-a-vital-role-in-protecting-australias-national-security

Tovatt, C., Bergman, M., Braunerhielm, C., Ejsing, C., Hellberg, L., & Sundberg, K. (2024). *Academic freedom in Sweden. A governmental assignment on higher education institutions’ work with academic freedom*. Available at https://www.uka.se/download/18.427c7de418f38533f7357/1715751054520/Akademisk%20frihet%20i%20Sverige.pdf.

Ulven, J. B., & Wangen, G. (2021). ‘A systematic review of cybersecurity risks in higher education’. *Future Internet, 13*(2), 39.

Universities UK. (2022). *Managing risks in Internationalisation: Security related issues*. Available at https://www.universitiesuk.ac.uk/what-we-do/policy-and-research/publications/managing-risks-internationalisation.

Usman Ali, M., & Ayub, R. (2012). *Cloud Computing as a Tool to Secure and Manage Information Flow in Swedish Armed Forces Networks* (Masters of Electrical Engineering Thesis, Blekinge Institute of Technology). Available at https://www.diva-portal.org/smash/get/diva2:833566/FULLTEXT01.pdf.

Wagner, Caroline S. (2018). *The Collaborative Era in Science: Governing the Network.* Verlag: Springer.

Walker-Munro, B. (2024a). ‘A Duty to Protect from Science? Interactions in International Law between Research Security and the Right to Science’. *Security Challenges*, available at https://regionalsecurity.org.au/article/a-duty-to-protect-from-science-interactions-in-international-law-between-research-security-and-the-right-to-science/.

Walker-Munro, B. (2024b). ‘Why isn’t Australia securing its critical research?’. *EduResearch Matters*, available at https://blog.aare.edu.au/why-isnt-australia-securing-its-critical-research/.

Walker-Munro, B. (2024c). ‘A student’s visa has been cancelled for links to ‘weapons of mass destruction’. What’s going on with Australian research security?’. *The Conversation*, available at https://theconversation.com/a-students-visa-has-been-cancelled-for-links-to-weapons-of-mass-destruction-whats-going-on-with-australian-research-security-230002.

Wilner, A., Beach-Vaive, S., Carbonneau, C., Hopkins, G., & Leblanc, F. (2022). ‘Research at risk: Global challenges, international perspectives, and Canadian solutions’. *International Journal, 77*(1), 26-50.

World Bank Group. (2024). *Military expenditure (% of GDP).* Available at https://data.worldbank.org/indicator/MS.MIL.XPND.GD.ZS.

Young, A. (2023). *Pandora’s Gamble: Lab Leaks, Pandemics, and a World at Risk.* New York: Center Street.

Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*, 2, 100025.

Zuo, Y., & Hu, W.-C. (2009). ‘Trust-based information risk management in a supply chain network’. *International Journal of Information Systems and Supply Chain Management, 2*(3), 19-34.

A relational view on Artificial Intelligence business value: A qualitative meta-analysis

Panagiotis Keramidis
Copenhagen Business School, Frederiksberg, Denmark, pke.digi@cbs.dk

Completed Research

Abstract

The business value of Artificial Intelligence (AI) is a prominent topic in Information Systems (IS) literature. As our knowledge around it becomes more nuanced, the intricacies of the relational aspects and their effects to value creation and capture become observable, particularly in interorganizational settings. This study sheds light on these aspects, by examining the factors that lead to value creation in partnerships around AI, but also the factors that impede value creation and capture. It follows a qualitative meta-analysis approach, drawing from the relational view on value creation. The study is founded on 20 empirical studies on AI business value and identifies the relational factors that are discussed as prominent when it comes to value creation and capture. The study informs both IS research and practice, by pointing to relevant factors that are influential, but also to factors not extensively discussed yet, thus providing a research agenda for future research.

Keywords: Artificial Intelligence, business value, value creation, value capture, value cocreation, relational view, AI business value, relational value

1.0 Introduction

The business value of Artificial Intelligence (AI) is one of the very prominent points of discussion in the current industrial setting. Enholm et al. (2022) conceptualize AI business value as the organizational impacts of AI usage on a process (e.g., improved efficiency, insights generation) and firm level (e.g., financial performance, customer satisfaction). Many organizations wonder how to benefit from AI investments, while many of them are partnering to cocreate value through AI (Enholm et al., 2022; Jacobides et al., 2021). Yet, typical with Information Systems (IS) investments, cocreating value is not an effortless endeavour, since it requires substantial alignment of both operations and strategic priorities to avoid conflicts (Trang et al., 2022). That makes the need to investigate the partnerships around AI value cocreation imperative.

Literature has discussed the topic of interorganizational partnerships of AI to a certain extent. Some studies have discussed the business ecosystems around AI value creation and capture, highlighting the relevance of diverse partnerships for that objective (Burström et al., 2021; Jacobides et al., 2021). Another studies have underlined the relevance of relational conditions for AI business value manifestation in interorganizational partnerships (Enholm et al., 2022; Keramidis & Shollo, 2022; Shollo & Vassilakopoulou, 2024). Yet, the relational factors that affect the value creation and capture in the context of partnerships around AI have not been specifically identified. This creates a challenge from both research and practice standpoint.

Research-wise, it should be noted that due to the complexity of the implementation of AI investments, interorganisational partnerships become fruitful avenues for value creation and capture (Wamba-Taguimdje et al., 2020). In fact, there are certain properties of recent advances in AI that make partnerships especially relevant. For example, Feuerriegel et al. (2024) note that generative AI requires a considerable amount of data, which can only be provided by a few big corporations. Yet, our understanding of the interactions in such settings, where one organization is dependent on another to gain access to the necessary for the generative AI application data, is limited (ibid.). By not capturing the knowledge related to the dynamics and interactions in partnerships, literature misses substantial knowledge regarding the AI value creation and capture (Enholm et al., 2022). On the practice side, industry stakeholders do not have a clear picture and thus guidance on the aspects that need to be considered when it comes to engaging in a successful partnership for AI value creation and capture.

This study addresses this need, answering the research question: "*What are the relational factors that affect the AI business value creation and capture in interorganizational partnerships?*". To answer this research question, this study follows a qualitative meta-analysis approach, drawing from the relational view on value creation and capture (Dyer & Singh, 1998; Dyer et al., 2018). While not as popular as their quantitative counterparts, qualitative meta-analyses have been employed in the context of a few prominent IS studies (Berente et al., 2022; Berente et al., 2019), since they constitute an appropriate method to synthesize knowledge from prior empirical research and provide new contributions (Berente et al., 2019). Considering the growing interest in the AI business value, the meta-analysis can re-interpret prior empirical findings to outline the relational factors associated with the value creation and capture, so that it can offer an alternative viewpoint that guides future studies. Specifically, this study draws conclusions based on empirical findings from 20 prior empirical studies and identifies the prominent relational factors for value cocreation in partnerships around AI, as they were mentioned in these studies.

This study contributes to IS research and practice by investigating the relational conditions that affect the creation and capture of AI business value in interorganizational settings and by specifying which of the factors elaborated in the relational view are more vividly discussed in the context of AI. The study also offers a research agenda regarding the existing gaps in the AI business value literature, extending the discussion on the relational conditions of AI business value. Further, it informs practitioners who wish to learn more about the strategic points of interest when it comes to partnering to cocreate value through AI.

The remainder of the paper is structured as follows: section 2 provides the theoretical background, presenting the status quo on AI business value research and the relational view. In Section 3, the methodological steps for the meta-analysis are explained. Section 4 presents the findings. In section 5, the findings are discussed and there is a research agenda offered. Section 6 concludes the study.

2.0 Background

This section provides the background of the study. It starts by providing the status quo on the AI business value research and it proceeds to elaborate on the relational view, which constitutes the theoretical viewpoint of this paper.

2.1 Artificial Intelligence Business Value

The business value of AI is a vividly discussed topic in IS literature, particularly in recent years, and it has been conceptually devised in different ways. Enholm et al. (2022) conceptualise the business value of AI in terms of first order (efficiency increase, insights generation and business process transformation) and second order effects (market, financial, operational and sustainability performance improvement). Borges et al. (2021) point to four different AI value creation sources, namely decision support, customer and employee engagement, automation and innovation in terms of products and services. Similarly, Collins et al (2021) have reported business value types of process automation, cognitive insight and cognitive engagement. Thus, we can assert that the business value of AI is a multifaceted concept, encompassing many beneficial qualities and objectives associated with such technologies.

Literature also delineates ontologically related concepts. Mikalef and Gupta (2021) have outlined the resources necessary to create AI capabilities so that organizations can generate value. Shollo et al. (2022) have described the shifting nature of the mechanisms for creating value with AI, while Jöhnk et al. (2021) have outlined the organizational readiness factors for creating value with AI. There are also studies which offer a more nuanced understanding of the value of AI across settings. For instance, Wang et al. (2021) have illustrated the differences between the AI value perceived in public and private organizations. Another example is Ruokonen and Ritala's (2023) study, which describes the different strategic approaches organizations may operationalize when investing in AI. The aforementioned studies have contributed to achieving a more nuanced understanding of the business value of AI, since they have elucidated its conditions in various organisational settings and have offered actionable insights for researchers and practitioners to further improve their understanding.

What has also contributed to better understanding of the business value of AI is the stream of literature dedicated to IS business value (Kohli & Grover, 2008; Schryen, 2013), since many definitions, measurements and conceptualisations are applicable to many IS investments. Indeed, the IS business value can be considered as the progenitor of the concepts dedicated to the business value of individual IS technological investments, such as AI. Yet, there are some peculiarities when it comes to AI investments, which are considered unlike other IS investments (Berente et al., 2021). Their agency and self-learning capabilities challenge existing assumptions (Baird & Maruping, 2021; Zhang et al., 2021). As a result, literature has been closely investigating the business value of AI investments, and not treating them as a typical IS investment.

2.2 The Relational View

The relational view supports that part of a firm's critical resources may span the boundaries of the firm and are founded in interorganizational collaborations (Dyer & Singh, 1998; Dyer et al., 2018). Originally pointing to relation-specific assets, knowledge-sharing, complementary resources and effective governance (Dyer & Singh, 1998), it substantially assisted the theorization of value cocreation. More recently, it is revisited, considering the dynamic nature of partnerships, and points to aspects related to the evolution of the partnership but also to the competition among partners (Dyer et al., 2018). Dyer et al. (2018, p. 3141) define value creation as the value created in an alliance beyond what the parties would achieve in traditional market relationships, and value capture as the percentage of the value appropriated by each party.

The relational view has informed several studies in the context of IS business value. Grover and Kohli (2012) have elaborated on IT capabilities for four layers of relational arrangements. Rai et al. (2012) have extended the knowledge on capabilities and elaborated on the role of communications in the relational view of IT value cocreation. More recently, Mandrella et al. (2020) have provided a meta-analysis on the interorganizational IT capabilities for value cocreation, informing about the effects and value moderators related to the interorganizational IT investments. These are just a few exemplary studies, since IS literature has drawn from the relational view to study more phenomena, such as digital transformation (Malik et al., 2024) and IT governance (Prasad et al., 2013).

The relationships among organizational entities have been noted as an important aspect of investigation when it comes to AI business value (Enholm et al., 2022). Indeed, recent empirical evidence showcases the relevance of relational conditions when it comes to AI business value (Keramidis & Shollo, 2024; Shollo & Vassilakopoulou, 2024). However, literature has not yet adopted a relational view to investigate AI business value creation and capture. Yet, innovative IS initiatives, such as AI, are often too complex to be provided by one organisation, and thus partnerships and alliances are fruitful ways of creating and capturing value (Wamba-Taguimdje et al., 2020). Such an observation has been recently discussed in the context of prominent AI technologies and their constitutive properties. For example, generative AI requires a considerable amount of data, which are often available only through big corporations (Feuerriegel et al., 2024). The relationships with such corporations that offer the data have not been investigated, and thus we cannot understand the economics of generative AI for organizations implementing such solutions (ibid.). Shedding light on the interactions and interdependencies when it comes to partnerships for AI value creation and capture, would thus offer valuable insights for both researchers investigating the concept of AI business value and practitioners actively pursuing it. This study draws from the relational view and fulfils this need.

3.0 Method

The study utilizes the qualitative meta-analysis method (Berente et al., 2019) to reinterpret some contextual details that emerge across the studies included (Berente et al., 2022; Noblit & Hare, 1988). These details relate to the partnerships formed around AI. A qualitative meta-analysis can result in meta-interpretations based on prior qualitative studies results and draw new inferences (Berente et al., 2019). In this case, prior qualitative studies' empirical data were used to draw inferences about the partnership conditions around AI systems. Figure 1 offers a visual presentation of the method.

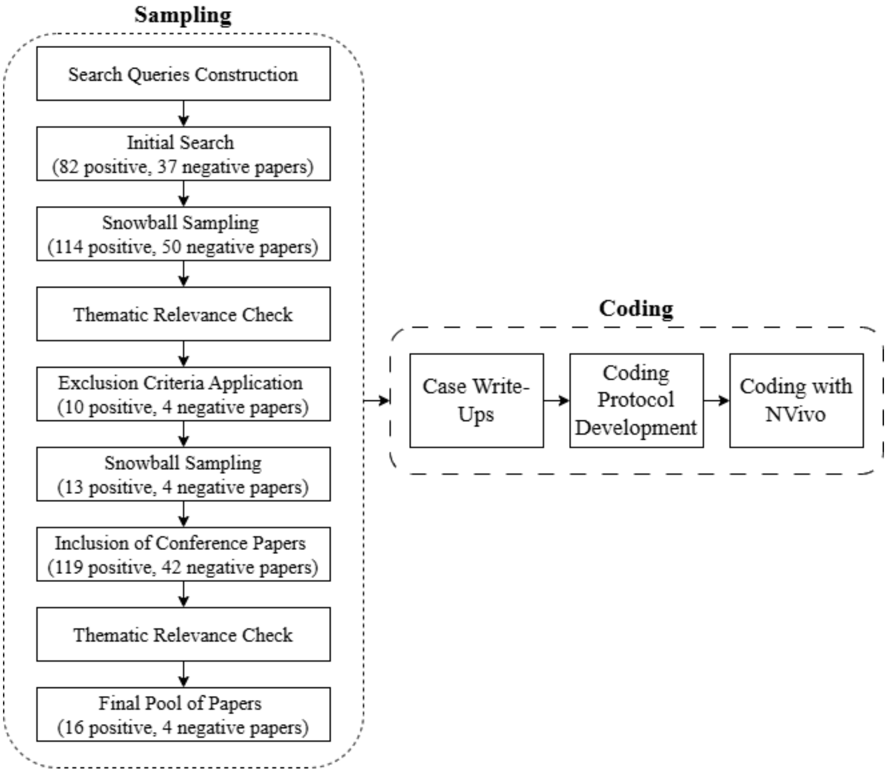

Sampling
Search Queries Construction
Initial Search
(82 positive, 37 negative papers)
Snowball Sampling
(114 positive, 50 negative papers)
Thematic Relevance Check
Exclusion Criteria Application
(10 positive, 4 negative papers)
Snowball Sampling
(13 positive, 4 negative papers)
Inclusion of Conference Papers
(119 positive, 42 negative papers)
Thematic Relevance Check
Final Pool of Papers
(16 positive, 4 negative papers)
Coding
Case Write-Ups
Coding Protocol Development
Coding with NVivo

Figure 1. The method followed.

3.1 Sampling

While a literature review focuses on synthesizing prior studies' knowledge and views, a qualitative meta-analysis reinterprets their findings regardless of the theoretical lenses used in each study (Berente et al., 2019). Consequently, the literature was systematically searched for qualitative studies that can inform the purposes of this study but used primarily the empirical material presented in them.

Since the objective was to investigate the relational determinants related to the business value of AI, there was a need for studies that examine the organizational adoption of AI, and the value generated (either positive or negative). Thus, there were two different search queries constructed, one for the positive instantiations of the business value of AI and one for the unintended and negative ones. The queries were inspired by Enholm et al. (2022) and their own search query, since this is a recent yet comprehensive literature review on the business value of AI. This study's research query, which was used in Scopus, can be found in Appendix 1.

Considering the pervasiveness of AI in not only IS research, but also in other neighbouring domains, it was decided to not limit the search to IS outlets, but also include prominent outlets from Strategy, Organization Studies and Management, since they can also be domains invested in analysing the business value of AI. It was also decided to include some executive outlets, since they are also frequently informative of organizational adoption cases related to AI and they are ample sources of empirical material. The full list of journals selected for this study's search can be found in Appendix 2.

The first search in Scopus resulted in 82 studies revolving around the positive instances of the AI business value and 37 studies for the negative ones (from 2013 until 2024). Snowball sampling was performed to expand the studies pool, to include relevant studies that were cited in these studies. Following Berente et al. (2019), only the cited papers that fit the initial query were included, while the duplicates were excluded. In the end, and after removing duplicates, there were 114 papers in the positive instantiations category and 50 for the negative one.

In the next stage the thematic relevance of the studies was checked. For that, the abstracts of the included papers were read. The exclusion criteria were similar to the ones followed in Berente et al. (2019). Specifically, the quantitative, conceptual and method papers were excluded. Further, the papers that did not focus on the organizational effects and the business value (either positive or negative) of AI were also excluded. These could be papers that briefly mention AI as an example but did not focus on it in the rest of the paper. It is worth noting that AI is a broad term, usually associated with technological approaches such as machine learning. Still, it is also discussed in the context of other terms, such as Big Data. These studies were included as well, since even if the artifact of analysis is not AI, but AI with another artifact, there are still some inferences that can be made about the strategic conditions that lead to value creation.

In addition, at this stage the second exclusion checking suggested by Berente et al. (2019) was performed, which was to exclude any papers that did not provide appropriate case descriptions and quotations. This resulted in 10 papers in the positive instantiations category and 4 in the negative ones. For the positive ones, there was once again a snowball sampling performed, as suggested by Berente et al. (2019), and there were 3 more studies gained. For the negative ones it was not necessary, since they were included in the first snowball sampling.

At this stage, the sampling was complemented with papers from relevant and reputable conferences. While this was not a step in the original methodology (Berente et al., 2019), it was decided to include conference papers for two reasons: first, because of the recent predominance of AI there might be insightful studies which have not yet reached publication but have been presented in conferences; second, since the number of qualitative studies in this topic is limited, any additional studies could benefit the analysis. Thus, the four largest IS conferences (ICIS, ECIS, AMCIS, PACIS) were included, since they often provide rigorous and rich in empirical material studies. The AIS library and the same two queries were used for all four of the conferences. The search provided 119 articles related to positive instantiations and 42 negative ones. Once again, the conference papers were checked for their thematic relevance and their provision of appropriate quotations. The conference papers were not submitted to snowball sampling, due to the oftentimes less comprehensive literature cited in the conference papers. The final articles added were 3 papers regarding positive instantiations.

The final pool of studies included were 16 studies for the positive instantiations of AI business value and 4 for the negative or unexpected ones. Table 1 presents the final set of studies.

Positive	
Engel et al., 2024	Jöhnk et al., 2021
Hopf et al., 2023a	Zhang et al., 2021
Badakhshan et al., 2022	Hopf et al., 2023b
Shollo et al., 2022	Magistretti et al., 2019
Huang et al., 2022	Ranjan & Foropon, 2021
Aversa et al., 2020	Someh et al., 2020
Wamba-Taguimdje et al., 2020	Keller et al., 2019
Shollo & Vassilakopoulou, 2024	Keramidis & Shollo, 2024
Negative	
Mayer et al., 2020	Allen & Choudhury, 2022
Cheng et al., 2022	Liang & Xue, 2022

Table 1. Studies included.

3.2 Coding

For the next phase, the papers were read and there were case write-ups written for each paper – which can be useful for interpreting the empirical material found in the studies (Berente et al., 2019). The case write-ups were brief (usually one page) descriptions of the empirical outcomes of each paper. They included three types of input: meta-information about the studies (such as the methods utilized and the research questions answered), information about the cases (such as the cases' industry, country and technologies) and some relational details about the partnerships found in the papers. The last one pertains to the theoretical scope of this study, and it follows the protocol which was also used for the coding in the next step.

The coding protocol emerged from theory (Berente et al., 2019). Specifically, the analysis draws from the relational view (Dyer & Singh, 1998; Dyer et al., 2018) to identify partnership details that affect the creation and capture of the AI business value. Dyer and colleagues' (2018) factors were used as literature-informed up-front codes (Boyatzis, 1998). Thus, the coding protocol was founded on three categories of codes: first, factors leading to initial value creation; second, factors leading to diminished value creation; third, factors leading to competition for value capture among partners. The subcodes followed faithfully the categories in each factor found in Dyer et al. (2018, p. 3143). The coding protocol, which was inspired by Kude & Huber (2024), with indicative codes and quotes can be found in Appendix 3.

NVivo was used for the coding of the studies. During the coding, the focus was on the ways in which the relational factors are discussed in the context of creating value through AI. Not all the factors enjoyed the same discussion in the studies. Still, there were interesting observations regarding the business value of AI found, reinterpreted in the context of partnerships.

4.0 Findings

The structure of the findings follows the three relational factors related to value creation and capture provided by Dyer et al. (2018, p. 3143). Accordingly, the factors that lead to initial value creation and capture in partnerships revolving around AI, the factors that lead to diminished value creation and the factors that lead to competition for value capture are presented.

4.1 Factors that lead to initial value creation

The factors leading to successful value creation in partnerships revolving around AI were notably discussed among the studies, and the complementary resources that these partnerships entail were the linchpin of these discussions. These complementary resources often included one company providing the algorithms while another one provided the data for the AI implementation. For example, a credit rating company providing an ML-based technology used other companies to acquire the data necessary for their services: “*With the expansion of our lending business, in particular, by developing different partnership arrangements with other companies, we are able to acquire a lot of new user data, which was not feasible before… I suppose the diversity does give us a lot more to play with.*” (Huang et al., 2022, p. 295). Another instance relates to a company partnering with an another, due to the capabilities of the latter’s technology (compared to other technologies): “*[…] if we automate with the imperative models, […] we have found […] there's a big part of the knowledge of the technical worker we didn't get that much advantage from. […] And that's why we decided to partner with them.*” (Keramidis & Shollo, 2024, p. 9).

Yet the complementary resources for the AI value creation in partnerships were not always referring to digital technologies. That was the case with an energy company providing the expertise necessary for an AI system to classify petroleum rock samples correctly: “*Two different geologists with different levels of experience will provide different levels of accuracy in their rock descriptions. With IBM Watson, we are ensuring that the description and interpretation is always at the expert level and that it will remain consistent throughout the years.*” (Wamba-Taguimdje et al., 2020, p. 1907). Another example can be spotted in the collaboration between an AI provider and a bank to provide loan services, where the former provides the AI loan processing system, and the latter provides the administrative operation. This arrangement was advantageous for both parties, as a manager of the AI provider states: “*We used to have [CleverLoan] shops. And that didn’t work out as well as we had imagined, and then consequently ... we withdrew from the area and left the value we could add completely to [Main Finance] with its branches and customer network.*” (Mayer et al., 2020, 242). Thus, we can observe that both the digital resources and the expertise and operations can be combined to form partnerships around AI value creation.

Still, this combination also requires appropriate knowledge sharing in order to successfully come to fruition. This necessity stems from the artefact peculiarity, which contra to traditional IT projects requires more domain knowledge: “*Business units know how classical IT projects are run: The IT department gets the requirements and then iteratively implements them. But data science projects require much more interaction between domain experts and Data Scientists, and it is important that the business knows and understands this*” (Hopf et al., 2023a, p. 32). Another reason for this need for knowledge sharing relates to the novelty of AI technologies, which makes exchanging ideas valuable: “*Generally, our expectations* *of the overall project originally were that […] we get those formalities, algorithms, thought-provoking impulses and maybe new ideas from the university, from the scientific perspective, and of course that we discuss this with the project partners*" (Keller et al., 2019, p. 9).

Knowledge sharing for AI is important also considering the business effects of its rarity. Indeed, knowledge sharing can decrease the dependence on experts within partnerships: "*We would like to talk to some specialists, but we would like to do it and apply it ourselves. So yeah, again, we want to be independent and to have as much knowledge about the [Company] solutions as possible, so that we can design, we can develop ourselves.*" (Keramidis & Shollo, 2024, p. 10). Such instances also serve as reminders of the need for appropriate governance structures, which are also vividly discussed as factors for value creation.

Governance structures are one of the defining factors when it comes to a successful partnership, especially considering the risk in the applications that often AI is utilized. For example, a stakeholder applying AI in credit management mentioned: "*Part of our selection (of partners) criteria is whether we can be in the driving seat in defining the industrial standard for risk management. For instance, what are the key parameters for risk assessment; what is the level of acceptable risk in relation to product specs, capital flow and its management, pricing methodology for risk management service, and so on?*" (Huang et al., 2022, p. 297). Data are once again brought up as a complicating factor regarding governance, as one participant mentioned: "*It is quite difficult to govern the entire process of getting the data right, as data are scattered across different systems owned by different teams*" (Engel et al., 2024, p. 103).

Overall, complementary resources (either tied closely to the AI technology or to its operationalization), domain knowledge and use cases sharing, as well as governance structures (both for the models and the data) are discussed as factors leading to successful value creation.

4.2 Factors that lead to diminished value creation

Still, partnerships revolving around AI do not guarantee value creation in perpetuity. There are some problems noted in the context of these partnerships that can potentially threaten the value creation. One very prominent problem relates to decrease in resource complementarity, a case in which one partner is more dependent on the resources brought by this partnership than the other. For instance, using AI models provided by a third party can lead to dependencies: "*In the AI industry in general, companies don't even recognise how important and significant third-party dependence they are building (...) but those who understand, they also understand that we are actually creating incredibly important third-party dependencies.*" (Shollo & Vassilakopoulou, 2024, p. 5). These dependencies are especially relevant to the recently emerged generative AI applications (ibid.), since there are not many companies providing such models.

However, the dependencies do not relate only to the technology components (data, models, infrastructure etc.). As one participant notes: "*If [CleverLoan] were to switch off, there would be chaos, because most people can't do that anymore. ... So, there are drawbacks: knowledge is lost; it is definitely gone; it is out of the company; it has been outsourced to the AI system*" (Mayer et al., 2020, p. 249). This is of particular importance. Certainly, operational dependence and expertise loss are relevant problems to other IS investments as well; but in this case the dependence is even more complicated, since the knowledge is outsourced to the AI system. Which implies that besides the two partner companies there is also one parameter that should be considered, and that is the AI system itself.

The problems also stretch beyond the partnership boundaries, since competition can replicate the partnership resources. This is the reason companies are in constant need of innovation, but also use these innovations to enhance their existing service provision: "*but also the degree of integration in linking these existing and new services to help us differentiate from competitors.*" (Huang et al., 2022, p. 298). Thus, it is not only the AI products in themselves that are the focus of the innovations, but also the infrastructure that connects them meaningfully.

The final factor which can impede value creation relates to the obsolescence of the partnership resources due to the environmental dynamism. This is indeed an issue that requires the AI models to be constantly updated: "*First time if we have a really good project in the past and then we stopped looking at it and we add more stuff and more stuff and then this project that we have done might be obsolete, then we don't really know how to do it anymore. But if we keep updating it, like every time we have an idea, if we have an update, so this project doesn't become like just a useless and we have to do it again from scratch.*" (Shollo et al., 2022, Appendix p. 9). And the customers' perceptions on the value of the AI tools play a prominent role in determining their relevance. While a partnership may bring fruits monetarily for the partners, the clients may be displeased, such as in the case of collaborating with third parties for optimizing riding routes through AI, or like a customer noted: "*If other companies want to buy the user information accumulated on the ridesharing platform with a high price for, it may lead to information transactions*" (Cheng et al., 2022, Appendix p. 359).

Overall, dependencies on both partners and specifically the AI systems, issues of replication and market dynamism are noted as threats when it comes to AI value creation in partnerships.

4.3 Factors that lead to competition for value capture

It is often the case that partners compete to capture the value, and this is also the case of AI. Specifically, one partner can create additional value-adding resources, increasing its bargaining power over its partners. In the case of AI, since they require significant upfront investment, this makes sense business model-wise: "*Given that massive investment has already been put in to come up with a template, adding new variations based on the existing offering is relatively straightforward and cost effective…. You can expand the product range by having [a] different loan period, loan amount and payback flexibility to fulfil different demands. You can also add more contexts where the loan can be used*" (Huang et al., 2022, p. 297). Yet, such an expansion of capabilities may intensify existing dependencies, or in this case: "*I suppose the beauty of our product (cell phone purchase loan) is that our partners are very keen to push it. Without our product, there is no alternative means in the market for some of the transactions to take place….*" (Huang et al., 2022, p. 293). Having an increasingly large share of the loan services in a market may increase the bargaining power of the partner.

More broadly, the factors regarding competition for value capture among partners is admittedly the part of the conceptual framework that enjoys the lowest coverage in the studies. Besides the reproducibility of the models to other value-adding purposes, which increases the bargaining power of one partner over the others, the rest of the factors are not extensively discussed.

5.0 Discussion

The findings showcase that certain factors discussed by Dyer et al. (2018) are vividly discussed in the context of AI business value creation. Regarding the success of partnerships for AI value creation, the relevance of complementary resources can be clearly identified. More specifically, data are very frequently discussed as the resource bringing together two organizations in such settings, which is indeed an observation relevant to AI interorganizational collaborations (Enholm et al., 2022; Papagiannidis et al., 2021). However, the role of governance was also underlined, especially considering that there are different systems that need to cooperate. In both formal and informal structures, governance in the context of partnerships has been noted as an important component for IT value cocreation (Findikoglu et al., 2021; Mandrella et al., 2020) and AI is no exception (Enholm et al., 2022). Finally, knowledge-sharing is also a notable factor, which corroborates prior literature insights on AI value creation (Olan et al., 2022).

But AI partnerships' sustainability may suffer from certain developments, such as issues of competition, which pertain to both internal and external to the partnership actors. Internally, certain partners may have more bargaining power over others. Indeed, there are specific strategies partners may employ to consolidate bargaining power, for instance controlling data or algorithms necessary for partnerships (Ruokonen & Ritala, 2023), thus unilaterally affecting the partnership's development. Externally, the very environmental dynamism and competition, which probably pushed the partners to the partnership (Enholm et al., 2022), is also a threat to the partnerships, in terms of replication.

Put together, these construct the relational factors that correspond to AI value creation and capture. Table 2 presents these factors, constructing a relational account on AI business value. The factors and their respective categories are based on Dyer and colleagues' (2018) framework, transferred to the context of AI through the meta-analysis results.

Factors leading to initial value creation	**Factors leading to diminished value creation**	**Factors leading to competition for value capture**
Complementary resources - Data – Algorithms complementation - Expertise - Operation & Administration	***Decrease in complementarity*** - Third party model dependencies - Knowledge lost (to other companies or AI systems)	***Development of additional resources by one partner*** - Replication of models to other domains
Knowledge sharing - Domain knowledge - Use cases	***Replication of alliance resources*** - Models, data and infrastructure as replicable	
Governance - Model parameters control	***Environmental dynamism*** - Models' obsolescence	

- Data governance	- Customers’ concerns	

Table 2. The relational factors for AI value creation and capture.

5.1 Research Agenda

It is visible that not all factors found in the framework provided by Dyer et al. (2018) are discussed in the context of AI partnerships. This can be interpreted as a need for more studies in the relational conditions that affect the AI business value creation and capture, particularly in interorganisational settings. Of course, that does not mean that further investigation in the already discussed factors is not warranted.

Starting with the factors leading to value creation, while complementary resources, knowledge-sharing and governance structures were discussed, the studies did not focus on the relation-specific assets built within the partnerships. This is not surprising, since factors such as IT governance and knowledge sharing have a stronger relationship to business value than relation-specific assets (Mandrella et al., 2020), so it is reasonable for studies to focus more extensively to them. Still, investigating more these assets is relevant, because they can be considered one element that keeps the partnerships together (since having invested in relation-specific assets may make partners not want to move away from a partnership and lose their investments). To give a few examples from the AI studies examined, it would be interesting to see how a jointly created consulting framework for business process management (Keramidis & Shollo, 2024), or a petrography proof of concept tool (Wamba-Taguimdje et al., 2020) would affect the AI partnership. And since Dyer et al. (2018) stress the importance of considering the dynamicity of the partnerships, it would be especially relevant to observe these effects over time.

Regarding the issues that may lead to decreased value creation over time, broadly, factors external to the alliance are more extensively discussed. Indeed, environmental dynamism and replication by competitors are more frequently discussed than decrease in complementarity, while relational inertia is not even mentioned. Yet if internal to the alliance factors were also discussed, the understanding of the failures in partnerships around AI would be more complete. For instance, Shollo et al. (2022) mention that due to market dynamism the teams had to constantly update models to avoid obsolescence, yet there is not much information about how the teams internally coordinated in order to avoid inertia. Similarly, Shollo & Vassilakopoulou (2024) elaborate on the dependencies on third-party generative AI tools, and it would be interesting to know what the effects of this “excessive trust” (as Stevens et al. (2015)

called it) were. That way there could be governance principles established that help avoiding cases of relational inertia.

Finally, the factors regarding competition for value capture in partnerships around AI are the least explored. Certainly, dependencies and asymmetries in dependencies have been brought up in several of the studies (Huang et al., 2022; Keramidis & Shollo, 2024; Mayer et al., 2020; Shollo & Vassilakopoulou; 2024), but the ways in which they manifest in the partners' relationships could be further elucidated. For example, Mayer et al. (2020) focus on how the AI provider uses the bank's customer network and operations, but it could be the case that eventually the bank could build their own AI capabilities and replicate the AI provider's resources. Keller et al. (2019) mention the collaboration of different stakeholders and the knowledge exchanged, but it may be the case that some of the partners dedicate more knowledge than others, with this asymmetry making their bargaining position less attractive. Huang et al. (2022) elaborate on the market access the retailers provide to the credit management provider, but if a newcomer to the market were to imitate the provider's system, they could approach the retailers and deprive the provider of the access to the market. These are mere examples of competition for value capture that could emerge through the cases, to illustrate the relevance of this category of factors for future research.

Table 3 includes a research agenda based on the meta-analysis findings. Of course, the agenda is not exhaustive, rather is based on the less investigated factors of the framework provided by Dyer et al. (2018). Once again, that does not imply that the rest of the factors do not warrant further research.

Factors	**Research Questions**
Factors that lead to initial value creation	
Relation-specific assets	What are the relation-specific assets that are created in the context of partnerships around AI?
	How do the relation-specific assets that are created in the context of partnerships around AI affect the partnerships?
	How do the AI relation-specific assets evolve over time?
Factors that lead to diminished value creation	
Relational inertia	How does the increase in relational inertia affect partnerships around AI?
	How does excessive trust affect the AI value creation over time?
	What are the governance determinants that prevent relational inertia in partnerships around AI?
Factors that lead to competition for value capture	
Superiority of one partner at absorbing-replicating the partners' complementary	How does the replication of complementary AI resources achieved by a partner affect the partnership?
	What are the mechanisms partners pursue to replicate AI

resources	resources in a partnership?
Asymmetries in partner investments	How do investment asymmetries in AI partnerships affect the capture of value?
	In what types of AI partner investments do asymmetries occur more frequently?
Competitors imitate one partner’s alliance-specific complementary resources	What complementary resources involved in AI value cocreation are most susceptible to replication by competitors?
	What are the mechanisms for protecting AI complementary partnership resources from being replicated by competitors?

Table 3. Research agenda for the relational factors in the context of AI value creation and capture.

5.2 Contributions

This study contributes to the stream of research dedicated to the business value of AI. Enholm et al. (2022) underline the relevance of interorganizational partnerships in the context of AI business value, Mikalef and Gupta (2021) call for more studies on the tensions involved among the different stakeholders involved in AI value creation, while Keramidis and Shollo (2024) invite more research on the relational determinants on AI business value creation in interorganizational settings. This study answers this call by offering a qualitative meta-analysis on the existing empirical studies on the business value of AI, and its positive, negative and unexpected manifestations. The findings answer the broader call for IS research in the relationships developing in the context of AI (Ågerfalk et al., 2022), targeting the relationships developed among organizations collaborating for value cocreation.

The findings point to specific relational conditions and factors that emerge from the reinterpretations of the current literature findings, highlighting their relevance for future studies to consider when studying the strategic initiatives around AI. The findings also point to specific factors that have not yet been discussed, but can potentially be influential for value cocreation, according to the relational view (Dyer & Singh, 1998; Dyer et al., 2018). Indeed, this study is attempting to bring the relational view into the forefront of IS research initiatives, along with other studies that focus on IS business value (Malik et al., 2024; Mandrella et al., 2020; Rai et al., 2012), while this study focuses on AI.

On the practice side, the study offers a comprehensive depiction of the influential relational factors related to AI business value creation, which is grounded on prior empirical knowledge. Indeed, through our findings, practitioners can have an indication on what relational components to pay attention to when it comes to partnering for AI value creation in order to succeed in value creation and capture over time. They can be summarized into complementary resources (of various sorts, refer to Table 2), knowledge-sharing (expertise and use cases), governance (data and model), and avoidance of dependence and replication of the aforementioned resources by both competitors and partners.

5.3 Limitations

This study has certain limitations. Typical with qualitative meta-analyses, the results of this study are constrained by the included studies' empirical material to draw interpretations (Berente et al., 2019). This points to the triple hermeneutic problem (Giddens, 1984), where the study relies on snippets of interpretations of the researchers, who rely on interpretations of their studies' participants. Inevitably, some of the contextual details are lost in the different layers of interpretation.

Further, by embracing AI and all the related technologies as terms for the keyword search, the study fails to address specific technological peculiarities and rather present the business value conditions for the broader technological area of AI. While a more specific analysis on AI technologies and their business value would be more effective, the study had to be rather inclusive with the technological approaches included, since the number of qualitative studies in the topic is relatively low. Future studies can examine more specific AI technologies.

Finally, it is important to note that this study does not focus specifically on one type of partnership around AI. The studies included present partnership models such as sourcing, consulting, and affiliate partnerships, but the study does not focus on the peculiarities of each partnership regarding the value creation and capture of AI. While this is certainly a limitation, it should be noted that the qualitative empirical studies are not numerous enough to warrant such comparative approaches on the partnership models, so it would not be feasible for the scope of this study to address this. Future studies can differentiate between partnership modes regarding AI value creation.

6.0 Conclusion

The business value of AI is a relevant research topic and recently there is particular emphasis put on its interorganizational manifestations and their relational determinants. In this context, the study follows a qualitative meta-analysis on the relational factors that affect the creation and capture of AI business value in interorganizational partnerships. It draws from the relational view, and it presents certain factors that are relevant in such partnerships. The study extends the discourse on AI business value and offers an agenda for future research.

Acknowledgements

This study is part of the DIREC project (grant number 9142-00001B Digital Research Centre – DIREC – P09B - Business Transformation and Organisational AI-based Decision Making) and is supported by Innovation Fund Denmark and DCR Solutions.

References

Ågerfalk, P. J., Conboy, K., Crowston, K., Eriksson Lundström, J., Jarvenpaa, S. L., Ram, S., and Mikalef, P. (2022) *Artificial intelligence in information systems: State of the art and research roadmap*, Communications of the Association for Information Systems, 50(1) 420-438.

Allen, R., and Choudhury, P. (2022) *Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion*, Organization Science, 33(1) 149-169.

Aversa, P., Cabantous, L., and Haefliger, S. (2020). *When decision support systems fail: Insights for strategic information systems from Formula 1*. In Strategic Information Management. Routledge. pp. 403-429.

Badakhshan, P., Wurm, B., Grisold, T., Geyer-Klingeberg, J., Mendling, J., and Vom Brocke, J. (2022) *Creating business value with process mining*, The Journal of Strategic Information Systems, 31(4) 101745.

Baird, A., and Maruping, L. M. (2021) *The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts*, MIS Quarterly, 45(1).

Berente, N., Gu, B., Recker, J., and Santhanam, R. (2021) *Managing artificial intelligence*, MIS Quarterly, 45(3).

Berente, N., Lyytinen, K., Yoo, Y., and Maurer, C. (2019) *Institutional logics and pluralistic responses to enterprise system implementation: a qualitative meta-analysis*, MIS Quarterly, 43(3) 873-902.

Berente, N., Salge, C. A. D. L., Mallampalli, V. K., and Park, K. (2022) *Rethinking project escalation: An institutional perspective on the persistence of failing large-scale information system projects*, Journal of Management Information Systems, 39(3) 640-672.

Borges, A. F., Laurindo, F. J., Spínola, M. M., Gonçalves, R. F., and Mattos, C. A. (2021) *The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions*, International Journal of Information Management, 57 102225.

Boyatzis, R. E. (1998) Transforming qualitative information: Thematic analysis and code development, Sage Publications.

Burström, T., Parida, V., Lahti, T., and Wincent, J. (2021) *AI-enabled business-model innovation and transformation in industrial ecosystems: A framework, model and outline for further research*, Journal of Business Research, 127 85-95.

Cheng, X., Su, L., Luo, X., Benitez, J., and Cai, S. (2022) *The good, the bad, and the ugly: Impact of analytics and artificial intelligence-enabled personal information collection on privacy and participation in ridesharing*, European Journal of Information Systems, 31(3) 339-363.

Collins, C., Dennehy, D., Conboy, K., and Mikalef, P. (2021) *Artificial intelligence in information systems research: A systematic literature review and research agenda*, International Journal of Information Management, 60 102383.

Dyer, J. H., and Singh, H. (1998) *The relational view: Cooperative strategy and sources of interorganizational competitive advantage*, Academy of Management Review, 23(4) 660-679.

Dyer, J. H., Singh, H., and Hesterly, W. S. (2018) *The relational view revisited: A dynamic perspective on value creation and value capture*, Strategic Management Journal, 39(12) 3140-3162.

Engel, C., Elshan, E., Ebel, P., and Leimeister, J. M. (2024) *Stairway to heaven or highway to hell: A model for assessing cognitive automation use cases*, Journal of Information Technology, 39(1) 94-122.

Enholm, I. M., Papagiannidis, E., Mikalef, P., and Krogstie, J. (2022) *Artificial intelligence and business value: A literature review*, Information Systems Frontiers, 24(5) 1709-1734.

Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P. (2024) *Generative AI*, Business & Information Systems Engineering, 66(1) 111-126.

Findikoglu, N. M., Ranganathan, C., and Watson-Manheim, M. B. (2021) *Partnering for prosperity: small IT vendor partnership formation and the establishment of partner pools*, European Journal of Information Systems, 30(2) 193-218.

Giddens, A. (1984) The constitution of society: Outline of the theory of structuration. Univ of California Press.

Grover, V., and Kohli, R. (2012) *Cocreating IT value: New capabilities and metrics for multifirm environments*, MIS Quarterly, 36(1) 225-232.

Hopf, K., Müller, O., Shollo, A., and Thiess, T. (2023a) *Organizational implementation of AI: Craft and mechanical work*, California Management Review, 66(1) 23-47.

Hopf, K., Weigert, A., and Staake, T. (2023b) *Value creation from analytics with limited data: a case study on the retailing of durable consumer goods*, Journal of Decision Systems, 32(2) 289-325.

Huang, J., Henfridsson, O., and Liu, M. J. (2022) *Extending digital ventures through templating*, Information Systems Research, 33(1) 285-310.

Jacobides, M. G., Brusoni, S., and Candelon, F. (2021) *The evolutionary dynamics of the artificial intelligence ecosystem*, Strategy Science, 6(4) 412-435.

Jöhnk, J., Weißert, M., and Wyrtki, K. (2021) *Ready or not, AI comes—an interview study of organizational AI readiness factors*, Business & Information Systems Engineering, 63(1) 5-20.

Keller, R., Stohr, A., Fridgen, G., Lockl, J., and Rieger, A. (2019) *Affordance-experimentation-actualization theory in artificial intelligence research: a predictive maintenance story*, In International Conference on Information Systems (ICIS 2019), AIS Proceedings.

Keramidis, P. and Shollo, A. (2024) *Perceptions of Artificial Intelligence Business Value in a Value Network*, In Proceedings of the European Conference on Information Systems (ECIS 2024), AIS Proceedings. 13.

Kohli, R., and Grover, V. (2008) *Business value of IT: An essay on expanding research directions to keep up with the times*, Journal of the Association for Information Systems, 9(1) 1.

Kude, T., and Huber, T. L. (2024) *Responding to platform owner moves: A 14-year qualitative study of four enterprise software complementors*, Information Systems Journal.

Liang, H., and Xue, Y. (2022) *Save face or save life: Physicians' dilemma in using clinical decision support systems*, Information Systems Research, 33(2) 737-758.

Magistretti, S., Dell'Era, C., and Petruzzelli, A. M. (2019) *How intelligent is Watson? Enabling digital transformation through artificial intelligence*, Business Horizons, 62(6) 819-829.

Malik, M., Andargoli, A., Clavijo, R. C., and Mikalef, P. (2024) *A relational view of how social capital contributes to effective digital transformation outcomes,* The Journal of Strategic Information Systems, 33(2) 101837.

Mandrella, M., Trang, S., and Kolbe, L. M. (2020) *Synthesizing and integrating research on IT-based value cocreation: A meta-analysis,* Journal of the Association for Information Systems, 21(2) 4.

Mayer, A. S., Strich, F., and Fiedler, M. (2020) *Unintended consequences of introducing AI systems for decision making*, MIS Quarterly Executive, 19(4).

Mikalef, P., and Gupta, M. (2021) *Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance,* Information & Management, 58(3) 103434.

Noblit, G. W., and Hare, R. D. (1988) Meta-ethnography: Synthesizing qualitative studies (Vol. 11), Sage.

Olan, F., Arakpogun, E. O., Suklan, J., Nakpodia, F., Damij, N., and Jayawickrama, U. (2022) *Artificial intelligence and knowledge sharing: Contributing factors to organizational performance*, Journal of Business Research, 145, 605-615.

Papagiannidis, E., Enholm, I. M., Mikalef, P., and Krogstie, J. (2021) *Structuring AI Resources to Build an AI Capability: A Conceptual Framework*, In Proceedings of the European Conference on Information Systems (ECIS 2021), AIS Proceedings.

Prasad, A., Green, P., and Heales, J. (2013) *On governing collaborative information technology (IT): A relational perspective*, Journal of Information Systems, 27(1) 237-259.

Rai, A., Pavlou, P. A., Im, G., and Du, S. (2012) *Interfirm IT capability profiles and communications for cocreating relational value: evidence from the logistics industry,* MIS Quarterly, 36(1) 233-262.

Ranjan, J., and Foropon, C. (2021) *Big data analytics in building the competitive intelligence of organizations*, International Journal of Information Management, 56 102231.

Ruokonen, M., and Ritala, P. (2023) *How to succeed with an AI-first strategy?,* Journal of Business Strategy, (ahead-of-print).

Schryen, G. (2013) *Revisiting IS business value research: what we already know, what we still need to know, and how we can get there*, European Journal of Information Systems, 22 139-169.

Shollo, A., and Vassilakopoulou, P. (2024) *Beyond Risk Mitigation: Practitioner Insights on Responsible AI as Value Creation,* In Proceedings of the European Conference on Information Systems (ECIS 2024), AIS Proceedings.

Shollo, A., Hopf, K., Thiess, T., and Müller, O. (2022) *Shifting ML value creation mechanisms: A process model of ML value creation,* The Journal of Strategic Information Systems, 31(3) 101734.

Someh, I., Wixom, B., & Zutavern, A. (2020) *Overcoming organizational obstacles to artificial intelligence value creation: propositions for research,* In Proceedings of the 53rd Hawaii International Conference on System Sciences.

Stevens, M., MacDuffie, J. P., and Helper, S. (2015) *Reorienting and recalibrating inter-organizational relationships: Strategies for achieving optimal trust*, Organization Studies, 36(9), 1237-1264.

Trang, S., Mandrella, M., Marrone, M., and Kolbe, L. M. (2022) *Co-creating business value through IT-business operational alignment in inter-organisational relationships: empirical evidence from regional networks*, European Journal of Information Systems, 31(2) 166-187.

Wamba-Taguimdje, S. L., Wamba, S. F., Kamdjoug, J. R. K., and Wanko, C. E. T. (2020) *Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects*, Business Process Management Journal, 26(7) 1893-1924.

Wang, C., Teo, T. S., and Janssen, M. (2021) *Public and private value creation using artificial intelligence: An empirical study of AI voice robot users in Chinese public sector*, International Journal of Information Management, 61 102401.

Zhang, D., Pee, L. G., and Cui, L. (2021) *Artificial intelligence in E-commerce fulfillment: A case study of resource orchestration at Alibaba's Smart Warehouse*, International Journal of Information Management, 57 102304.

Zhang, Z., Yoo, Y., Lyytinen, K., and Lindberg, A. (2021) *The unknowability of autonomous tools and the liminal experience of their use*, Information Systems Research, 32(4) 1192-1213.

Appendix 1

This appendix includes the search queries used for both the search and the citation analysis.

Positive value query
TITLE-ABS-KEY (("Artificial intelligence" OR "(AI)" OR " AI " OR "cognitive technology" OR "robotic automation" OR "cognitive insight" OR "process automation" OR "machine learning" OR "(ML)" OR " ML " OR "deep learning" OR "(DL)" OR " DL " OR "large language model" OR "LLM" OR "large-language model" OR "cognitive automation" OR "neural network" OR "supervised learning" OR "unsupervised learning" OR "natural language processing" OR "(NLP)" OR " NLP " OR "computer vision" OR "machine vision" OR "expert systems" OR "cognitive application" OR "image recognition" OR "reinforcement learning" OR "deep mind technologies" OR "adaptive algorithms" OR "recurrent neural networks" OR "heuristic search techniques" OR "decision tree" OR "data mining" OR "cluster analysis" OR "classification" OR "chatbots" OR "semantic analysis" OR "Bayesian learning system" OR "rule-based AI" OR "rule-based engine" OR "symbolic AI") AND ("Business value" OR "business benefits" OR "business process redesign" OR "organizational change" OR "firm performance" OR "organizational performance" OR "competitive advantage" OR "process innovation" OR "business transformation" OR "business gains" OR "business performance" OR "competitive performance" OR

"business efficiency" OR "reduce business costs" OR "commercial value" OR "business value proposition" OR "business growth" OR "business success" OR "customer value" OR "corporate value" OR " value creation" OR "value capture" OR "create value" OR "capture value" OR "business value manifestation" OR "firm performance" OR "strategic value" OR "competitive value" OR "value-creation" OR "value co-creation" OR "value cocreation" OR "creating value" OR "cocreate value" OR "cocreating value" OR "co-create value" OR "co-creating value" OR "create business value" OR "creating business value" OR "cocreate business value" OR "cocreating business value" OR "co-create business value" OR "co-creating business value" OR "value capturing" OR "capturing value" OR "capture business value" OR "capturing business value")) AND PUBYEAR > 2013
Negative/unintended value query
TITLE-ABS-KEY (("Artificial intelligence" OR "(AI)" OR " AI " OR "cognitive technology" OR "robotic automation" OR "cognitive insight" OR "process automation" OR "machine learning" OR "(ML)" OR " ML " OR "deep learning" OR "(DL)" OR " DL " OR "large language model" OR "LLM" OR "large-language model" OR "cognitive automation" OR "neural network" OR "supervised learning" OR "unsupervised learning" OR "natural language processing" OR "(NLP)" OR " NLP " OR "computer vision" OR "machine vision" OR "expert systems" OR "cognitive application" OR "image recognition" OR "reinforcement learning" OR "deep mind technologies" OR "adaptive algorithms" OR "recurrent neural networks" OR "heuristic search techniques" OR "decision tree" OR "data mining" OR "cluster analysis" OR "classification" OR "chatbots" OR "semantic analysis" OR "Bayesian learning system" OR "rule-based AI" OR "rule-based engine" OR "symbolic AI") AND ("value destruction" OR "destroy value" OR "destroy business value" OR "value destructive" OR "value destructing" OR "negative value" OR "negative impact*" OR "dark side" OR "unexpected" OR "unintended")) AND PUBYEAR > 2013

Appendix 2

This appendix includes the journals selected for the searching strategy.

Information Systems outlets
Decision Support Systems
European Journal of Information Systems
Information & Management

Information and Organization
Information Systems Journal
Information Systems Research
Journal of the AIS
Journal of Information Technology
Journal of MIS
Journal of Strategic Information Systems
MIS Quarterly
Strategy outlets
Strategic Management Journal
Strategic Organization
Journal of Business Strategy
Organization Studies outlets
Organization Studies
Organization Science
Organizational Research Methods
Management outlets
Academy of Management Journal
Academy of Management Review
Administrative Science Quarterly
Management Science
Journal of Management
Executive outlets
MISQ Executive
California Management Review
Harvard Business Review
MIT Sloan Management Review

Appendix 3

This appendix includes the coding table.

Concept	**Definition**	**Codes**	
		Indicative Subcodes	***Indicative Quotes***
Factors that lead to initial value creation	As noted in Dyer et al. (2018, p. 3143), initially there are four factors that (all else being equal) lead to value creation: complementary resources, relation specific assets, knowledge-sharing routines and effective governance. These reflect the need for interdependence and complementarity. In the context of AI, the resources might refer	*Complementary resources*	
		Data + algorithms and platforms as resources	"We had a short discussion with the IBM rep about how we could pursue this aspiration. We started by working together with their data scientists to repeatedly train the platform. Then we prepared a small sample set for them to analyze as a proof of concept." (Wamba-Taguimdje et al., 2020)
		Users as resource	"It was extremely hard to reach users in this market, where you just have to tap

	to data, algorithms, infrastructure and market segment, as well as the knowledge and governance necessary to combine them.		into existing social networks to get your product noticed.” (Huang et al., 2022)
		Interpretation of data by the business analysts	“The improvements as they come out of the analysis is something that we don’t own but we facilitate. One of the key roles is the business analyst because first you think the tool is doing everything for you and you would see all areas and detect improvements and automation, but you need business experts to interpret the data and findings; somebody who explores, analyzes, and presents, and interprets the data” (Badakhshan et al., 2022)
		Relation-specific assets	
		N/A	
		Knowledge-sharing routines	
		Constraint mitigation as main knowledge-sharing	“We exchanged ideas and consulted each other. It was very interesting to see that many of the participating companies face the same problems” (Keller et al., 2019)
		Knowledge sharing can lead to independence	“We would like to talk to some specialists, but we would like to do it and apply it ourselves. So yeah, again, we want to be independent and to have as much knowledge about the [Company] solutions as possible, so that we can design, we can develop ourselves.” (Keramidis & Shollo, 2024)
		Spending time talking with the people with knowledge is important	“Spending time to talk to people with knowledge of the underlying processes is critical. It is beneficial to use the listening muscle and show enthusiasm for what they are doing. It

			allows to collect and understand business requirements, which are essential in order to develop solutions that are aligned with business needs” (Engel et al., 2024)
		Effective governance	
		Partners' processes understanding	“we put our initial effort in getting a strong understanding as [to] where the complexity lies. For example, with a new partner, you need to know what procedures are carried out online and what is done offline” (Huang et al., 2022)
		Debt counselling agencies verify AI decisions	“Talking to debt counselling agencies, we never heard that customers were overindebted because of CleverLoan decisions. And, for us, that was always the greatest compliment. This has always reflected quite positively on the fact that our loan decision algorithms work well.” (Mayer et al., 2020)
		Responsible governance can position trustworthiness	“[by ensuring responsible practices the company can] get some goodwill because you are like a player that does some good for a large amount of the population. So, it makes sense to position yourself as a trustworthy partner” (Shollo & Vassilakopoulou, 2024)
Factors that lead to diminished value creation	Over time, the value creation afforded by the partnership decreases (Dyer et al. 2018). This is due to both internal (decrease in resource complementarity, relational inertia) and environmental	*Decrease in resource complementarity*	
		Knowledge outsourced to the AI outside of organization	“A disadvantage is that profound loan decision knowledge is outsourced—it is now transferred into some kind of algorithm” (Mayer et al., 2020)
		Third party AI models create dependences	“[But as] companies want to take into use

	(replication from competitors, obsolescence of alliance resources) factors. In the case of AI, the dependence on third party algorithms has been discussed as a major impediment. Further, the replication of models and infrastructure is also a notable problem, while trust and knowledge are important to consider over the evolution of the partnership in order for that to sustain.		Generative AI opportunities [(StellaAI) they know that there are] uncertainties and unclarities in this area [and] there are also very significant third party dependencies" (Shollo & Vassilakopoulou, 2024)
		Increase in relational inertia	
		N/A	
		Replication of alliance resources	
		AI programming as the competitive advantage	"I can't tell you exactly how our AI system is programmed because this is our competitive advantage, but I can tell you that there is not ONE SINGLE decision determining criterion" (Mayer et al., 2020)
		Product and market segment are imitable	"Once you have a product in the market that shows high growth potential, you also quickly notice that it won't take long for others to imitate and better what you can do.… In this market, you just need to be very quick on your feet. So that you are always several steps ahead and let others play the catching up game" (Huang et al., 2022)
		Obsolescence of alliance resources	
		Trust is lost from stakeholders	"So I think it was running for three months and then there was sort of a drift in the performance. And then [the stakeholder] didn't trust us as much as after that" (Shollo et al., 2022, Appendix A.4)
		Knowledge sharing is difficult because results are individual	"Our first learning concerns the transferability of individual results, which is extremely complex because the companies are very different." (Keller et al., 2019)
Factors that lead to	Competition within the alliance may	*Superiority of one partner at absorbing/replicating the partner's complementary resources*	

competition for value capture	change over time and that can affect the value capture (Dyer et al. 2018). That corresponds to both internal (resource absorbing superiority, additional resources development, asymmetries in investments) and external (competitors' imitation of resources) to the partnership factors. The discussion around AI focuses more extensively on the creation of additional resources associated with AI by one partner.	N/A	
		One partner develops additional resources	
		Partner keen on adding more similar ventures	"Operating a fully linked product system enables us to cover the entire life cycle scenarios and at the same time generates a lot of synergy. These full life cycle scenarios ranging from are marketing, interaction, transaction, payment, and data, providing marketing antifraud, application antifraud, transaction antifraud, account security, and data anticrawling, etc., are shared amongst all the products. Given the shared nature of product, technologies and services developed to support the products can be highly modularized. This structure has multiple benefits. On the one hand, we can meet multiple business needs with minimal resource requirements. On the other hand, the more products we have, the more sharing can be materialized." (Huang et al., 2022)
		Resource interoperability as focus in entering the partnership	"We are at heart a technology company. We know our strength and opportunities lie in our ability to apply the same technology to solve many problems in the market, and create and meet the demand of many unserved areas in the market" (Huang et al., 2022)
		Asymmetries in partner investments	
		N/A	
		Competitors imitate one partner's alliance-specific complementary resources	
		N/A	

Exploring AI-powered Digital Innovations from A Transnational Governance Perspective: Implications for Market Acceptance and Digital Accountability

Claire Li
Royal Holloway, University of London
David Peter Wallis Freeborn
Northeastern University London

Completed Research

Abstract

This study explores the application of the Technology Acceptance Model (TAM) to AI-powered digital innovations within a transnational governance framework. By integrating Latourian actor-network theory (ANT), this study examines how institutional motivations, regulatory compliance, and ethical and cultural acceptance drive organisations to develop and adopt AI innovations, enhancing their market acceptance and transnational accountability. We extend the TAM framework by incorporating regulatory, ethical, and socio-technical considerations as key social pressures shaping AI adoption. Recognizing that AI is embedded within complex actor-networks, we argue that accountability is co-constructed among organisations, regulators, and societal actors rather than being confined to individual developers or adopters. To address these challenges, we propose two key solutions: (1) internal resource reconfiguration, where organisations restructure their governance and compliance mechanisms to align with global standards; and (2) reshaping organisational boundaries through actor-network management, fostering engagement with external stakeholders, regulatory bodies, and transnational governance institutions. These approaches allow organisations to enhance AI accountability, foster ethical and regulatory alignment, and improve market acceptance on a global scale.

Keywords: Digital innovations, Digital accountability, Ethical AI, Technology acceptance model, Transnational governance, Latourian Actor-Network Theory

1.0 Introduction

This study examines how the Technology Acceptance Model (TAM) can be adapted to evaluate AI innovations in a transnational context. It provides insights into the reasons for organizations to develop and adopt AI technologies, and the ways in which they can navigate regulatory, ethical, and cultural landscapes. Additionally, it explores how organisations can enhance global market acceptance and strengthen accountability for digital innovations.

Technology is one piece of the puzzle that must be solved for companies to remain competitive in a digital world (Rivard 2004). Organisations also require adequate strategies (Bharadwaj et al 2013; Matt et al 2015) as well as suitable internal structures (Selander and Jarvenpaa 2016), processes (Carlo et al 2012), and culture (Karimi and Walter 2015) to yield the capability to generate new paths and innovations for value creation (Svahn et al 2017). Existing studies offer valuable insights into value creation (Vial, 2021), yet the organisational motivations behind innovation development and strategies for improving the process remain underexplored. Moreover, there is a lack of understanding regarding how these influences persist after digital innovations have been adopted in other nations. Therefore, our study addresses two research questions from a transnational governance perspective: "*Why do organisations develop and adopt AI powered digital innovations?*" and "*How do organisations manage the process of improving the market acceptance of their AI-powered digital innovations?*".

To address these gaps, our study employs TAM as an underlying analytical framework. TAM, initially developed by Davis, Bagozzi and Warshaw (1989), provides a theoretical framework for analysing the factors that influence users' adoption of new technologies. However, the transnational nature of AI innovations introduces complex governance challenges, arising from varying regulatory frameworks and cultural norms. Cultural differences can shape different individuals' perspectives and priorities, leading to significantly varying values and approaches across different regions (Toon, 2024). AI developers may unwittingly bring their cultural perspectives and cognitive biases into the process of AI development, for example by using unrepresentative training data or an imperfect algorithmic structure (Fazelpour and Danks, 2021, Athota et al., 2023). Lacking complete information can cause bias by excluding certain groups or sections of the population. The rise of *transnational governance* (Djelic & Quack, 2007; Djelic & Sahlin-Andersson, 2006) has heightened the need for a more nuanced understanding of how AI-driven innovations are accepted across borders, taking into account the diversity of cultural, ethical, and regulatory environments.

This theoretical lens helps us explain how certain factors of digital transformation become critical in market acceptance in transnational governance. However, it neither tells us about the motivations that drive organisations towards digital transformation, nor how organisations can respond to these factors, leading to an increase in market acceptance and digital accountability. Social pressures are especially important in driving organisations toward digital transformation (Gegenhuber et al 2022; Saarikko, Westergren and Blomquist 2020; Zhu et al 2006). When organisations face social pressure, they are more likely to prioritise digitalisation to maintain their legitimacy, reputation, and market position (Lee, Pak and Roh 2024). Additionally, responding to societal expectations can improve corporate social performance, making organisations more responsible and adaptive to new digital trends. In times of social or economic crises, organisations often accelerate digital initiatives to meet new demands, demonstrating that social pressure is a significant catalyst for transformation (Khurana, Dutta and Ghura 2022). We use the concept of social pressure to understand the institutional motivations, particularly regulatory, ethical and cultural pressures, for organisations to develop and adopt AI-powered digital innovations.

Based on these insights, our study proposes two solutions to enhance the transnational accountability and market acceptance of AI-powered digital innovations: reconfiguring internal governance and reshaping organisational boundaries through actor-network management. Drawing on Latour’s (1987, 2005) actor-network theory (ANT), which views technological systems as shaped by the interactions of both human and non-human actors, we argue that accountability is co-constructed through dynamic relationships between AI developers, regulatory bodies, users, and governance institutions.

ANT helps to illustrate that AI-powered innovations do not operate in isolation but are embedded within broader socio-technical networks that influence their development, deployment, and societal impact. In this light, organisations must actively engage with these networks to ensure ethical and regulatory alignment. We contend that AI or digital innovations themselves are not inherently biased or culturally insensitive; rather, bias arises from the ways in which organisations develop, train, and direct AI systems. Therefore, from a transnational governance perspective, organisations must assume transnational accountability for digital innovations, as they control the AI development processes, training platforms, and systemic design choices that directly affect market acceptance and ethical considerations. By integrating ANT with transnational governance principles, we highlight the importance of both internal organisational restructuring and external boundary management in fostering responsible, compliant, and globally accepted AI innovations.

Our study aims to make the following contributions to the literature on digital transformation. First, we investigate why and how organisations develop and adopt AI-powered digital innovations in transnational governance. Prior research has focused on how organisations develop digital technologies by building up information architecture (Tan, Abdaless and Liu 2018; Tan, Liu and White 2013). Related studies have explored how digital innovations can become affordable and acceptable by the market via organisational semiotics (Pan et al 2018; Hafezieh and Eshraghian 2022; Hafezieh and Pollock 2023; Nambisan et al 2017). We differ from this stream of literature by exploring how organisations can adapt their AI-powered innovations in a transnational context. In this vein, we further demonstrate how organisations show differences in AI-powered digital transformation when facing different regulatory and ethical pressures. We contend that regulatory, ethical and cultural acceptance pressures are important factors in technology acceptance in transnational governance.

Second, we contribute to the existing literature on the motivations driving organisations toward digital transformation. Previous research has highlighted several key factors that influence the adoption of digital innovations, including perceived ease of use, perceived usefulness, trust, security risks, costs, privacy concerns, cultural context, and social influence (Pan et al., 2018; Kim, Mirusmonov, & Lee, 2010; Shin, 2010; Koenig-Lewis et al., 2015; Lu et al., 2011; Arvidsson, 2014; Slade, Williams, & Dwivedi, 2013; Mallat et al., 2009). We add another layer to the literature by explaining that regulatory, ethical and cultural acceptance pressures are important factors driving digital transformation, drawing ethical AI principles and cultural acceptance discussions.

Third, we respond to Vial's (2021) call to explore the ways in which organisations improve their market acceptance in digital transformation. Prior research has used search as an approach to inform organisations to renew their innovative products, innovation processes and redefine their value propositions (Hafezieh and Eshraghian 2022; Hafezieh and Pollock 2023). We extend this discussion by arguing that digital innovation developers and adopters must assume transnational accountability for their innovations, as AI technologies operate within complex socio-technical networks that transcend national borders. Drawing on Latourian actor-network theory (ANT), we highlight how AI-powered innovations are shaped not only by technological advancements but also by interactions between regulatory bodies, industry stakeholders, and end-users. Accountability, therefore, is not static but co-constructed within these evolving actor-networks. We accordingly propose two solutions for organisations to enhance market acceptance and accountability in AI-powered digital transformation – internal resource reconfiguration and reshaping organisational boundaries through actor-network management.

2.0 Motivations of Developing and Applying AI-powered Digital Innovations

2.1 Profit-driven Motivation

Organisations are increasingly adopting AI-powered digital innovations to enhance their profitability (Fountaine, McCarthy and Saleh 2019). These technologies allow businesses to automate routine tasks, reduce operational costs, and increase efficiency. AI systems can process vast amounts of data, enabling more accurate forecasting and decision-making, which directly contributes to improved financial performance (Olan et al 2022). Moreover, AI-driven innovations offer personalized services to customers, enhancing customer satisfaction and driving revenue growth (Usman et al 2024). The competitive advantage that AI provides has become a significant motivation for organisations to continuously invest in and develop such technologies.

2.2 Market Acceptance and Technology Acceptance Model (TAM)

Market acceptance refers to the process by which consumers or businesses adopt and use a product or technology (Gao et al 2013). It is crucial because it determines the commercial success of a product, influencing profitability and long-term sustainability. Without market acceptance, even the most innovative products may fail due to low adoption rates.

The Technology Acceptance Model (TAM) provides a framework for understanding how AI-powered innovations gain market acceptance (Davis 1989; Silva 2015). TAM is an information systems theory developed to explain how users come to accept and use technology. It is based on two key variables. The first variable is *Perceived Usefulness*, referring to the degree to which an individual believes that using a particular technology or system will enhance their job performance or productivity. If a user perceives that a system or technology is useful and can improve their tasks or roles, they are more likely to accept and use it. This concept highlights the importance of demonstrating clear benefits to users in order to encourage technology adoption. AI-powered innovations are adopted more readily if users perceive that the technology improves their performance or adds value. In fields like accounting, AI tools that enhance efficiency and accuracy can lead to higher acceptance rates.

The second variable is *Perceived Ease of Use*, refers to the degree to which a person believes that using a particular technology or system will be free of effort. If a user perceives that a system is simple and straightforward to use, they are more likely to accept and use it. Perceived ease of use is essential because it reduces the cognitive load and learning curve associated with new technologies, increasing the likelihood of widespread adoption. The simpler and more intuitive AI technologies are to operate; the more likely users are to adopt them. Advanced user interfaces and seamless integration into existing workflows reduce friction in adoption, leading to broader market acceptance.

In the case of AI-powered digital innovations, organisations perceive these technologies as highly useful due to their ability to automate complex processes and deliver real-time insights. At the same time, advancements in user interfaces and AI integration make these innovations easier to use, thereby fostering broader acceptance within organisations.

2.3 Extended TAM and Transnational Governance

Surprisingly, despite the increasing importance of AI-powered digital innovations, there has been relatively little exploration of its institutionalisation impacts, particularly in the context of transnational economic and market governance (Arnold 2009a; 2009b; Mehrpouya and Salles-Djelic 2019; Friedrich, Kunkel and Thiemann 2024). Our study extends TAM by exploring the institutionalisation impacts of AI-powered digital innovations from the transnational governance perspective.

The *transnational governance perspective* refers to the collaborative efforts of countries, organisations, and stakeholders to regulate, manage, and oversee emerging technologies or global issues beyond national borders (Roger and Dauvergne 2016). This perspective recognizes that challenges such as AI, climate change, and cyber security require international coordination and governance mechanisms due to their global impact.

In the context of AI, transnational governance involves creating frameworks that standardize rules, ensure ethical usage, and mitigate risks across different jurisdictions. It relies on cooperation between states, international organisations, and private sector actors to establish regulations that reflect shared values, such as accountability, transparency, and fairness. For instance, the European Union's regulatory model for AI is one of the leading examples of successful AI governance at the transnational level. Transnational governance is critical because AI’s global nature means that decisions in one region can have ripple effects worldwide. This necessitates diplomatic engagement, international law, and shared policy frameworks to manage both the risks and opportunities of AI.

An extended version of TAM can be applied to consider factors relevant to transnational governance when examining AI-powered innovations. While *Perceived Usefulness* and *Perceived Ease of Use* remain critical, institutional variables such as *regulation compliance* and *cultural and ethical acceptance* should be considered. Organisations operating across borders must ensure that their AI systems comply with various international regulations, such as data privacy laws (e.g., GDPR in Europe). The ease with which AI innovations can be adapted to meet these governance requirements can significantly impact their acceptance. Furthermore, organisations must consider the cultural context in which these technologies are deployed, ensuring they are compatible with local norms and ethical principles.

2.4 Social pressures as motivations

Our study uses the concept of *social pressures* to explore the motivations for organisations to develop and adopt AI-powered digital innovations. *Social pressures* refer to the external influences exerted by society, stakeholders, or peers on organisations or individuals to conform to certain norms, regulations, or ethical standards (Bursztyn and Jensen 2017). These pressures can arise from regulatory bodies, customers, cultural expectations, and societal values, prompting organisations to align their strategies and practices with accepted norms to maintain legitimacy, reputation, and competitiveness.

2.4.1 AI Regulations and Digital Innovation

Organisations face growing social pressure to comply with AI regulations, which are being developed across different jurisdictions (OECD, 2021). Regulatory frameworks focus on issues such as data privacy, transparency, and accountability in AI development. Failure to adhere to these laws can result in penalties, reputational damage, and loss of consumer trust. Therefore, organisations are motivated to develop and adopt AI innovations that comply with these regulations to maintain market access and ensure sustainable growth.

AI regulations play a crucial role in shaping the trajectory of digital innovations by setting the boundaries within which AI systems operate. Effective regulation is essential to ensure that AI advancements are aligned with societal values and ethical standards, thereby fostering trust and widespread adoption.

AI regulations can significantly influence the pace and direction of digital transformation. Rules-based regulations, such as the European Commission AI Act, passed on March 13, 2024 (EU, 2024), provide certainty and clarity, offering uniform applications across the EU. This approach ensures that AI systems meet specific predefined standards, thus safeguarding public interests and maintaining high levels of trust. However, the rigidity of rules-based regulations can stifle innovation and hinder technological advancements, as they may not be adaptable to the rapid pace of AI development.

On the other hand, principles-based regulations such as those proposed in the UK's White Paper on AI regulation (UK Government, 2023) adopt a more flexible and adaptive approach. This framework focuses on desired outcomes rather than specific rules, encouraging innovation while addressing key principles such as safety, security, robustness, transparency, fairness, accountability, and governance. This flexibility encourages rapid adaptation to new technological developments and diverse applications of AI. However, the inherent ambiguity and subjectivity in principles-based regulations can lead to inconsistent interpretations and applications, posing challenges for legal settlements, and potentially allowing unethical behaviour.

Regulations/rules compliance is a social norm of crime avoidance and can form social pressures for organisations to comply with AI regulations/rules. Incompliance with AI regulations can result in possible crimes and negative social influences for organisations, reducing their technology acceptance level. For this reason, we contend that AI regulation compliance is an important factor to motivate organisations to develop and adopt AI-powered innovations because social pressures push them to comply with social norms and regulations.

By grounding regulations in a robust theoretical understanding of risk, it is possible to create more effective, adaptable, and ethically sound AI governance structures that support sustainable digital innovation.

2.4.2 Ethical and cultural acceptance of AI-powered innovations

Ethical and cultural acceptance is another critical social pressure driving organisations to adopt AI responsibly (Lobschat et al 2021). Societies increasingly demand ethical AI usage, with concerns around fairness, bias, and job displacement. Companies that address these ethical concerns may be seen as more socially responsible, and thus more likely to build public trust and customer loyalty. This may motivate organisations to develop AI systems that align with ethical and cultural standards, ensuring acceptance and integration into broader society.

There is widespread agreement that AI regulatory frameworks, whether rules-based or principles-based, should incorporate ethical principles into their design (Ashok et al 2022). These ethical principles are needed to provide a normative foundation for AI governance and to define what constitutes responsible development and use of AI systems. Yet, there are substantial disagreements about which precise ethical principles should be incorporated into regulations governing AI. These disagreements stem from underlying cultural differences, ethical disagreements, as well as uncertainties about how to respond to a rapidly changing technological landscape. Key areas of contention include balancing risks and opportunities, privacy versus security, and transparency against system efficacy and integrity.

Nonetheless, Jobin et al. (2019) identify an emerging consensus around five core ethical principles in guidance documents for ethical AI. We outline these principles below.

The first principle is *Transparency,* meaning that AI systems should be explainable, interpretable, and open to human scrutiny (Zednik, 2021). However, the level and nature of transparency may vary depending on the stakeholder and the specific application. For example, an AI spam-email detection system might offer high transparency to regulators ensuring proper data use, but more limited transparency to individuals to prevent exploitation of system vulnerabilities.

The second principle is *Justice, Fairness and Equity*, meaning that AI systems should not discriminate or create unfair outcomes for different groups (Johnson, 2020; Fazelpour and Danks, 2021). In particular, AI should not perpetuate or exacerbate existing societal biases and inequalities. For instance, a responsible AI recruiting tool should avoid discrimination, whether direct or incidental, based on protected characteristics like gender or ethnicity.

The third principle is *Non-maleficence*, meaning that AI should not cause harm to humans, either deliberately or inadvertently (Floridi and Cowls, 2019). This demands rigorous risk assessment and robust safety measures. For example, an autonomous vehicle should be designed with protection against deliberate or dangerous misuse and multiple fail safes to prevent accidents.

The fourth principle is *Responsibility*, meaning that AI systems should be accountable to humans, and responsible to human oversight (Matthias, 2004). This may involve clear liability frameworks for AI developers and specific human oversight requirements. For instance, AI-driven financial trading systems should have a well-defined human accountability chain for errors, even when decisions are made autonomously.

The fifth principle is *Privacy,* meaning that AI systems must be responsible in their use of personal data and information (Nissenbaum, 2010). This might involve data protection measures and respect for privacy rights. For example, a responsible AI financial assessment tool should implement strict safeguards to protect users' personal data, potentially including data anonymisation or deletion protocols.

These ethical principles provide a framework for identifying the ethical problems that arise in the context of digital innovations driven by AI. However, these principles can sometimes conflict with each other (Blanchard et al., 2024; Sanderson et al., 2023). For example, transparency and privacy may conflict in credit rating algorithms. While individuals and regulators may request explainability in the outcomes, strict privacy laws may demand data minimisation, limiting what can be disclosed. Similarly, fairness and non-maleficence can create conflicts in AI hiring tools. Algorithms designed to ensure demographic fairness may adjust selection criteria, but this can reduce accuracy, leading to less optimal hiring decisions. These inherent tensions will often necessitate some tradeoffs between the ethical principles.

AI systems are not merely technical tools; they are embedded in cultural and ethical contexts that vary across different societies, and AI-driven digital innovations are becoming more integrated into daily life. As such, organisations need to navigate the challenge of aligning their AI innovations with diverse cultural expectations, whilst also addressing ethical concerns. Under this social expectation, a form of *accountability* emerges for organisations that create or deploy AI. Rather than being solely driven by regulatory compliance, companies are increasingly compelled to take responsibility for how their AI-driven platforms influence culture, identity, and representation. Public scrutiny, ethical debates, and the demand for responsible AI have pushed organisations to be more transparent about their design choices, data sources, and decision-making processes. This accountability is not just about preventing harm; it is also about gaining legitimacy and trust. When organisations demonstrate ethical responsibility in AI development, they enhance their reputation and increase the likelihood of widespread acceptance of their technologies.

What makes this accountability even more significant is that it transcends national borders, creating *transnational accountability* for digital innovations. AI is a global phenomenon—its impact is not confined to the country where it is developed but extends across multiple regions and cultures. As a result, we contend that the responsibility for ensuring ethical AI use does not belong to any single government or organisation. Instead, multinational organisations, international regulatory bodies, and cross-border advocacy groups all play a role in shaping AI governance. Transnational accountability arises when organisations must answer not only to their home governments but also to global stakeholders, including consumers, civil society organisations, and policymakers from different countries. This interconnected accountability structure drives the push for shared standards, ethical frameworks, and collaborative governance efforts to ensure that AI is developed and used in ways that respect cultural diversity and human rights worldwide.

We contend that AI regulations, ethical and cultural acceptance form the main social pressures for organisations to develop and adopt AI-powered digital innovations. From a transnational governance perspective, we propose that this social pressure not only motivates organisations in technology advancement and market acceptance but also creates mutual and transnational accountability for digital innovations.

Based on these discussions, we propose the following framework of motivations for organisations to develop and adopt AI-powered digital innovations (Figure 1).

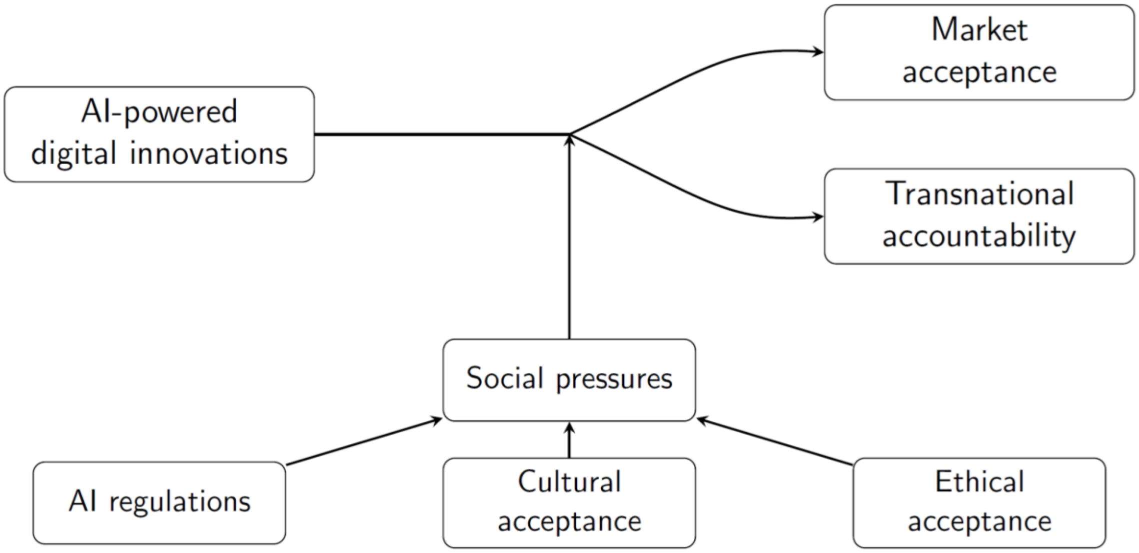


Figure 1. Motivations for organisations to develop and adopt AI-powered digital innovations.

3.0 Proposed solutions: Enhancing Transnational Accountability for AI-powered Digital Innovations

3.1 Theoretical Foundations: Latourian Perspective, TAM, and Transnational Governance

We propose solutions based on the previously discussed theoretical foundations, the Technology Acceptance Model (TAM), transnational governance, and Latourian actor-network theory (ANT). Each provides a complementary lens for analysing how organisations navigate AI accountability and market acceptance across global contexts.

From the perspective of ANT, accountability is not a fixed or top-down process but is instead co-constructed through interactions between human and non-human actors (Latour, 2005). In AI governance, this means that AI systems, regulatory institutions, users, and organisations all form dynamic networks of accountability, shaping how AI is perceived and accepted. ANT also highlights that governance is a continuous process of assembling and reassembling networks, meaning organisations must actively shape their AI ecosystems rather than merely conform to existing rules.

The Technology Acceptance Model (TAM) provides a framework for understanding why users adopt technology. It emphasises two primary factors: perceived usefulness (the degree to which a user believes a technology enhances their work) and perceived ease of use (the extent to which a technology is easy to operate). While TAM has traditionally focused on individual users, we extend it to organisational and transnational contexts, where regulatory compliance, ethical concerns, and social pressures shape AI adoption.

Transnational governance refers to the collective regulatory, institutional, and normative frameworks that transcend national borders (Djelic & Sahlin-Andersson, 2006). Unlike traditional governance structures that operate within national jurisdictions, transnational governance ensures that AI-powered digital innovations comply with cross-border legal, ethical, and social expectations. This governance model aligns with TAM by influencing Perceived Usefulness (through regulatory incentives and ethical guidelines) and Perceived Ease of Use (by harmonizing AI standards across different jurisdictions).

By combining these three perspectives, we argue that organisations must adopt a dual strategy, internally restructuring their AI governance while externally managing their actor-network relationships, to improve accountability and ensure ethical and market acceptance of these technologies at a transnational level.

3.2 Internal Reconfiguration and Optimisation of Resources

A key approach to ensuring transnational accountability is internal reconfiguration, which involves realigning organisational structures, governance mechanisms, and workforce competencies to meet evolving regulatory, ethical, and social pressures. Drawing from Latour's concept of "matters of concern" (Latour, 2004), organisations should not treat AI adoption as a mere matter of fact but rather as a socially embedded process requiring continuous accountability and ethical reflexivity. From a TAM perspective, internal reconfiguration enhances perceived usefulness by ensuring that AI systems are optimised for regulatory compliance and cultural adaptability. Embedding ethical considerations such as bias mitigation, data transparency, and fairness into AI design processes strengthens user trust and market acceptance across diverse regions (Ko and Leem, 2021; Kelly, Kaye, & Oviedo-Trespalacios, 2023). Additionally, perceived ease of use improves when governance structures facilitate seamless compliance with varying regulations, reducing friction in cross-border AI deployment (World Bank Group, 2024).

Transnational governance further requires organisations to harmonise their internal AI policies with international frameworks such as the EU AI Act, GDPR, and OECD AI principles. By proactively aligning internal practices with evolving regulatory landscapes, organisations minimise legal risks, enhance legitimacy, and shift AI adoption from a reactive response to market pressures to a deliberate, ethically informed process. To meet these demands, organisations must align governance structures, workforce skills, and technology investments with the evolving AI development and ethical standards. By fostering a culture of continuous learning and agility, organisations improve responsiveness to regulatory changes and societal expectations. Developing internal capabilities in AI ethics and compliance ensures adherence to principles such as fairness and transparency (Song, Lee, and Khanna, 2016). Moreover, optimising resource allocation leads to more efficient AI development, reducing time-to-market while meeting regulatory and market standards.

By optimising internal resources, including governance structures and workforce competencies, organisations can align AI innovations with both market and ethical expectations. This speeds up the development and deployment process, ensures regulatory compliance, and enhances trust with consumers. Organisations that integrate ethical considerations into their AI practices foster higher market acceptance, while efficient resource use strengthens competitiveness and accelerates consumer adoption (Deloitte, 2024).

In the transnational governance landscape, where AI regulations and ethical standards vary across countries, organisations must optimise internal resources to ensure compliance with global and local laws. Aligning governance structures with diverse regulatory frameworks helps organisations navigate cross-border compliance complexities, ensuring their AI-powered innovations meet ethical and legal standards across multiple jurisdictions. This approach builds global trust and enhances market acceptance in regions with differing governance structures and expectations.

3.3 Reshaping Boundaries and Managing Actor-Network Dynamics

The second proposed solution involves reshaping organisational boundaries and managing actor-network dynamics, acknowledging that AI-powered innovations operate within complex socio-technical networks. From the perspective of ANT, accountability is not confined to a single entity but is instead co-constructed through interactions between human and non-human actors, including regulatory bodies, consumers, AI platforms, and governance institutions. This dynamic, relational process extends beyond organisational boundaries, requiring collaborative engagement to ensure responsible AI adoption.

To enhance market acceptance and legitimacy, organisations must actively engage in collaborative AI governance with regulators, industry bodies, and civil society organisations. Participation in multi-stakeholder initiatives, such as the UN’s AI for Good Summit or the Partnership on AI, ensures alignment with international ethical standards and responsiveness to evolving societal concerns. This approach builds trust, fosters compliance, and allows organisations to gain valuable insights into market needs and expectations.

Managing actor-network dynamics also helps organisations navigate transnational accountability by aligning with global AI governance structures. Rather than treating AI adoption as a top-down technological imposition, organisations must engage in public dialogues, open-source collaborations, and regulatory co-creation processes. This participatory approach enhances AI’s cultural acceptance and fosters an inclusive, globally responsible AI ecosystem.

Additionally, organisations should form strategic partnerships and alliances to share resources, expertise, and influence in shaping industry standards. AI-powered innovations are not developed in isolation but within a network of social, technological, and regulatory actors. Adapting strategies to accommodate these broader socio-technical systems helps mitigate social pressures while improving market acceptance by aligning innovations with societal and stakeholder values.

From a governance perspective, organisations that proactively shape actor-network relationships can anticipate regulatory shifts, co-create ethical standards, and strengthen consumer trust. By engaging with international regulatory bodies, local governments, and global stakeholders, companies can ensure compliance with diverse cultural and regulatory environments. This collaborative approach not only fosters AI accountability but also ensures that innovations remain globally accepted, ethically sound, and aligned with transnational governance expectations.

By integrating Latourian theoretical insights with TAM and transnational governance principles, we argue that internal reconfiguration and actor-network management provide robust solutions for organisations to navigate the complex landscape of AI accountability and market acceptance. In a world where AI-driven technologies transcend national borders, organisations must move beyond passive compliance and embrace active engagement with governance ecosystems. These two solutions position organisations as responsible AI stewards, ensuring that AI-powered digital innovations are not only technologically advanced but also ethically sound, legally compliant, and socially accepted on a global scale.

4.0 Conclusion

The study examines the development and adoption of AI-powered digital innovations through the lens of the Technology Acceptance Model (TAM), extended to incorporate the transnational governance perspective. In particular, the study focuses on the motivations for organisations to develop and adopt AI-powered digital innovations, as well as how organisations can enhance the market acceptance of such innovations.

We recognise that AI-powered digital innovations are not inherently biased or culturally insensitive; rather, their development, deployment, and market acceptance are shaped by the organisations that control AI training platforms and direct system operations. We argue that from a transnational governance perspective, organisations must take transnational accountability for their digital innovations, ensuring they align with regulatory, ethical, and cultural expectations. Regulatory and ethical pressures serve as key factors driving technology acceptance, compelling organisations to comply with social norms and legal frameworks. However, ensuring ethical AI use is not the responsibility of any single entity; instead, multinational organisations, international regulatory bodies, and cross-border advocacy groups collectively shape AI governance. Social pressures, including AI regulations, ethical considerations, and cultural acceptance, play a crucial role in motivating organisations to advance technology while fostering market acceptance and mutual accountability. To navigate these challenges, organisations must adopt a dual strategy: internally restructuring governance through resource reconfiguration while externally managing actor-network relationships. By integrating Latourian theoretical insights, TAM, and transnational governance principles, we propose that internal reconfiguration and actor-network management offer robust solutions to strengthen AI accountability and acceptance. In an era where AI operates beyond national borders, organisations must move beyond passive compliance and engage proactively in governance ecosystems to ensure AI is technologically innovative, ethically sound, legally compliant, and socially accepted on a global scale.

Future research could expand upon this work in several promising directions. First, empirical dataset could be developed to support the theoretical framework proposed in this study. Second, this work proposed internal resource reconfiguration; however, we must highlight that the infrastructure accretes during this process (Power 2015). Future research could investigate the alignment between organisational strategies and infrastructure accretion development, in particular information and financial infrastructure (Tan, Abdaless and Liu 2018; Tan, Liu and White 2013). Third, this work also proposed reshaping boundaries and managing the actor-network dynamics as a solution. Future research might examine the mechanisms and search strategies to manage such actor-network dynamics. Fourth, there is a difference between market acceptance and consumer affordance (El Amri ans Akrout 2020; Hafezieh and Eshraghian 2017). Future research can discuss the opportunities and challenges of turning AI-powered digital innovations into affordable products for consumers. Fifth, whilst this study outlines several strategies conceptually, translating them into concrete steps requires further investigation. Further research should focus on developing practical, actionable guidelines to support organisations in implementing the proposed solutions of resource reconfiguration and actor-network management. Empirical case studies might examine best practices for restructuring internal governance mechanisms to enhance AI accountability, as well as frameworks for effectively managing actor-network relationships across transnational regulatory environments.

Finally, it is important to acknowledge that AI-powered digital innovations vary significantly in their design and functionality. For example traditional symbolic or rules-based AI systems operate on a predefined logic, making them relatively predictable but limited in their adaptability, whereas modern neural AI models – such as those used in generative AI and predictive learning systems – leverage vast datasets and continuous learning, often rendering them as black box systems, impenetrable to human understanding (Zednik, 2021). These differences will also likely influence how organisations develop and adopt these technologies, which provides a further promising avenue for future research.

References

Arnold, P. J. (2009a). Global financial crisis: The challenge to accounting research. *Accounting, Organizations and Society*, 34(6-7), 803-809.

Arnold, P. J. (2009b). Institutional perspectives on the internationalization of accounting. In C. S. Chapman, D. J. Cooper, & P. B. Miller (Eds.), Accounting, Organizations & Institutions: Essays in Honour of Anthony Hopwood (pp. 47-64). Oxford University Press.

Arvidsson, N. (2014). Consumer attitudes on mobile payment services–results from a proof of concept test. *International Journal of Bank Marketing*, 32(2), 150-170.

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management*, *62*, 102433.

Athota, V. S., Pereira, V., Hasan, Z., Vaz, D., Laker, B., & Reppas, D. (2023). Overcoming financial planners’ cognitive biases through digitalization: A qualitative study. *Journal of Business Research, 154*, 113291.

Bharadwaj, A., El Sawy, O., Pavlou, P., & Venkatraman, N. (2013). Digital business strategy: toward a next generation of insights. *MIS Quarterly*, 37(2), 471-482.

Blanchard, A., Thomas, C., & Taddeo, M. (2024). Ethical governance of artificial intelligence for defence: Normative tradeoffs for principle to practice guidance. AI & Society.

Bursztyn, L., & Jensen, R. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, *9*(1), 131-153.

Carlo, J. L., Lyytinen, K., & Boland Jr, R. J. (2012). Dialectics of collective minding: contradictory appropriations of information technology in a high-risk project. *MIS Quarterly*, 36(4), 1081-1108.

Davis, F. D. (1989). Technology acceptance model: TAM. *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption*, *205*, 219.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982-1003.

Deloitte. (2024). Technology Trust Ethics: Leadership, governance, and workforce decision-making about ethical AI. Deloitte Insights.

Djelic, M. L., & Quack, S. (2007). Overcoming path dependency: path generation in open systems. *Theory and Society*, 36, 161-186.

Djelic, M. L., & Sahlin-Andersson, K. (Eds.). (2006). Transnational governance: Institutional dynamics of regulation. Cambridge University Press.

El Amri, D., & Akrout, H. (2020). Perceived design affordance of new products: Scale development and validation. *Journal of Business Research*, *121*, 127-141.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*, L168, 12 July 2024.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8), e12760. https://doi.org/10.1111/phc3.12760

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).

Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-powered organization. *Harvard Business Review*, *97*(4), 62-73.

Friedrich, J., Kunkel, T., & Thiemann, M. (2024). Becoming influential: Strategies of control, expertise, and socialisation in transnational governance of accounting regulation. *Accounting, Organizations and Society*, 113, 101566.

Gao, T. T., Rohm, A. J., Sultan, F., & Pagani, M. (2013). Consumers un-tethered: A three-market empirical study of consumers' mobile marketing acceptance. *Journal of Business Research*, *66*(12), 2536-2544.

Gegenhuber, T., Logue, D., Hinings, C. B., & Barrett, M. (2022). Institutional perspectives on digital transformation. In *Digital Transformation and Institutional Theory* (Vol. 83, pp. 1-32). Emerald Publishing Limited.

Hafezieh, N., & Eshraghian, F. (2017, June). Affordance theory in social media research: systematic review and synthesis of the literature. In *25th European Conference on Information Systems (ECIS 2017)*.

Hafezieh, N., & Eshraghian, F. (2022). Adopting a 'Search' Lens in Exploration of How Organizations Transform Digitally. In *Proceedings of the 2022 European Conference on Information Systems.* Association of Information Systems.

Hafezieh, N., & Pollock, N. (2023). Digital consumers and the new 'search' practices of born digital organizations. *Information and Organization*, 33(4), 100489.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389-399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, G. (2020). Algorithmic bias: on the implicit biases of social technology. *Synthese*, 198, 9941-9961.

Karimi, J., & Walter, Z. (2015). The role of dynamic capabilities in responding to digital disruption: a factor-based study of the newspaper industry. *Journal of Management Information Systems*, 32(1), 39-81.

Kelly, S., Kaye, S. A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*, 101925.

Khurana, I., Dutta, D. K., & Ghura, A. S. (2022). SMEs and digital transformation during a crisis: The emergence of resilience as a second-order dynamic capability in an entrepreneurial ecosystem. *Journal of Business Research*, *150*, 623-641.

Kim, C., Mirusmonov, M., & Lee, I. (2010). An empirical examination of factors influencing the intention to use mobile payment. *Computers in Human Behavior*, 26(3), 310-322.

Ko, Y., & Leem, C. S. (2021). The influence of AI technology acceptance and ethical awareness towards intention to use. *Journal of Digital Convergence*, *19*, 217-225.

Koenig-Lewis, N., Marquet, M., Palmer, A., & Zhao, A. L. (2015). Enjoyment and social influence: predicting mobile payment adoption. *The Service Industries Journal*, 35(10), 537-554.

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press.

Latour, B. (2004). Why has critique run out of steam? From matters of fact to matters of concern. *Critical Inquiry*, *30*(2), 225-248

Latour, B. (2005). Reassembling the Social: An Introduction to Actor-Network Theory. Oxford University Press.

Lee, M. J., Pak, A., & Roh, T. (2024). The interplay of institutional pressures, digitalization capability, environmental, social, and governance strategy, and triple bottom line performance: A moderated mediation model. *Business Strategy and the Environment*, 33(6), 5247-5268.

Lobschat, L., Mueller, B., Eggers, F., Brandimarte, L., Diefenbach, S., Kroschke, M., & Wirtz, J. (2021). Corporate digital responsibility. *Journal of Business Research*, *122*, 875-888.

Lu, Y., Yang, S., Chau, P. Y., & Cao, Y. (2011). Dynamics between the trust transfer process and intention to use mobile payment services: A cross-environment perspective. *Information & Management*, 48(8), 393-403.

Mallat, N., Rossi, M., Tuunainen, V. K., & Öörni, A. (2009). The impact of use context on mobile services acceptance: The case of mobile ticketing. *Information & Management*, 46(3), 190-195.

Matt, C., Hess, T., & Benlian, A. (2015). Digital transformation strategies. *Business & Information Systems Engineering*, 57(5), 339-343.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175-183.

Mehrpouya, A., & Salles-Djelic, M. L. (2019). Seeing like the market; exploring the mutual rise of transparency and accounting in transnational economic and market governance. *Accounting, Organizations and Society*, 76, 12-31.

Nambisan, S., Lyytinen, K., Majchrzak, A., & Song, M. (2017). Digital innovation management: Reinventing innovation management research in a digital world. *MIS Quarterly*, 41(1), 223-238.

Nissenbaum, H. (2010). Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford University Press.

OECD. (2021). State of implementation of the OECD AI Principles: Insights from national AI policies. *OECD Digital Economy Papers,* No. 311. OECD Publishing.

Olan, F., Arakpogun, E. O., Suklan, J., Nakpodia, F., Damij, N., & Jayawickrama, U. (2022). Artificial intelligence and knowledge sharing: Contributing factors to organizational performance. *Journal of Business Research*, *145*, 605-615.

Pan, Y. C., Jacobs, A., Tan, C., & Askool, S. (2018). Extending technology acceptance model for proximity mobile payment via organizational semiotics. In K. Liu, K. Nakata, W. Li, & C. Baranauskas (Eds.), *Digitalisation, Innovation, and Transformation: 18th IFIP WG 8.1 International Conference on Informatics and Semiotics in Organisations, ICISO 2018* (pp. 43-52). Springer International Publishing.

Power, M. (2015). How accounting begins: Object formation and the accretion of infrastructure. *Accounting, organizations and society*, *47*, 43-55.

Rivard, S. (2004). Information technology and organizational transformation: Solving the management puzzle. Routledge.

Roger, C., & Dauvergne, P. (2016). The rise of transnational governance as a field of study. *International Studies Review*, *18*(3), 415-437.

Saarikko, T., Westergren, U. H., & Blomquist, T. (2020). Digital transformation: Five recommendations for the digitally conscious firm. *Business Horizons*, 63(6), 825-839.

Sanderson, C., Douglas, D., & Lu, Q. (2023). Implementing responsible AI: Tensions and trade-offs between ethics aspects. arXiv preprint

Selander, L., & Jarvenpaa, S. L. (2016). Digital action repertoires and transforming a social movement organization. *MIS Quarterly*, 40(2), 331-352.

Shin, D. H. (2010). Modeling the interaction of users and mobile payment system: Conceptual framework. *International Journal of Human-Computer Interaction*, 26(10), 917-940.

Silva, P. (2015). Davis' technology acceptance model (TAM)(1989). *Information seeking behavior and technology adoption: Theories and trends*, 205-219.

Slade, E. L., Williams, M. D., & Dwivedi, Y. K. (2013). Mobile payment adoption: Classification and review of the extant literature. *The Marketing Review*, 13(2), 167-190.

Song, J., Lee, K., & Khanna, T. (2016). Dynamic capabilities at Samsung: Optimizing internal co-opetition. *California Management Review*, *58*(4), 118-140.

Svahn, F., Mathiassen, L., & Lindgren, R. (2017). Embracing digital innovation in incumbent firms: How Volvo Cars managed competing concerns. *MIS Quarterly*, 41(1), 239-253.

Tan, C., Abdaless, S., & Liu, K. (2018). Norm-based abduction process (NAP) in developing information architecture. In K. Liu, K. Nakata, W. Li, & C. Baranauskas (Eds.), *Digitalisation, Innovation, and Transformation: 18th IFIP*

WG 8.1 International Conference on Informatics and Semiotics in Organisations, ICISO 2018 (pp. 33-42). Springer International Publishing.

Tan, C., Liu, K., & White, E. (2013). Information architecture for healthcare organizations: the case of a NHS hospital in UK. In Proceedings of the 34th International Conference on Information Systems.

Toon, N. (2024). *How AI Thinks: How we built it, how it can help us, and how we can control it*. Random House.

UK Government. (2023). A pro-innovation approach to AI regulation: White Paper. Department for Science, Innovation and Technology.

Usman, F. O., Eyo-Udo, N. L., Etukudoh, E. A., Odonkor, B., Ibeh, C. V., & Adegbola, A. (2024). A critical review of ai-driven strategies for entrepreneurial success. *International Journal of Management & Entrepreneurship Research*, *6*(1), 200-215.

Vial, G. (2021). Understanding digital transformation: A review and a research agenda. In S. Nambisan (Ed.), *Managing Digital Transformation* (pp. 13-66). Springer.

World Bank Group. (2024). *Global trends in AI governance: Evolving country approaches*. Washington, DC: World Bank.

Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 34, 265-288.

Zhu, K., Dong, S., Xu, S. X., & Kraemer, K. L. (2006). Innovation diffusion in global contexts: determinants of post-adoption digital transformation of European companies. *European journal of information systems*, *15*(6), 601-616.

Surveillance and transparency as parallel systems of workplace analytics

Oliver G Kayas and Niki Panteli

Liverpool John Moores University, Lancaster University

Research in progress

Abstract

While surveillance and transparency have each received extensive attention on their own, there is a paucity of research integrating these concepts to produce a more nuanced analysis of their effects when deployed through workplace analytics. This developmental paper proposes a conceptual framework that integrates surveillance and transparency as parallel effects of workplace analytics in order to produce new and deeper insights into their impact on employee experience, and specifically on intra-organisational trust dynamics and employee engagement. Guided by the proposed conceptual framework, a future empirical study will be undertaken to examine the interplay between surveillance and transparency and their subsequent impact on intra-organisational trust and employee engagement.

Keywords: workplace analytics, surveillance, transparency, trust, engagement

1 Introduction

Despite workplace analytics being a nascent area of research (Coolen et al., 2023), studies have linked workplace analytics to two contrasting perspectives: surveillance and transparency (Viola & Laidler, 2022). On the one hand, workplace analytics can provide managers with a variety of surveillance measures, including digital recruitment tools to headhunt and filter candidates; telephone monitoring to assess call waiting time, idle time, and number of calls completed; and performance monitoring to generate actionable insights through the evaluation of employees' performance. These surveillance measures (and others) collect and analyse data related to employees and managers outside the organisation (e.g., demographic data, education, social network participation), their positions in the organisation (e.g., position status, salary, benefits, date of promotion), the work they have undertaken in the organisation (e.g., individual performance, performance evaluations, sentiments, message content), and the employees themselves (e.g., personality traits, cognitive abilities, skills, health, expertise, training completed) (Fernandez & Gallardo-Gallardo, 2020). Although these data can provide valuable insights to inform workplace decisions, research suggests that surveillance can reduce productivity, heighten employees' stress levels, undermine privacy, escalate acts of resistance, exacerbate counterproductive work behaviours, intensify opportunism among employees, and heighten employees' fear of management.

On the other hand, workplace analytics can provide insights on employees' behaviour, and their daily and routine practices, contributing to transparency and increased awareness of employees' practices that may need to be adjusted e.g., too much reliance on specific individuals while other key collaborators may be ignored. This transparency can ultimately lead to corrective action and a more positive employee experience (John et al., 2023). Transparency is a multifaceted concept that is viewed as a public value or norm of behaviour designed to counter corruption, inefficiency, and incompetence while also enhancing accountability to ensure organisational members behave adequately and appropriately through the act of being open (Meijer, 2009; Michener & Bersch, 2013). It is also portrayed as an antidote to the issues associated with workplace surveillance, ensuring organisations, managers, and employees comply with expectations and make informed decisions that evoke a sense of justice, responsibility, and fairness (Johnson & Regan, 2014). While surveillance and transparency have each received extensive attention on their own, the dynamic interplay between both has received little attention, with their relationship often taken for granted (Johnson & Regan, 2014; Viola & Laidler, 2022).

This developmental paper integrates surveillance and transparency as parallel systems of workplace analytics to propose a conceptual framework (Figure 1) that aims to produce new and deeper multi-level insights, while also answering previous calls to study their impact on employee experience and notably on intra-organisational trust dynamics as well as employee engagement (Kayas, 2023; Viola & Laidler, 2022). Moving forward, the conceptual framework will guide an empirical examination of how the surveillance and transparency embedded in workplace analytics affects intra-organisational trust dynamics and employee engagement. The next section presents and discusses the literature underpinning this study's proposed conceptual framework. The paper concludes by outlining how the project will move forward through an empirical investigation.

2 Theoretical foundations

2.1 Workplace analytics as a parallel system of surveillance and transparency

Workplace analytics turns insights into action by continuously monitoring and measuring the abilities, aptitudes, behaviour, health and fitness, performance, personal characteristics, personality traits, psychological disposition, sentiment, and skills of employees and managers to determine whether organisational expectations have been achieved (Fernandez (Fernandez & Gallardo-Gallardo, 2020). Although workplace analytics is a relatively new monitoring practice (Ball, 2021; Fernandez & Gallardo-Gallardo, 2020), it is becoming widespread within organisations being enabled by technological advancements and digital workplace transformation but also the popularity of alternative modes of work, such as remote and hybrid work (John et al., 2023). Within this context, employees may perceive the surveillance embedded in workplace analytics as an acceptable part of working life (Ball, 2021). They may even view it as a positive organisational practice if it benefits employees, informs decisions around remuneration and promotion, and exposes antisocial behaviour like favouritism (Kayas, 2023; Kayas et al., 2019). However, if the surveillance embedded in workplace analytics is perceived as too intensive or personalised (Ball, 2021; Sewell et al., 2012), violates boundaries by reaching into the personal lives of employees (Kayas, 2023), collects and analyses data beyond employees' behaviour and performance (Kayas et al., 2019; Sewell et al., 2012), then it can become a controversial organisational practice with negative implications for employees, managers, and organisations (Ball, 2021; Kayas, 2023).

In addition to providing organisations with the mechanisms needed to implement surveillance, workplace analytics also provides the means to implement transparency practices, with both surveillance and transparency producing accounts that are used to scrutinise the watched through information technology systems that collect data from internal and external sources to determine whether they are behaving as expected

(Michener & Bersch, 2013; Viola & Laidler, 2022). Despite their similarities, both have different rationales, with transparency becoming a mobilising idea for resisting or overcoming the negative consequences of surveillance. Defined as ‘the ability to look clearly through the windows of an institution’ (den Boer, 1998, p. 105), and the idea ‘that something is happening behind curtains and once these curtains are removed, everything is out in the open and can be scrutinized’ Meijer (2009, p. 258).’ The aphorism of transparency thus being ‘sunlight disinfects’ (Johnson & Regan, 2014).

Public debates have argued that improved transparency can induce better oversight and decisions, while restoring relations damaged through surveillance (Brin, 1998). This is achieved by exerting pressure on institutions, organisations, leaders, managers, and employees to ensure they behave as expected by their constituents (Johnson & Regan, 2014). By giving people better information that can be used to contribute to the rationalisation of organisations, transparency can reduce corruption, inefficiency, and incompetence, while also enhancing accountability, and opening institutions and organisations to ensure their members act adequately and appropriately (Meijer, 2009; Michener & Bersch, 2013). Proponents of transparency claim that those who are subjected to it are less likely to betray the trust of their constituents or neglect their responsibilities, while opponents claim that if transparency is unidirectional, unstructured, and decontextualised, then it will not benefit society, and could lead to a loss of trust by undermining freedom and threatening privacy (Meijer, 2009).

By developing a conceptual framework that integrates the surveillance and transparency embedded in workplace analytics, it provides a more comprehensive lens through which their impact on employee experience can be examined. This study focuses on two aspects of employee experience in the workplace that have been identified by existing research as vital for employee well-being, fulfilment, and performance (e.g., Chamakiotis et al., 2021; John et al., 2023) i.e., intra-organisational trust dynamics and the engagement of employees, both of which have been overlooked in the workplace analytics literature.

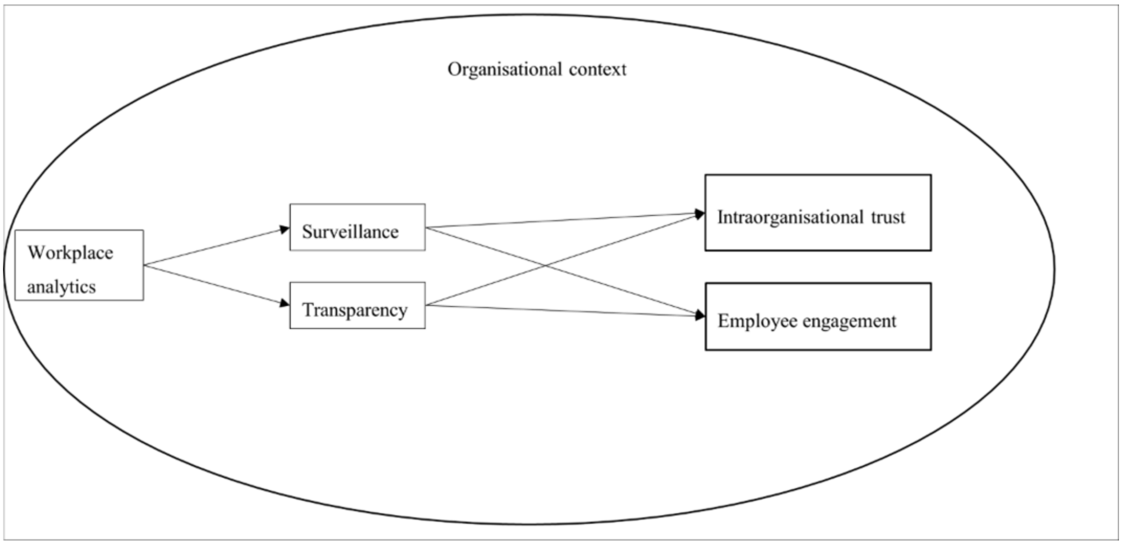


Figure 1. Conceptual framework

2.2 Employee engagement

Employees who are engaged in their work are said to have a sense of strong connection and identification with the work as well as the organisation (Gruman & Saks, 2011). Kahn (1990, p. 694) defined employee engagement as ‘the harnessing of organisation members’ selves to their work roles.’ Although there has been a growing literature on employee engagement (Bakker & Demerouti, 2008; Monje-Amor et al., 2020), research has predominantly focused on traditional and permanent work arrangements in collocated organisations. Alternative forms of work, such as virtual teams and hybrid work, are challenging conventional understandings of employee engagement due to varied employee experience and increased dependency on information and communication technologies that may affect employees’ connection and identification with the organisation (Panteli et al., 2019). John et al. (2023) has shown that employee engagement is likely to improve when work analytics are used in hybrid work context due to the transparency provided especially through insights on employees’ online behaviour and communication practices.

2.3 Intra-organisational trust

The concept of trust has received considerable attention among organisational researchers, leading to a confusing potpourri of definitions that have been applied to different units and levels of analysis; thus, making it a particularly difficult concept to define (Connell & Mannion, 2006). Mayer et al. (1995, p. 172) develop a widely held definition, asserting that trust is a psychological state in which there is a ‘willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.' Crucially, they highlight how the context within which a trustor perceives a relationship of trust can affect the need for trust and the evaluation of trustworthiness. A change in the sociopolitical context within an organisation or a perceived violation of the trustee can thus lead to the re-evaluation of trustworthiness. Should the introduction of an organisational control system provide managers with the means to deploy invasive surveillance measures, then this change in context could undermine trustworthiness, leading to a re-evaluation of intra-organisational relationships that may affect employees emotional state and consequently their productivity. Studies of trust thus necessitate an understanding of context to ascertain how it affects perceptions of trustworthiness.

3 Conclusions and implications for further research

This developmental paper proposes a conceptual framework that integrates the literatures related to the impact of the surveillance and transparency embedded in workplace analytics, including intra-organisational trust and engagement. Moving forward, this project will continue to develop the conceptual framework to ensure it produces novel and nuanced insights into workplace analytics. Ultimately, the conceptual framework will guide an empirical examination of the affects the surveillance and transparency embedded in workplace analytics have on intra-organisational trust and employee engagement, which we intend to present at the conference.

4 References

Bakker, A. B., & Demerouti, E. (2008). Towards a model of work engagement. *Career Development International*, *13*(3), 209-223. https://doi.org/10.1108/13620430810870476

Ball, K. (2021). *Electronic monitoring and surveillance in the workplace: literature review and policy recommendations*. Publications Office of the European Union.

Brin, D. (1998). *The transparent society: will technology force us to choose between privacy and freedom?* Perseus Books.

Chamakiotis, P., Panteli, N., & Davison, R. M. (2021). Reimagining e-leadership for reconfigured virtual teams due to Covid-19. *International Journal of Information Management*, *60*(102381). https://doi.org/10.1016/j.ijinfomgt.2021.102381

Connell, N. A. D., & Mannion, R. (2006). Conceptualisations of trust in the organisational literature: some indicators from a complementary perspective. *Journal of Health Organization and Management*, *20*(5), 1477-7266. https://doi.org/10.1108/14777260610701795

Coolen, P., van den Heuvel, S., Van De Voorde, K., & Paauwe, J. (2023). Understanding the adoption and institutionalization of workforce analytics: a systematic literature review and research agenda. *Human Resource Management Review*, *33*(4), 100985. https://doi.org/10.1016/j.hrmr.2023.100985

den Boer, M. (1998). Steamy Windows: Transparency and Openness in Justice and Home Affairs. In V. Deckmyn & I. Thomson (Eds.), *Openness and Transparency in the European Union* (pp. 91-105). European Institute of Public Administration.

Fernandez, V., & Gallardo-Gallardo, E. (2020). Tackling the HR digitalization challenge: key factors and barriers to HR analytics adoption. *Competitiveness Review*, *31*(1), 162-187. https://doi.org/10.1108/CR-12-2019-0163

Gruman, J. A., & Saks, A. M. (2011). Performance management and employee engagement. *Human Resource Management Review*, *21*(2), 123-136. https://doi.org/10.1016/j.hrmr.2010.09.004

John, B., Zeena, A. i., & Panteli, N. (2023). Enhancing Employee Experience in the Era of Hybrid Work: The Case of Microsoft Viva. *IEEE Software*, *40*(2), 70-79. https://doi.org/10.1109/MS.2022.3229956

Johnson, D. G., & Regan, P. M. (2014). Introduction. In D. G. Johnson & P. M. Regan (Eds.), *Transparency and surveillance as sociotechnical accountability: a house of mirrors* (pp. 1-24). Routledge.

Kahn, W. A. (1990). Psychological conditions of personal engagement and disengagement at work. *Academy of Management Journal*, *33*(4), 692-724. https://doi.org/10.2307/256287

Kayas, O. G. (2023). Workplace surveillance: A systematic review, integrative framework, and research agenda. *Journal of Business Research*, *168*, 114212. https://doi.org/10.1016/j.jbusres.2023.114212

Kayas, O. G., Hines, T., McLean, R., & Wright, G. H. (2019). Resisting government rendered surveillance in a local authority. *Public Management Review*, *21*(8), 1170-1190. https://doi.org/10.1080/14719037.2018.1544661

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709-734.

Meijer, A. (2009). Understanding modern transparency. *Review of Administrative Sciences*, *75*(2), 255-269. https://doi.org/10.1177/0020852309104175

Michener, G., & Bersch, K. (2013). Identifying transparency. *Information Polity 18*, 233-242. https://doi.org/10.3233/IP-130299

Monje-Amor, A., Vázquez, J. P. A., & Faíña, J. A. (2020). Transformational leadership and work engagement: exploring the mediating role of structural empowerment. *European Management Journal*, *38*(1). https://doi.org/10.1016/j.emj.2019.06.007

Panteli, N., Yalabik, Z. Y., & Rapti, A. (2019). Fostering work engagement in geographically dispersed and asynchronous virtual teams. *Information Technology & People*, *32*(1), 2-17.

Sewell, G., Barker, J. R., & Nyberg, D. (2012). Working under intensive surveillance: When does 'measuring everything that moves' become intolerable? *Human Relations*, *65*(2), 189-215. https://doi.org/10.1177/0018726711428958

Viola, L. A., & Laidler, P. (2022). On the relationship between trust, transparency and surveillance. In L. A. Viola & P. Laidler (Eds.), *Routledge Studies in Surveillance: Trust and transparency in an age of surveillance* (pp. 3-18). Routledge.

Beyond Regulatory Compliance: Towards a multi-layered responsible cybersecurity perspective

Boineelo R Nthubu
Lancaster University, b.nthubu1@lancaster.ac.uk

Niki Panteli
Lancaster University, n.panteli1@lancaster.ac.uk

Konstantinos Mersinas
Lancaster University, Konstantinos.Mersinas@rhul.ac.uk

Research In progress

Abstract

The paper develops a "multi-layered responsible cybersecurity" perspective as a holistic and collective approach to protecting people, organisations, supply chains and societies. This perspective posits that responsible cybersecurity extends beyond regulatory compliance, to the extent that it encompasses different layers of responsibilities that span across techno-centric, human-centric, organisational (intra and inter) and societal perspectives. Our theoretical development emerges from raw data through an exploratory study that involved qualitative interviews with senior cybersecurity professionals and consultants. The "responsible cybersecurity" perspective generates significant implications. First, it has implications for cybersecurity research in that it provides an integrative and balanced approach to viewing the multiple and diverse stakeholders who might be impacted by potential attacks that expand beyond regulations and the organisation. Second, it has implications for digital responsibility research in that responsible cybersecurity can be viewed from different layers each exposing different stakeholders who may be affected as well as different responsibilities.

Keywords: Cybersecurity, Compliance, Regulation, Responsible, Exploratory Study

1.0 Introduction

With growing digitalisation and the acceleration of digital transformation, cybersecurity attacks are becoming increasingly prominent and are a constant threat to individuals, organisations and societies at large. Such incidents do not just cause unnecessary disruptions to organisations and their business operations, but they also contribute to huge financial and reputational costs to the organisations involved (Safa et al., 2016) and society more widely (Agrafiotis, Nurse et al. 2018). Moreover, cybersecurity is not limited to organisations' employees and end-users, but relates to essentially every individual in digitally advanced societies. A privacy violation of a single individual's personal data can have devastating effects for the wellbeing of that person; a leak of confidential information or a denial of service due to a ransomware attack can have catastrophic consequences for a company and its employees; consequently, the exploitation of a vulnerability in any of the healthcare, transportation, energy, financial etc. sectors can have catastrophic societal impacts.

Although regulations are needed in order to compel organisations to protect their data and systems from cyber attacks, some cyberattacks are not covered by regulations. Srinivas et al., (2019) reviewed cybersecurity regulations adapted by several federal governments and highlights that there is a tendency to focus on specific industries such as healthcare, homeland security, finance etc. Further, regulations may not be enough to protect an organisation and its business partners in the same supply chain.

Accordingly, in this paper, we posit for the need for developing an understanding of cybersecurity from a responsible perspective, one that goes beyond being compliant. The need for responsible cybersecurity derives from an increased realisation that cyber threats and attacks have implications beyond the individuals and organisations that may be directly affected and impact societies at large. Following from these, the driving question of the study is '*How can organisations move beyond regulatory compliance to a responsible and more holistic cybersecurity perspective?*

In what follows we review relevant literature on the role of compliance in cybersecurity and present literature on responsible digital and the responsible perspective more generally. Following this, we present the research design and methodology of the empirical study and the analytical approach adopted. We then show the findings to-date and broadly discuss the implications of the study.

2.0 Literature Review

2.1 Cybersecurity - Beyond Compliance

Due to the interconnectedness and dependencies on other organisations' services, cybersecurity is no longer a concern limited to isolated organisations or specific sectors. Instead, it impacts every layer of society, from individuals and organisations to supply chains and the society at large. Despite the interconnectedness and spread impact, regulations tend to focus on organisations' internal responsibilities within their sector even though e,g. third party risks come from beyond the sector (Didenko, 2020). Further, Srinivas et al (2019)'s review of federal government regulations in cyber security has shown that regulations leave out some sectors. These regulation challenges means that compliance can leave gaps in security and organisations cannot rely on compliance alone to protect themselves and their supply chains. A compliant organisation may still have vulnerabilities that aren't covered by regulatory standards (Marotta and Madnick, 2020). Although essential, compliance can be incomplete when it comes to providing comprehensive protection (Harris and Martin 2021). Additional proactive measures are needed beyond what regulations mandate (Harris and Martin, 2021). In light of this, Didenko, (2020) highlight the necessity for what they call a "*cross-sectoral"* cybersecurity framework to address risks that extend beyond industry sectors. Similarly, Tropina and Callanan, (2015) emphasise the importance of self-regulation to augment the limitations of regulations in cybersecurity. Together, these studies emphasise the importance of moving beyond compliance which has an "organisation" focus, towards a holistic framework that encompasses different layers and scope of responsibilities.

2.2 The Responsible Perspective

Researchers interested in the promoting a responsible perspective in the digital era often highlight ethical concerns such as the reinforcement of existing bias, lack of transparency whilst proposing an urgent need for regulation (Trocin et al, 2023). Within this body of literature, digital responsibility is viewed as the ethical and accountable use of digital technologies, including, ethical decision making, online behaviour, and protecting one's privacy and security (Zhang and Hon, 2020) with other researchers referring to a perspective that is ethical, sustainable, and respectful of human values and society (Pappas et al., 2023). Reasons for adopting a responsible perspective includes a need to promote fairness and equality in the design, implementation and use of digital technologies (Trocin et al, 2023), as well as a need to minimize any potential negative impacts on users' wellbeing and the society in general (Dignum, 2019), and benefit multiple-stakeholders (Pappas et al. 2023). Pappas et al (2023) argue that digital initiatives need to be designed and implemented in a way that benefits multiple-stakeholders. They posit that the value of such digital initiatives should be both co-created and shared. The expectation is that when digital (and other) initiatives are built on responsible principles negative outcomes are avoided whilst individuals, organisations and societies experience great positive impacts (Dignum, 2019).

3.0 Research Methodology

We carried out a series of semi-structured interviews with cyber leaders and other members of the senior management team to explore understandings and attributes of responsible cybersecurity and the role of the organisation and cyber leaders in promoting this cybersecurity perspective. Interviewees were encouraged to share their understanding of responsible cybersecurity and contribute towards the co-design of a framework for fostering a responsible cybersecurity mindset. Sample interview questions included: *In your view, what are the fundamental principles (dimensions) that responsible cybersecurity should encompass? What challenges does your organisation face in adhering to the fundamental principles of responsible cybersecurity? Are there specific opportunities or best practices that contribute to fostering a responsible cybersecurity approach?*

In total, 20 interviews were conducted and included 15 male and 5 female participants who held leadership positions in cybersecurity e.g Directors, Chief Information Security Officer (CISO) etc. The participants were chosen because of their responsibility in managing and directing cybersecurity operations. Our participants represented a range of sectors including finance, IT, transport, consultancy and government. Their experience in the cybersecurity sector varied from 5 to 30 years. The duration of the interviews which took place online (via Teams) was between 25 to 70 minutes and they were all audio recorded and transcribed. NVivo was used for data analysis as it enabled systematic coding, organisation and retrieval of data. We followed the approach outlined by Gioia, Corley, & Hamilton (2013) to discover the dimensions of responsible cybersecurity, which we later named "layers", and how to foster responsibility in each layer.

4.0 Findings

The findings reveal that responsible cybersecurity requires a collective commitment where all stakeholders act as stewards, not only of their data but also of their supply chain and the broader wellbeing of individuals and society. The interviews highlighted five core layers of responsibility: **techno-centric**, focusing on technological defenses; **human-centric**, emphasising security solutions designed with users in mind and safeguarding the wellbeing of security professionals; **intra-organisational**, stressing the role of team collaboration and leadership buy in, in promoting a strong security culture; **inter-organisational**, concerning the security of supply chains and third-party partners; and **societal**, recognising the ethical implications of security solutions on a broader societal scale. This multi-layered approach emphasises the scope of responsibilities beyond the organisation and compliance.

5.0 Implications and Potential Contributions

We are currently at the stage of further analysing the data. At the time of writing, we expect at least two theoretical contributions to derive from the study: First, it provides an integrative and balanced approach of not only different views that can be represented in the responsibility domain but also the multiple and diverse stakeholders who have an interest in cybersecurity and who may be affected by potential attacks. This approach confirms that compliance is not enough for ensuring robust cybersecurity in organisations. Second, it expands literature on responsible digital and digital responsibility by showing that responsible cybersecurity can be viewed from different layers each exposing different stakeholders who may be affected as well as different responsibilities. Further, the study provides opportunities for practical implications and in particular for decision-makers and organisational leaders who can be encouraged to identify security practices for not just their own organisation but for peer organisations, entities in the supply chain, and the broader security ecosystem.

6.0 Future Research Directions

While the current study has made significant progress in exploring responsible cybersecurity, several key research activities remain to be undertaken to enhance the understanding of responsible cybersecurity. First, we plan to conduct a co-participatory workshop with senior cybersecurity professionals to further develop a responsible cybersecurity framework. Following this, additional qualitative coding and validation of emerging themes will be necessary to refine the resulting framework.

References

Agrafiotis, I., Nurse, J. R., Goldsmith, M., Creese, S., & Upton, D. (2018). A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *Journal of Cybersecurity*, *4*(1), 1-15.

Dignum, V. (2019). Responsible Artificial Intelligence*: How to Develop and Use AI in a Responsible Way*. Springer International Publishing. https://doi.org/10.1007/978-3-030-30371-6.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational research methods*, *16*(1), 15-31.

Harris, M. A., & Martin, R. (2019). Promoting cybersecurity compliance. In *Cybersecurity education for awareness and compliance* (pp. 54-71). IGI Global.

Pappas, I. O., Mikalef, P., Dwivedi, Y. K., Jaccheri, L., & Krogstie, J. (2023). Responsible Digital Transformation for a Sustainable Society. Information Systems Frontiers, 1-9.

Marotta, A., & Madnick, S. (2020). Analysing the interplay between regulatory compliance and cybersecurity (Revised).

Trocin, C., Mikalef, P., Papamitsiou, Z. *et al.* Responsible AI for Digital Health: a Synthesis and a Research Agenda. *Inf Syst Front* 25, 2139–2157 (2023). https://doi.org/10.1007/s10796-021-10146-4

Tropina, T., & Callanan, C. (2015). *Self-and co-regulation in cybercrime, cybersecurity and national security* (p. 25). Heidelberg: Springer.

Safa, N. S., & Von Solms, R. (2016). An information security knowledge sharing model in organizations. *Computers in Human Behavior*, *57*, 442-451.

Safa, N. S., Von Solms, R., & Futcher, L. (2016). Human aspects of information security in organisations. *Computer Fraud & Security*, *2016*(2), 15-18.

Srinivas, J., Das, *A*. K., & Kumar, N. (2019). Government regulations in cyber security: Framework, standards and recommendations. *Future generation computer systems*, *92*, pp.178-188.

Zhang, J., & Hon, H. W. (2020). Towards responsible digital transformation. *California Management Review Insights*.

Environnemental Impacts of Blockchain for Patient Consent

Isabelle BOURDON
MRM, Université de Montpellier, Montpellier, France
Negar ARMAGHAN
MRM, Université de Montpellier, Montpellier, France
Sylvain MARTINEZ
Scalian, France

Completed Research

Abstract

The aim of this paper is to measure the environmental impact of an innovative solution using blockchain in healthcare. Thanks to this technology, new healthcare services such as the traceability of patient consents are being developed with the aim of reducing costs and improving quality of care. However, despite their virtual nature, digital technologies, and blockchain technologies in particular, are based on physical infrastructures that generate significant externalities for the environment and consume a lot of energy during their manufacture, use and end-of-life. These externalities are one of the main criticisms of these technologies. As part of a case study, we carried out a comparative Life Cycle Assessment of two solutions for managing patient consent in clinical trials. In this article, we establish that the proposed innovative solution based on blockchain technology has an average reduction in environmental impact of 44% compared with the current solution, while performing the same function.

Keywords: Blockchain, Environmental Impact, Life Cycle Assessment (LCA), Health, Patient Consent, Sustainability

1.0 Introduction

IS plays a crucial role in addressing environmental challenges by enabling organizations to implement sustainable practices reducing greenhouse gas emissions, optimizing resource consumption (Seidel et al., 2017). It can contribute by automating processes, enhancing data-driven decision-making, facilitating organizational transformation, providing standardized and scalable infrastructures (Dao et al., 2011). Environmental sustainability is the first most well-known sustainability outcome in IS literature, especially to reduce the use of natural resources and adopting measures that improve the planet's general health (Melville, 2010).

Elkington (Elkington, 1994, 2004) defined sustainability by three components: natural environment, society and economic performance which together form the Triple

Bottom Line (TBL). As illustrated in Figure 1, the TBL approach suggests that organizations should go beyond economic performance and actively engage in initiatives that positively impact both society and environment (Dao et al., 2011).

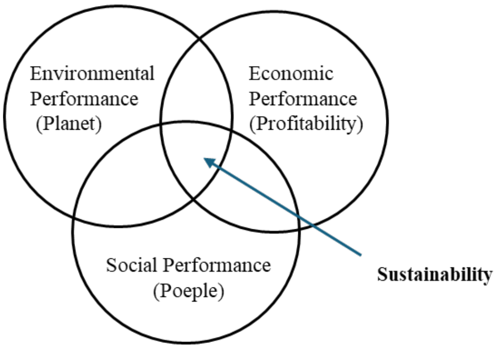

Environmental Performance (Planet)
Economic Performance (Profitability)
Social Performance (Poeple)
Sustainability

Figure 1. The triple bottom line of sustainability

Some research has demonstrated that long-term profitability and organizational sustainability are best achieved when firms balance economic goals with social and environmental responsibilities (Porter & Kramer, 2007). A shift in focus is needed, using IS scholars to consider environmental impact alongside traditional measures of efficiency and effectiveness. Emphasizing the importance of interdisciplinary collaboration, engagement with policymakers, and the broader role of IS research, sustainability should be integrated as a fundamental aspect rather than being treated as a niche topic. However, the research community has not yet fully integrated sustainability as a core element in IS studies (Seidel et al., 2017).

As with all digital technologies blockchain raises concerns regarding its environmental impact. Faced with these growing impacts, the link between IS and environmental sustainability development (Melville, 2010; Watson et al., 2010) is increasingly mobilized to respond to these challenges.

In environmental applications, blockchain is used for tracking sustainable supply chains, trading renewable energy, and automating climate-related transactions. It promotes the development of renewable energy by facilitating energy exchanges between producers and consumers through smart contracts, eliminating intermediaries and improving market efficiency. It plays a key role by facilitating the recording and verification of transactions related to ecological initiatives (Arshad et al., 2023).

A systematic literature review was conducted to analyse blockchain's role in ESG[1] (Environmental, Social, and Governance) and sustainability policies. While blockchain offers potential for transparency and efficiency, challenges remain in policy alignment, energy consumption, and standardization. Blockchain holds promises for sustainability, energy efficiency, and transparent supply chains, but policy alignment, energy efficiency, and standardization remain major challenges (Mulligan et al., 2024). The study of Fan (Fan et al., 2022, p. fan) examines the impact of blockchain technology on supply chains, considering consumer awareness of product traceability. It analyzes two scenarios: one where blockchain is adopted and one where it is not, focusing on optimal pricing strategies and conditions for adoption. The findings indicate that blockchain enhances transparency and consumer trust, but its implementation depends on factors such as operational costs, cost-sharing among supply chain members, and consumer awareness levels.

Furthermore, blockchain technology has emerged as a transformative IS tool, particularly in healthcare (Sadeghi R. et al., 2022). Therefore, the relevance of blockchain in data privacy and consent management is another raising topic in IS. Given the constant evolution of privacy regulations, blockchain offers solutions for managing consent and protecting user data (Anderson et al., 2023; Kakarlapudi & Mahmoud, 2021). One of its applications is securing and enhancing the confidentiality of Electronic Health Records (EHR), which is a major challenge in medical data sharing (Sadeghi R. et al., 2022).

Blockchain plays a crucial role in enhancing EHRs and patient consent management by providing a secure, decentralized, and transparent framework for data exchange. Unlike traditional EHR systems, which often suffer from interoperability issues and security breaches, blockchain enables seamless access to medical records while ensuring data integrity and patient ownership. Patients can control who accesses their health information through smart contracts, allowing real-time consent management without intermediaries. This not only enhances privacy and security but also improves healthcare efficiency by reducing administrative burdens and enabling accurate,

[1] Policymakers are introducing mandatory ESG reporting standards, including (Mulligan et al., 2024).

- European Sustainability Reporting Standards (ESRS)
- International Sustainability Standards Boards (ISSB)
- SEC climate related disclosure rules (USA)
- Corporate Sustainability Reporting Directive (CSRD)

This study shows that despite blockchain's potential, few studies align solutions with these regulatory frameworks:

tamper-proof medical histories across multiple healthcare providers (AbdelSalam, 2023)

Despite its potential benefits, few studies have explored the environmental impacts of blockchain in health sector (Medaglia & Damsgaard, 2020; Zhang et al., 2020), while blockchain presents advantages in data security and transparency, its environmental impact, particularly in the healthcare sector, remains underexplored.

Two main lines of reflection are emerged in the literature (Kotlarsky et al., 2023): IS for green (leveraging digital technologies to reduce the environmental impact of other industries) and green IS (developing of digital technologies with lower environmental impact). We are part of the reflections on Green IS in this paper. In their editorial, Kotlarsky et al. (2023) develop the concept of digital sustainability, emphasizing the role of IS in enhancing environmental, social and economic welfare. In this paper we are focused more on environmental impact of blockchain in healthcare. Therefore, given the growing adoption of blockchain in healthcare and its potential environmental implications, this research addresses the following question:

How does the use of blockchain for patient consent management in clinical trials impact environmental sustainability, particularly in terms of energy consumption? To explore this question, we conducted a unique case study in France on a project called Consent Chain. Through observations and interviews, we carried out an inventory of patient consent collection practices in the context of clinical trials. We then developed a prototype of an innovative solution within the framework of design science methodology (detailed phases not presented in this paper). Finally, we conducted a Life Cycle Assessment (LCA) to evaluate the environmental impact of the innovative solution based on blockchain.

The paper is structured as follows: Section 1 presents the conceptual framework relating to patient consent challenges and the contribution of blockchain. This section also presents how to assess the environmental impacts of digital services through LCA. Section 2 details the methodology adopted in this study. The final section summarizes the findings and discusses implications.

2.0 Theoretical Background

The purpose of this section is to provide a brief overview of the different conceptual frameworks used for the case study. We present patient consent issue and the contributions of blockchain technologies to ensure better traceability.

2.1 Patient Consent and Health Records

Issue of patient consent's traceability for medical procedures and more particularly for clinical trials is crucial for healthcare institutions, in particular due to several strong regulatory constraints (Goldsmith et al., 2008). In accordance with the Public Health Code, a person participating in a clinical trial must provide written consent. The notification of this consent is generally carried out in the institution's information system, through several steps including the printing of paper sheets and their scanning. This task is time-consuming and is often carried out with a time gap (digitization of consent forms completed and signed by patients). The digitalization of records allows to create further opportunities for analyzing medical trends and evaluation of the quality of care, relevant technologies are still emerging, like blockchain, which can offer various benefits for the support service (Prokofieva & Miah, 2019). Specific consent for use of the data is also required. The clinical study is conducted by a sponsor or a Clinical Research Organization (CRO). For each sponsor or CRO, method used for consent process (collection, storage, transmission of data) are different. As a result, healthcare staff involved in these trials must manage multiple different processes and supports. While records are digitized, they are isolated in locally centralized data storages which present a strong impediment to further developments.

Various studies have shown that the fulfilment of patient's consents and their traceability are not always carried out, or cannot be proven (Benchoufi et al., 2017). The U.S. Food and Drugs Administration (FDA) reported that 53% of patient records involved in clinical research did not include full informed consent (Seife, 2015). Recent studies have shown the usefulness and feasibility of blockchain in solving these problems (Kakarlapudi & Mahmoud, 2021).

2.2 Blockchain Technologies For E-Health and Sustainability

Blockchain technologies are gaining increasing attention from both practitioners and academics. Many applications exist in healthcare for medical records and data management, drug traceability, or research and clinical trials (Madhoun et al., 2024).

Blockchain is a duplicated and shared ledger of transactions (data), consisting of blocks, each containing the record of all exchanges made between its participants. The data is encrypted and organized into blocks, with each block cryptographically linked to the previous block (Dinh et al., 2018; Petre & Haï, 2018). Blockchain, originally designed for cryptocurrencies, offers advantages such as enhanced security, transparency, and efficiency in data management. While blockchain technique has been known for several years in financial sector, recently, researchers and practitioners have become interested in the applications of blockchain in healthcare sector, especially for consent management (Kakarlapudi & Mahmoud, 2021). This technology offers several advantages, including improved data security, interoperability, transparency, and efficiency. It optimizes medical record management, reduces costs, minimizes fraud, and strengthens pharmaceutical supply chains by ensuring traceability and authentication of medications (Haleem et al., 2021). The potential applications of blockchain technology were developed in the healthcare sector through a systematic literature review of studies published between 2008 and 2019 (Prokofieva & Miah, 2019). This study identifies key applications, including secure medical data storage and sharing, smart contracts for administrative automation, improved drug supply chain traceability, and enhanced integrity in clinical trials. Implementation of blockchain improves the connectivity while security and privacy and reduces the costs are also ensured for the parties (Prokofieva & Miah, 2019). Blockchain-based systems can increase patient trust, reduce costs, and improve disease management. It plays a positive role in empowering patients and facilitating the sharing of electronic medical records and could enhance patient engagement while reducing healthcare sector costs (Hajian et al., 2023).

Despite its promise, challenges such as privacy concerns, technical complexity, scalability issues, and resistance to adoption remain significant barriers (Prokofieva & Miah, 2019).

Blockchain make it possible to automate consents, through smart contracts (Coiera & Clarke, 2004; Dinh et al., 2018) Genestier et al.(2017) provide use cases for blockchain-based consent management system. The use of this technology raises questions, concerning ecological cost (Truby, 2018), because validating transactions and saving ledger requires a lot of energy in the context of a public blockchain. However, few research have carried out with their environmental impacts, even if recent work has talked about sustainable blockchain technologies (Khurshid et al., 2023).

3.0 Research Project and Methods

To explore the environmental impacts of blockchain solutions, we conducted a unique case study. We were particularly interested in the use of blockchain to optimize the traceability of patient consents. The Consent Chain project, called "CC", is a project led by a university and a consortium of partners in France. It has been started in 2019, and the LCA was carried out in 2023.

3.1 Consent Chain Project

CC is a project that aims to design, build, and test a solution based on a blockchain, allowing to optimize the traceability of patient consents in a transparent and secure environment by supporting the control of data by citizens (patients) while ensuring access control, according to health sector regulations in France (RPPS identifiers of healthcare professionals, National Health Identifiers (NHIs) for citizens). The innovative blockchain-based solution was designed through a process of Action Design Research (this stage of the case study will not be described in the paper), which is a methodology that solves practical problems and creates artifacts. The first step of problem formulation allowed the analysis of the existing situation through interviews and observations with stakeholders involved in the patient consent process during clinical trials. This phase made it possible to describe in detail the different steps or actors involved in the process and to identify the weaknesses of the existing process to propose the innovative solution. The second phase of construction, intervention, and evaluation made it possible to conceptualize the value propositions for each actor, and to formalize the target process proposed by the solution. During this phase, finally, the technical analysis of the solution was carried out, the mock ups modelled, and a prototype was developed.

3.2 Life Cycle Assessment (LCA) Methodology

To assess the environmental impact of an activity or a product, many methods exist: quantitative and mono-criteria methods that evaluate greenhouse gas emissions, such as the French Carbon® Footprint, or the Water Footprint, qualitative methods, such as energy labels or labels, and Life Cycle Assessment (LCA) method, which is both quantitative and multi-criteria. Used since the 1990s, it has been standardized in the ISO 14040:2006 and ISO 14044:2006 series of standards (ISO, 2006a, 2006b) . This method makes it possible to measure the ecological impact of a product or service throughout the life cycle. Melville (Melville, 2010) first raised how LCAs can benefit IS researchers for measuring environmental impacts of IS use. The LCA literature has produced a rich contextual understanding of how to assess the externalities of a product or service (Finnveden et al., 2009). LCA is an approach that makes it possible to measure the impact of product‘s or service's externalities on environmental indicators (global warming, resource depletion, air pollution, water pollution, ecotoxicity,...) by integrating general impacts at all stages of the equipment life cycle, from the extraction of raw materials for the production of the technology, to the end-of-life of the technology, including energy consumption during the use phase. According to ISO standards, there are four steps that must be carried out during LCA : (1) definition of goal and scope, (2) inventory analysis, (3) impact assessment and (4) interpretation (ISO, 2006a). Many impact calculation methodologies have emerged since the creation of LCA (CML, Recipe, Impact, etc.). To standardize the approach and results, the European Commission has decided to develop a hybrid methodology based on expert recommendations. This started with a first version named ILCD (Wolf et al., 2011), followed in 2021 by the Environmental Footprint (European Commission, 2021).

The purpose of the analysis is to compare the environmental impact of the existing solution, and the solution developed by blockchain to ensure the function of collecting, processing, and storing patient consents in the context of clinical trials. We carried out a functional LCA to compare different technical solutions. We describe the characteristics of LCA: The functional unit is the unit of measurement used to evaluate the service provided. Here, we consider the consents collected during 1 year in France: 2600 clinical trials and 420,000 patients (reference year 2021). The perimeter is France. The duration of the analysis is one year.

First, a life-cycle inventory was carried out: for each solution, data are collected from the project (such as the number of clinical trials/year, or the number of A4 sheets per patient), others are calculated (such as the annual consumption of storage in kwh, or the consumption of scanning, consumption of one PC per day) and hypotheses are made (such as the distances of travel made by the investigator centers, the percentage of trips made by train or car, the length of time consent scans are stored, the number of servers per rack). Then, for each equipment that ensures the function, environmental impacts are modelled using Simapro software and Ecoinvent database. Each environmental indicator is quantified and characterized. The selected impact indicators are presented in the following table, from Environmental Footprint methodology, in its 3rd version.

Impact Indicators	Unit
Climate change	kg CO2 eq
Ozone depletion	kg CFC11 eq
Ionising radiation	kBq U-235 eq
Photochemical ozone formation	kg NMVOC eq
Particulate matter	disease inc.
Human toxicity, non-cancer	CTUh
Human toxicity, cancer	CTUh
Acidification	mol H+ eq
Eutrophication, freshwater	kg P eq
Eutrophication, marine	kg N eq
Eutrophication, terrestrial	mol N eq
Ecotoxicity, freshwater	CTUe
Land use	Pt
Water use	m3 depriv.
Resource use, fossils	MJ
Resource use, minerals and metals	kg Sb eq

Table 1. Impact Indicators and Units of Measurement

The environmental impact on each indicator corresponds to the impact of each modelled data, each piece of equipment considered, and this on all stages of the life cycle (extraction of raw materials to management at the end of life).

4.0 Findings

We present the results of the LCA and compare the current paper-based solution with the blockchain-based solution.

4.1 Results for Current Paper-Based Solution

We present the results related to the environmental impacts of the current solution based on paper consent forms. To do this, a description of the current system is presented, integrating assumptions made regarding the solution, we then describe the contribution measures on analyzed indicators and finally present the actions to be carried out that would limit these impacts.

Description of the system of the current paper solution

To conduct LCA, a description of the current one-year solution is summarized below:

- The time horizon studied is 1 year, the geographical area "studied" is France
- There are 2600 clinical trials per year, which corresponds to 420,000 patients (reference year 2021).
- There are 3 investigator centers per clinical case
- 3 trips per clinical case are mandatory on the part of the investigator centers
- 2 people are needed per trip
- Some of these journeys are made by car (60% hypothesis), the others are made by train (40%)
- An average car trip corresponds to 760 km round trip (assumption)
- An average train trip corresponds to 860 km round trip (assumption)
- 3 paper copies are required per patient (1 for the patient, 1 for the CHU, 1 for the sponsor). Each copy contains 2 sheets recto/verso
- Each copy is scanned and stored after the patient's signature. The legal storage period is 10 years
- PCs and servers are not considered as they are not used exclusively for this activity

Contribution analysis of the current paper solution

The results of the LCA on the current solution are presented below:

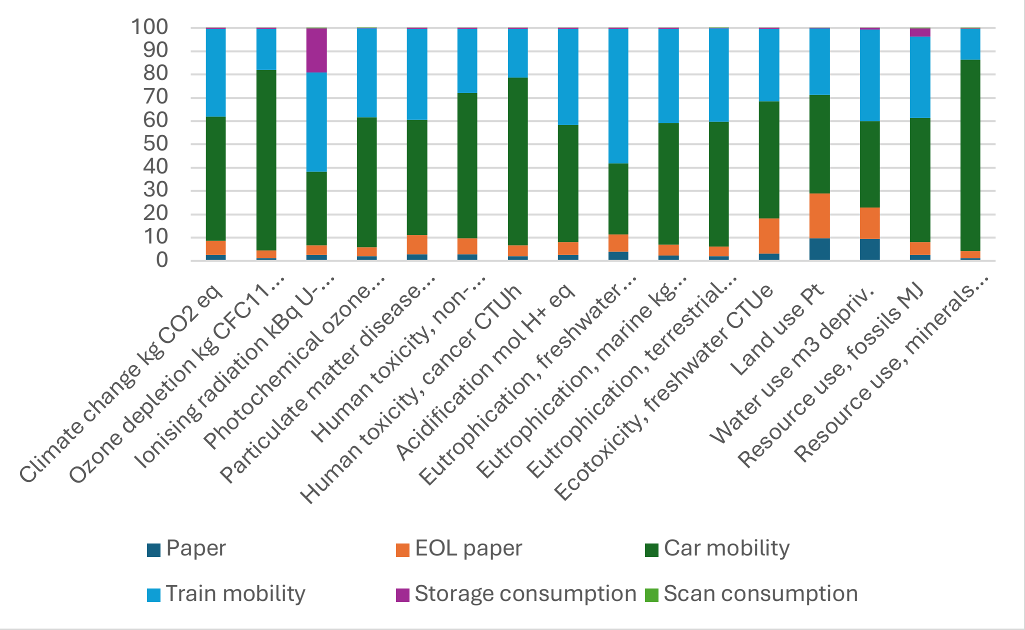


Figure 1. Contribution Analysis - Current Solution

- Car trips made by investigators have the strongest environmental impact on 13 indicators (16 indicators in total): The average contribution of these car trips, to the total environmental impact of the paper solution, is 53.4%. Train travel by investigators has the strongest environmental impact on 3 indicators (16 indicators in total): The average contribution of these train journeys to the total environmental impact of the paper solution is 34.4%
- Papermaking has a relatively small contribution to the total environmental impact of the paper solution (average contribution of 7.2%). End-of-life paper has a very small contribution to the total environmental impact of the paper solution (average contribution of 3.3%). The consumption of paper scanning has a very small contribution to the total environmental impact of the paper solution (average contribution of 0.03%)

To reduce the environmental impacts of the current patient consent traceability solution, the following actions are proposed: Use recycled paper instead of virgin paper and use renewable electricity for servers storing data. Several parameters are imposed by regulatory constraints (number of copies per patient, number of trips, etc.). Therefore, they can't be the subject of proposals for improvement at this time.

After measuring the environmental impacts of the current paper solution regarding the traceability function of patient consents in clinical trials, the LCA have been conducted on the innovative solution proposed in the CC.

4.2 Results for Innovative Blockchain-Based Solution

We present the results related to the environmental impacts of the innovative solution based on a blockchain technology. To do this, a description of the innovative system is presented, then we describe the contribution measures on the analyzed indicators and present the actions to be carried out that would limit these impacts. To conduct the LCA on the innovative solution, we present a one-year description of the blockchain solution.

Description of the innovative solution system

- The time horizon studied is 1 year, the geographical area "studied" is France
- There are 2600 clinical trials per year, which corresponds to 420,000 patients (reference year 2021) There are 3 investigator centers per clinical case
- 3 trips per clinical case are mandatory on the part of the investigator centers, but once every 2 years
- 2 people are needed per trip
- Some of these journeys are made by car (60% hypothesis), the others are made by train (40%)
- An average car trip corresponds to 760 km round trip (assumption)

- An average train trip corresponds to 860 km round trip (assumption)
- 20 PCs are needed 24 hours a day to produce the blockchain
- 20 servers of small size, are needed to run the blockchain solution. These 20 servers are 100% dedicated to this blockchain solution, they are not shared

The legal storage of the data is 10 years. In this study, we will test 2 scenarios, legal storage of 10 years, and possible storage of 50 years. The results of the LCA on the innovative solution are presented below, first for a consents' storage over 10 years and over a long period of 50 years:

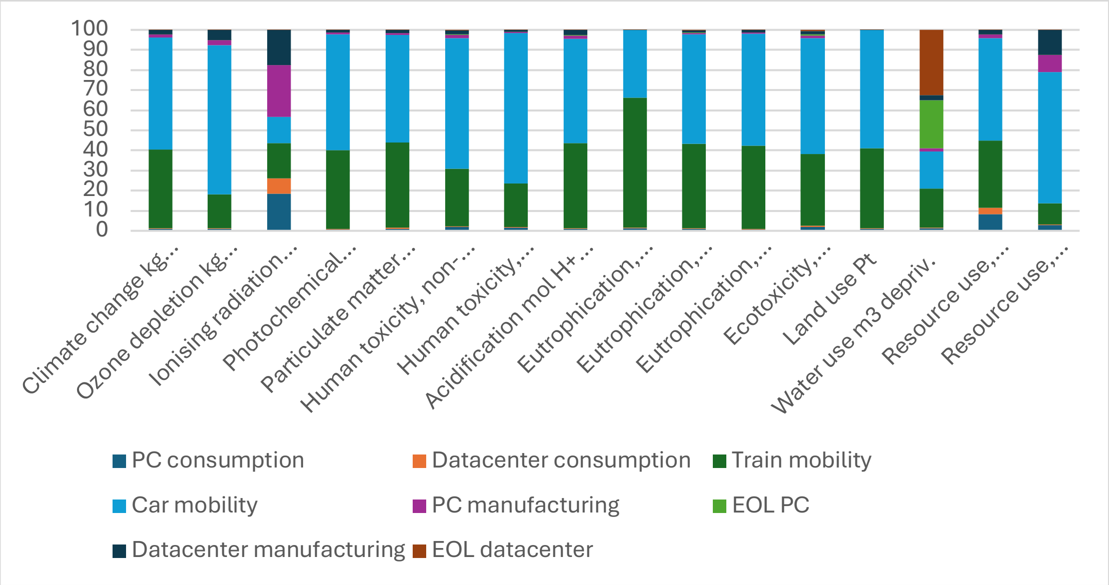


Figure 2. Contribution Analysis - 10-year innovative solution

- Car trips made by investigators have the strongest environmental impact on 13 indicators (16 indicators in total): The average contribution of these car trips to the total environmental impact of the paper solution is 52.5%. Train travel by investigators has the highest environmental impact on 2 indicators (16 indicators in total): The average contribution of these train journeys to the total environmental impact of the paper solution is 33.4%.
- Manufacturingof PCs has a very low contribution to the total environmental impact of the blockchain solution (average contribution of 3.2%). End-of-life of PCs has a very small contribution to the total environmental impact of the paper solution (average contribution of 1.6%)

To reduce the environmental impact of the current solution, the following actions are proposed: Use renewable electricity for use of PCs producing and maintaining the blockchain, as well as servers storing for blockchain. As with the paper solution, several parameters are imposed by regulatory constraints. Therefore, they cannot be subject of proposals for improvement at this time.

We have carried out a comparison of 10-year and 50-year blockchain storage of the innovative solution. The contribution of electricity required to run PCs and servers is very low 3.9%. With a storage period of 50 years, the average contribution of electricity consumption is only 6.7%. Thus, comparing these 2 storage durations, out of the 16 environmental indicators, the average difference is 3.4%.

Paper vs Blockchain Comparison

In this last section, we present a comparison between the current paper solution and blockchain solution by first integrating travel, then excluding travel (travel is mandatory in both technologies, and the impact of these trips being very high, they do not allow us to visualize the specificities of the 2 solutions).

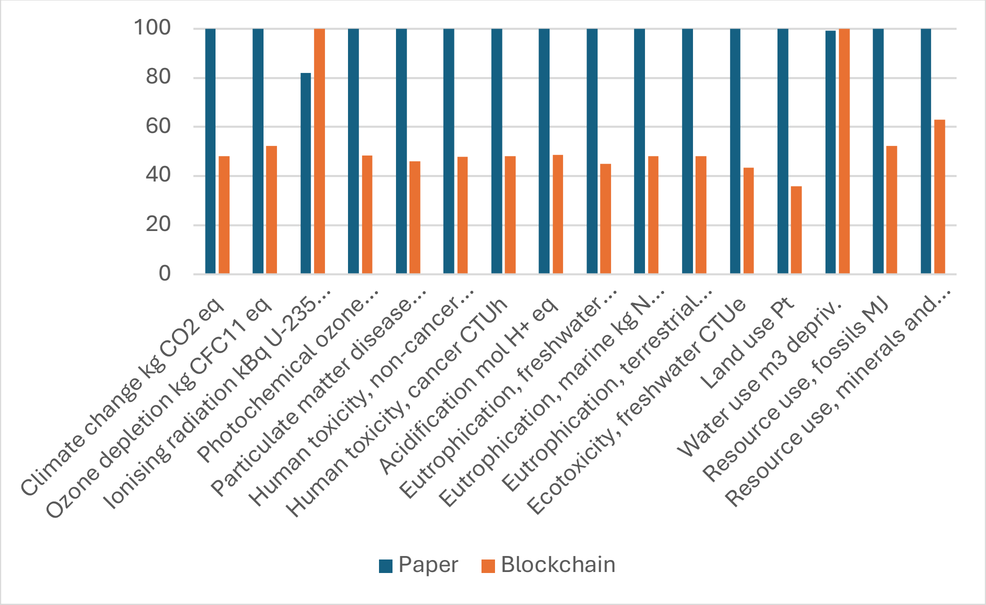


Figure 3. **Contribution comparison – paper solution/blockchain solution (with travel)**

The paper-based solution has a higher environmental impact than the blockchain solution on 14 of the 16 environmental indicators. The average gap between the 2 solutions is 44%, in favor of the blockchain solution. The blockchain solution requires 2 times less travel than the paper solution. The blockchain solution, in addition to the technological advantages it presents, also has strong environmental advantages compared to the current solution used to record and store a patient's consent during a clinical trial.

5.0 Conclusion and perspective

Blockchain offers notable advantages in the field of healthcare to manage patient consents. However, blockchain technology has its drawbacks, particularly in terms of environmental impacts. The purpose of this paper was to measure the environmental impact of a blockchain solution for patient consent in the context of clinical trials. It appears that, like any digital service, ensuring the traceability of consent have impact on the environment, beyond electricity consumption and carbon footprint; It is possible to measure its impact on natural resources, water consumption or biodiversity. In conducted study, we were able to compare the current paper-based solution for managing patient consents with a solution based on blockchain. We have shown that the paper solution has a higher impact than the blockchain solution on 14 of the 16 environmental indicators and that the average gap between the 2 solutions is 44%, in favor of the blockchain solution. In addition, the blockchain solution requires 2 times less travel than the paper solution.

Blockchain has a significant positive impact on sustainability by enhancing transparency, efficiency, and accountability in environmental initiatives. It improves supply chain traceability, allowing businesses to monitor carbon footprints and reduce emissions. By eliminating intermediaries, it fosters corruption-free climate governance, making climate financing and emissions tracking more reliable. However, blockchain also has negative environmental consequences, it faces regulatory and legal uncertainties, slowing its integration into sustainability frameworks. The quality of recorded data is another concern, as blockchain ensures immutability but cannot verify the accuracy of inputs, making reliable data verification essential. Furthermore, technological complexity and adoption barriers, particularly in developing nations, hinder large-scale implementation (Arshad et al., 2023).

The results show that blockchain technology has a significant positive impact on our study, particularly in improving the efficiency of patient consent digitalization. It facilitates the transition from paper-based to digital consent processes, enhancing traceability for all stakeholders involved. Additionally, it reduces the need for CRO travel, thereby lowering energy consumption and optimizing time and resource management. Moreover, it strengthens patient data security while improving accessibility and transparency, enabling patients to manage their own data independently and reducing reliance on intermediaries.

According to the TBS model presented by Dao et al. (2011), sustainability results from the intersection of environmental, social, and economic performance. Thus, cost reduction and improved patient satisfaction are key components of sustainability.

The findings of this study demonstrate that reducing the travel of Contract CROs offers a dual benefit: on hand, it lowers the environmental footprint, particularly in terms of energy consumption; on the other hand, it reduces costs. Furthermore, process automation and the facilitation of electronic medical record sharing among stakeholders contribute to enhanced patient satisfaction. These results confirm the environmental and economic benefits of integrating blockchain technology into the healthcare sector.

r.

However, a key challenge to its implementation lies in integrating European Union laws and regulations into the blockchain system, which may introduce technological complexities.

This study has other limitations, in particular our study is limited by virtue of our single case methodology that limits the generalizability of our results. Secondly, LCA is a very data-sensitive method and its quality.

References

AbdelSalam, F. M. (2023). *Blockchain Revolutionizing Healthcare Industry: A Systematic Review of Blockchain Technology Benefits and Threats.* Perspectives in Health Information Management, 20(3), 1b.

Anderson, C., Carvalho, A., Kaul, M., & Merhout, J. W. (2023). *Blockchain innovation for consent self-management in health information exchanges.* Decision Support Systems, 174, 114021. https://doi.org/10.1016/j.dss.2023.114021

Arshad, A., Shahzad, F., Ur Rehman, I., & Sergi, B. S. (2023). *A systematic literature review of blockchain technology and environmental sustainability: Status quo and future research.* International Review of Economics & Finance, 88, 1602-1622. https://doi.org/10.1016/j.iref.2023.07.044

Benchoufi, M., Porcher, R., & Ravaud, P. (2017). *Blockchain protocols in clinical trials: Transparency and traceability of consent.* F1000Research, 6. https://doi.org/10.12688/F1000RESEARCH.10531.5

Coiera, E., & Clarke, R. (2004). e-*Consent: The Design and Implementation of Consumer Consent Mechanisms in an Electronic Environment.* Journal of the American Medical Informatics Association, 11(2), 129-140. https://doi.org/10.1197/jamia.M1480

Dao, V., Langella, I., & Carbo, J. (2011). F*rom green to sustainability: Information Technology and an integrated sustainability framework.* The Journal of Strategic Information Systems, 20(1), 63-79. https://doi.org/10.1016/j.jsis.2011.01.002

Dinh, T. T. A., Liu, R., Zhang, M., Chen, G., Ooi, B. C., & Wang, J. (2018). *Untangling Blockchain: A Data Processing View of Blockchain Systems.* IEEE Transactions on Knowledge and Data Engineering, 30(7), 1366-1385. https://doi.org/10.1109/TKDE.2017.2781227

Elkington, J. (1994). Towards the sustainable corporation: Win-win-win business strategies for sustainable development. *California Management Review,* 90-100.

Elkington, J. (2004). *Enter the Triple Bottom Line. In The Triple Bottom Line- Does it All Add Up*? (Adrian Henriques and Julie Richardson, p. 208). https://doi.org/10.4324/9781849773348

European Commission. (2021). Commission Recommendation (EU) 2021/2279 of 15 December 2021 on the use of the Environmental Footprint methods to measure and communicate the life cycle environmental performance of products and organisations. In OJ L (Vol. 471). http://data.europa.eu/eli/reco/2021/2279/oj/eng

Fan, Z.-P., Wu, X.-Y., & Cao, B.-B. (2022). C*onsidering the traceability awareness of consumers: Should the supply chain adopt the blockchain technology?* Annals of Operations Research, 309(2), 837-860. https://doi.org/10.1007/s10479-020-03729-y

Finnveden, G., Hauschild, M. Z., Ekvall, T., Guinée, J., Heijungs, R., Hellweg, S., Koehler, A., Pennington, D., & Suh, S. (2009). *Recent developments in Life Cycle Assessment.* Journal of Environmental Management, 91(1), 1-21. https://doi.org/10.1016/j.jenvman.2009.06.018

Genestier, P., Zouarhi, S., Limeux, P., Excoffier, D., Prola, A., Sandon, S., & Temerson, J.-M. (2017). *Blockchain for Consent Management in the eHealth Environment: A Nugget for Privacy and Security Challenges*. Journal of the International Society for Telemedicine and eHealth, 5, (GKR);e24:(1-4).

Goldsmith, L., Skirton, H., & Webb, C. (2008). *Informed consent to healthcare interventions in people with learning disabilities – an integrative review.* Journal of Advanced Nursing, 64(6), 549-563. https://doi.org/10.1111/J.1365-2648.2008.04829.X

Hajian, A., Prybutok, V. R., & Chang, H.-C. (2023). A*n empirical study for blockchain-based information sharing systems in electronic health records: A mediation perspective.* Computers in Human Behavior, 138, 107471. https://doi.org/10.1016/j.chb.2022.107471

Haleem, A., Javaid, M., Singh, R. P., Suman, R., & Rab, S. (2021). *Blockchain technology applications in healthcare: An overview.* International Journal of Intelligent Networks, 2, 130-139. https://doi.org/10.1016/j.ijin.2021.09.005

ISO. (2006a). ISO 14040:2006 (en), Environmental management—Life cycle assessment—Principles and framework. https://www.iso.org/obp/ui/#iso:std:iso:14040:ed-2:v1:en

ISO. (2006b). ISO 14044:2006(en), Environmental management—Life cycle assessment—Requirements and guidelines. https://www.iso.org/obp/ui/#iso:std:iso:14044:ed-1:v1:en

Kakarlapudi, P. V., & Mahmoud, Q. H. (2021). *A Systematic Review of Blockchain for Consent Management.* Healthcare, 9(2), Article 2. https://doi.org/10.3390/healthcare 9020137

Khurshid, M., Zahid, R. M. A., & Rehman, W. ul. (2023). *Sustainable Blockchain Technologies in the Circular Economy.* In Emerging Trends in Sustainable Supply Chain Management and Green Logistics (p. 174-193). IGI Global. https://doi.org/10.4018/978-1-6684-6663-6.ch008

Kotlarsky, J., Oshri, I., & Sekulic, N. (2023). *Digital Sustainability in Information Systems Research: Conceptual Foundations and Future Directions*. Journal of the Association for Information Systems, 24(4), 936-952. https://doi.org/10.17705/1jais.00825

Madhoun, N. El, Hammi, B., Blockchain, B. H., & El Madhoun, N. (2024). Blockchain technology in the healthcare sector: Overview and security analysis.

Medaglia, R., & Damsgaard, J. (2020). *Blockchain and the United Nations Sustainable Development Goals: Towards an Agenda for IS Research.* PACIS 2020 Proceedings. https://aisel.aisnet.org/pacis2020/36

Melville, N. P. (2010). I*nformation Systems Innovation for Environmental Sustainability.* MIS Quarterly, 34(1), 1-21. https://doi.org/10.2307/20721412

Mulligan, C., Morsfield, S., & Cheikosman, E. (2024). *Blockchain for sustainability: A systematic literature review for policy impact.* Telecommunications Policy, 48(2), 102676. https://doi.org/10.1016/j.telpol.2023.102676

Petre, A., & Haï, N. (2018). *Opportunities and challenges of blockchain technology in the healthcare industry.* Medecine sciences: M/S, 34(10), 852-856. https://doi.org/10.1051/MEDSCI/2018204

Porter, M. E., & Kramer, M. R. (2007). *Strategy and society: The link between competitive advantage and corporate social responsibility.* Strategic Direction, 23(5). https://doi.org/10.1108/sd.2007.05623ead.006

Prokofieva, M., & Miah, S. J. (2019). *Blockchain in healthcare*. Australasian Journal of Information Systems, 23. https://doi.org/10.3127/ajis.v23i0.2203

Sadeghi R., J. K., Prybutok, V. R., & Sauser, B. (2022). *Theoretical and practical applications of blockchain in healthcare information management.* Information & Management, 59(6), 103649. https://doi.org/10.1016/j.im.2022.103649

Seidel, S., Chandra Kruse, L., Watson, R. T., Albizri, A., Boudreau, M.-C. (Maric), Butler, T., Chandra Kruse, L., University of Liechtenstein, Guzman, I., Karsten, H., Lee, H., Melville, N., & Watts, S. (2017). *The Sustainability Imperative in Information Systems Research.* Communications of the Association for Information Systems, 40, 40-52. https://doi.org/10.17705/1CAIS.04003

Seife, C. (2015). R*esearch Misconduct Identified by the US Food and Drug Administration: Out of Sight, Out of Mind, Out of the Peer-Reviewed Literature.* JAMA Internal Medicine, 175(4), 567-577. https://doi.org/10.1001/JAMAINTERNMED.2014.7774

Truby, J. (2018). Decarbonizing Bitcoin: *Law and policy choices for reducing the energy consumption of Blockchain technologies and digital currencies.* Energy

Research & Social Science, 44, 399-410. https://doi.org/10.1016/J.ERSS.2018.06.009

Watson, R. T., Boudreau, M.-C., & Chen, A. J. (2010). *Information Systems and Environmentally Sustainable Development: Energy Informatics and New Directions for the IS Community.* MIS Quarterly, 34(1), 23-38. https://doi.org/10.2307/20721413

Wolf, M.-A., Chomkhamsri, K., Brandao, M., Pant, R., Ardente, F., Pennington, D., Manfredi, S., De, C. C., & Goralczyk, M. (2011, janvier 18). International Reference Life Cycle Data System (ILCD) Handbook—General guide for Life Cycle Assessment—Detailed guidance. JRC Publications Repository. https://doi.org/10.2788/38479

Zhang, A., Zhong, R. Y., Farooque, M., Kang, K., & Venkatesh, V. G. (2020). *Blockchain-based life cycle assessment: An implementation framework and system architecture.* Resources, Conservation and Recycling, 152, 104512. https://doi.org/10.1016/J.RESCONREC.2019.104512

Automating Business Process to Enhance Organisational Efficiency and Productivity: Using Academic Intervention at Salford Business School as a Case Study

Yun Chen, Kate Han, Charlotte Seager
Salford Business School, The University of Salford

Abstract

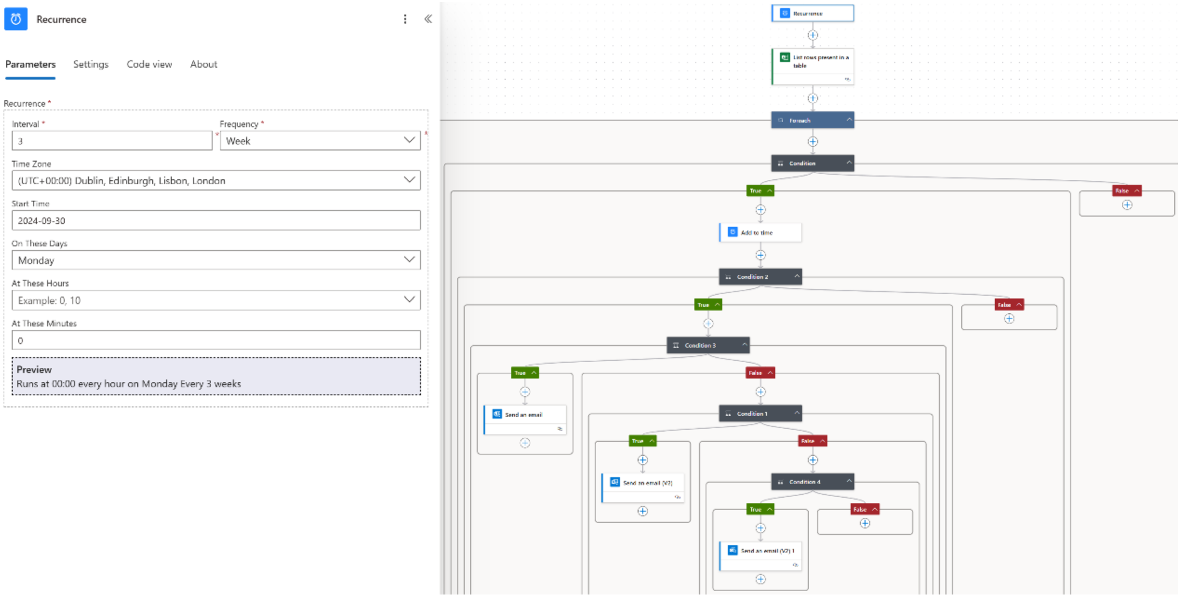
Organisations, particularly large ones with multiple stakeholders involved in their business processes, frequently face challenges related to process efficiency and productivity, leading to wasted resources, time, and effort, ultimately reducing overall output and performance. These issues can stem from poor workflow design or a lack of automation. Higher Education Institutions (HEIs) are no exception. For example, in academic interventions, current process models often fail to account for the complexity of the organisational context. At Salford Business School, Academic Personal Tutors (APTs) play a key role in academic intervention to support student success, by monitoring student activity and intervening when necessary. However, manually consolidating data from various platforms is time-consuming and limits the team's ability to focus on personalised student support. This paper presents an ongoing project aimed at optimising processes by utilising AI-driven data analysis, automation, and the Microsoft ecosystem for process modelling and data integration. In this developing paper, we will discuss the preliminary work we have completed on the project and outline the plans for its future development.

Keywords: Process Automation, Organisational Efficiency, Productivity, Academic Intervention, Digital Workplace

1.0 Introduction

Many organisations, especially the large ones involving multiple stakeholders, struggle with business processes that are not automated or optimised, leading to inefficiencies, delays, and resource strain (Brás and Moro, 2023; Dumas, 2018). HEIs are no exception, facing similar challenges in managing complex processes that involve academia, professional services staff, students, and external partners, such as student admissions, curriculum management, research funding applications, (Postgraudate Research) PGR review processes, and partnership agreements with industry or other institutions etc. These processes often require coordination across multiple departments, making it difficult to track progress, avoid delays, and ensure timely completion without automated systems in place. To address this, some HEIs have experimented with using the Microsoft 365 ecosystem to streamline the process (Goedde, 2024). At Salford Business School

(SBS), we have implemented automated reminders for admissions tutors using Microsoft Power Automate. These reminders are triggered by the current status of an application and predefined deadlines, helping to reduce application backlogs. The system is integrated with Microsoft SharePoint and Outlook to streamline the process efficiently (Figure 1). This demonstrates that even minor automation within a standard licensed system (i.e. Microsoft 365) can result in substantial enhancements in managing complex, multi-stakeholder processes in HEIs.



Recurrence

Parameters | Settings | Code view | About

Recurrence *

Interval *: 3

Frequency *: Week

Time Zone: (UTC+00:00) Dublin, Edinburgh, Lisbon, London

Start Time: 2024-09-30

On These Days: Monday

At These Hours: Example: 0, 10

At These Minutes: 0

Preview
Runs at 00:00 every hour on Monday Every 3 weeks

Figure 1. PGR Automative Reminder

To further investigate the potential of automation in enhancing business process efficiency and decision-making for organisational productivity, we are collaborating with the APT team to develop a more effective academic intervention process, which will be discussed in this paper. The project is based on ongoing action research at SBS, aimed at improving student success through data-driven strategies (Chen & Han, 2024; Chen et al., 2024). The preliminary research has uncovered key challenges faced by the SBS APT team, as identified through interviews. Addressing these challenges through successful implementation could lead to scaling the system across other schools and processes within the University, or even adapting it for use in other HEIs. Our vision is to create an optimised data-driven approach that enhances the ability of the School APT team to support student success, serving as the showcase for improving efficiency and productivity in organisations. By automating data collection and key aspects of the process, we aim to reduce the manual workload on APTs, enabling them to focus more on strategic, personalised, and timely interventions. Currently, the APT team spends 600 hours per year on manual data administration, equivalent to the time they spent delivering personalised workshops. This project will deliver a system that automates these administrative tasks and seamlessly integrates them into the APT process, providing a monitoring mechanism based on their Key Trackable Priorities (KTP). This will improve intervention accuracy and timeliness while optimising efficiency and productivity across stakeholders.

As this is a research project in progress, the remainder of this paper will outline the project brief, detail the work completed thus far, and present our future plans.

2.0 Project Discussion

2.1 Project Aim and Objectives

The project aims to automate the APT interventions with students through process modelling, data integration, and automation, leveraging Microsoft ecosystems to enhance process efficiency and employee productivity.

To achieve the aim, research objectives include:

1. To collaborate with the APT team to map the current student intervention process, analyse existing data sources, and identify challenges and opportunities for automation.
2. To develop and implement machine learning models, particularly ensemble learning, to collect data from various systems (e.g., Blackboard, QlikView) and integrate the data into KTP tracking system.
3. To optimise the process by minimising the time and complexity involved in managing and interpreting student data.
4. To refine the process through prototype testing and roll out the system to the APT team, providing them with real-time insights and alerts to enable timely and effective interventions.

2.2 Business Requirements

To understand the challenges that need to be addressed within the APT team for process optimisation, an initial meeting was held with the APT lead in October 2024. The initial challenges identified within the APT team centre on three key areas: exportation and consolidation of key data from across multiple sources; time spent on individual communications with students across all UG programmes, but also with those students identified as a priority target group for support; and finally, manual tracking of ongoing data to identify triggers for academic intervention and prevent resource waste.

At the start of the academic term, for the APT team to effectively organise the students into priority groups for support, they are required to draw data from multiple sources. The sources are not all organised in the same way and collation of the data is therefore a manual exercise. The APT's are required to export from student attendance and engagement platforms, student admissions records, progression and attrition trackers, as well as submission and resubmission data from previous academic years. This data then needs to be manually collated and organised to determine those priority groups which may need additional academic support. The collated data is then used as a starting point, from which to map the on-going support needs and academic interventions which may be organised and tracked throughout the year. The manual extraction and collation of data presents challenges around human error, resource intensity and time.

Once the priority groups have been established, the APT team must then send comms, according to the position in the academic year, to all UG students and all priority group students to book their academic support meeting. If a student does not attend their booked meeting, there is an additional responsibility on the APT to chase up the rationale for the non-attendance and reschedule the support. Further, in the lead up to assessment submission, mandated submission support comms are sent in the week leading up to the submission deadline to all UG students. These individual targeted comms to different priority groups, and at different points of the year to all UG students, is administratively time consuming. Further, in cases where the APT team is required to chase a student who has not attended their academic support meeting, the resource of personal support is wasted two-fold.

Throughout the academic year, the APT team is required to track assessment component submissions, so that targeted interventions can be executed in a timely manner. At the point where a student has not submitted by the deadline, the APT team must reach out to support late submission within a seven-day period. This requires the manual extraction of the submission data which is to then be mapped against the student data for the APT to reach out. The nature of this manual tracking presents a challenge around organising interventions in time to be effective: by the time the APT team has extracted and mapped the submission data, then reached out to the individual students who did not submit on time, and organised the intervention meeting they are likely to be a few days into the late submission window already, hindering the effectiveness of the intervention. The time-consuming manual efforts of the APT team are depicted in Figure 2 below.

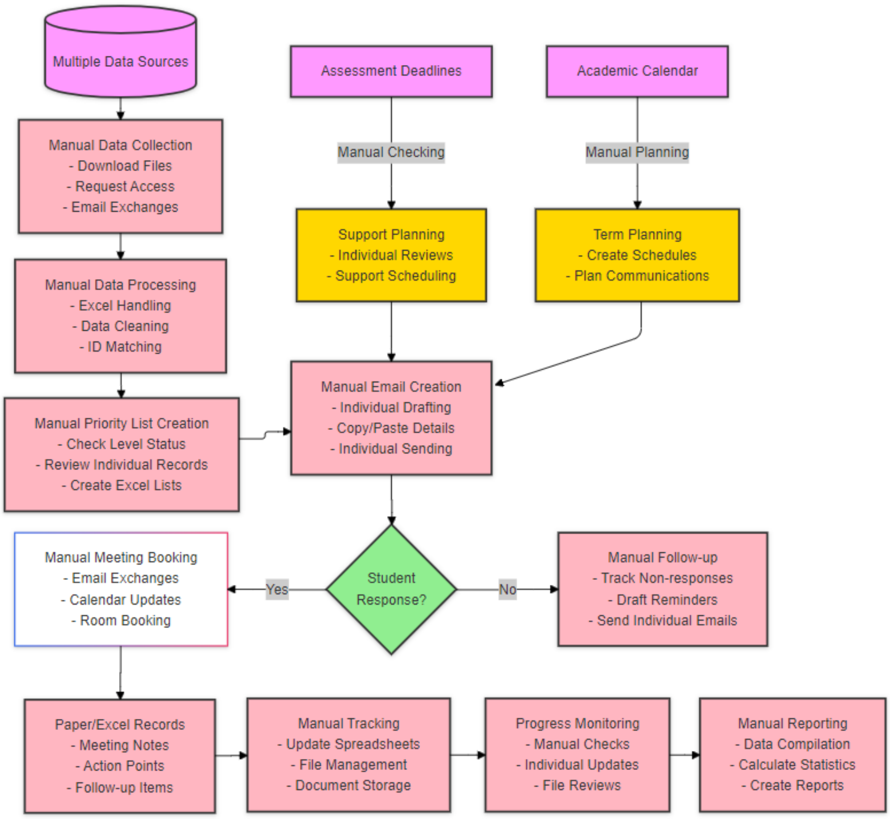


Figure 2. APT Manual Workflow

To address these challenges, we are aiming to automate much of the initial data exportation and consolidation required to ascertain the priority groups. The project also aims to automate, where possible, the comms pieces which are sent routinely to the students as well as the administration which follows the non-attendance at a support meeting. Finally, automating the continual drawdown and mapping of submission data and the subsequently triggered intervention comms, seeks to overcome the challenge of wasted resource within the APT team.

2.3 Process Mapping

To gain a clearer understanding of the system, process mapping is being conducted to assess the current processes within the APT team. As shown in Figure 3 below, the APT team currently relies heavily on manual efforts for data collection, integration, and tracking of support tasks.

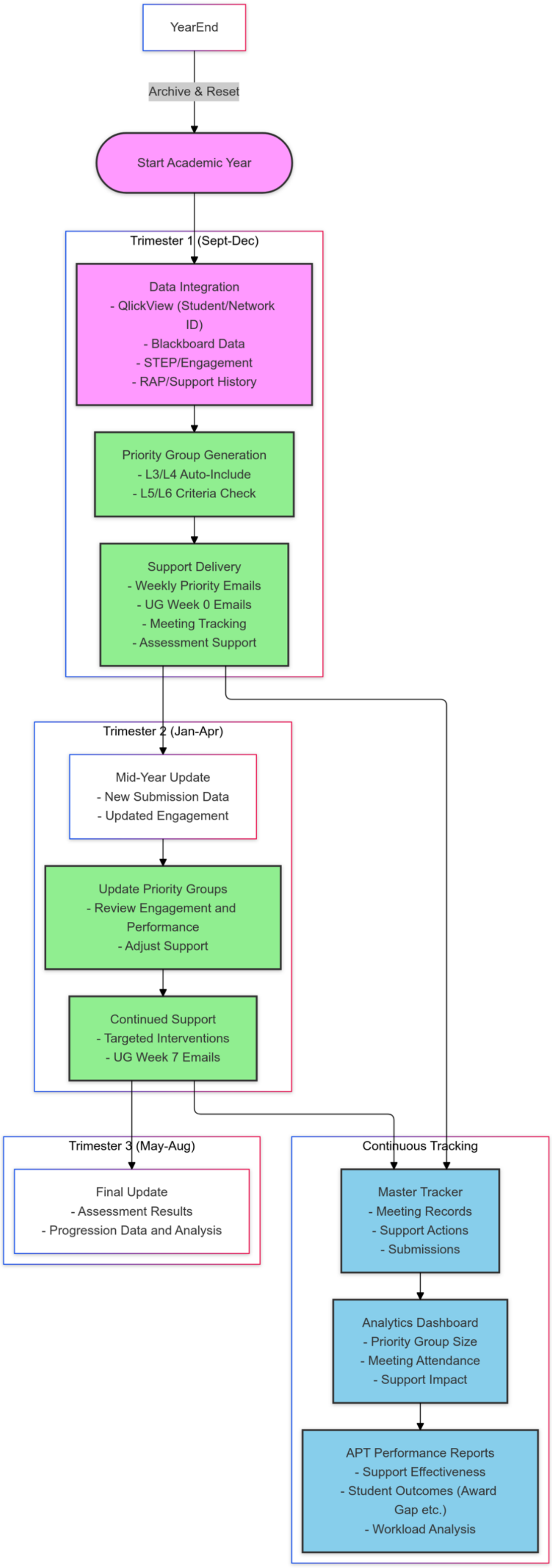

YearEnd
Archive & Reset
Start Academic Year
Trimester 1 (Sept-Dec)
Data Integration
- QlickView (Student/Network ID)
- Blackboard Data
- STEP/Engagement
- RAP/Support History
Priority Group Generation
- L3/L4 Auto-Include
- L5/L6 Criteria Check
Support Delivery
- Weekly Priority Emails
- UG Week 0 Emails
- Meeting Tracking
- Assessment Support
Trimester 2 (Jan-Apr)
Mid-Year Update
- New Submission Data
- Updated Engagement
Update Priority Groups
- Review Engagement and Performance
- Adjust Support
Continued Support
- Targeted Interventions
- UG Week 7 Emails
Trimester 3 (May-Aug)
Final Update
- Assessment Results
- Progression Data and Analysis
Continuous Tracking
Master Tracker
- Meeting Records
- Support Actions
- Submissions
Analytics Dashboard
- Priority Group Size
- Meeting Attendance
- Support Impact
APT Performance Reports
- Support Effectiveness
- Student Outcomes (Award Gap etc.)
- Workload Analysis

Figure 3. Business Process Mapping: Current ATP Workflow

To address the challenges of manual workloads, the proposed automation includes the following components:

- Data clean, integration and entrance automation for *Priority Groups Generation,* as indicated in Figure 4 below.
- APT support processes automation for *APT Tracker*, *Student Performance Analysis* and *Report Generation,* as indicated in Figure 5 below.

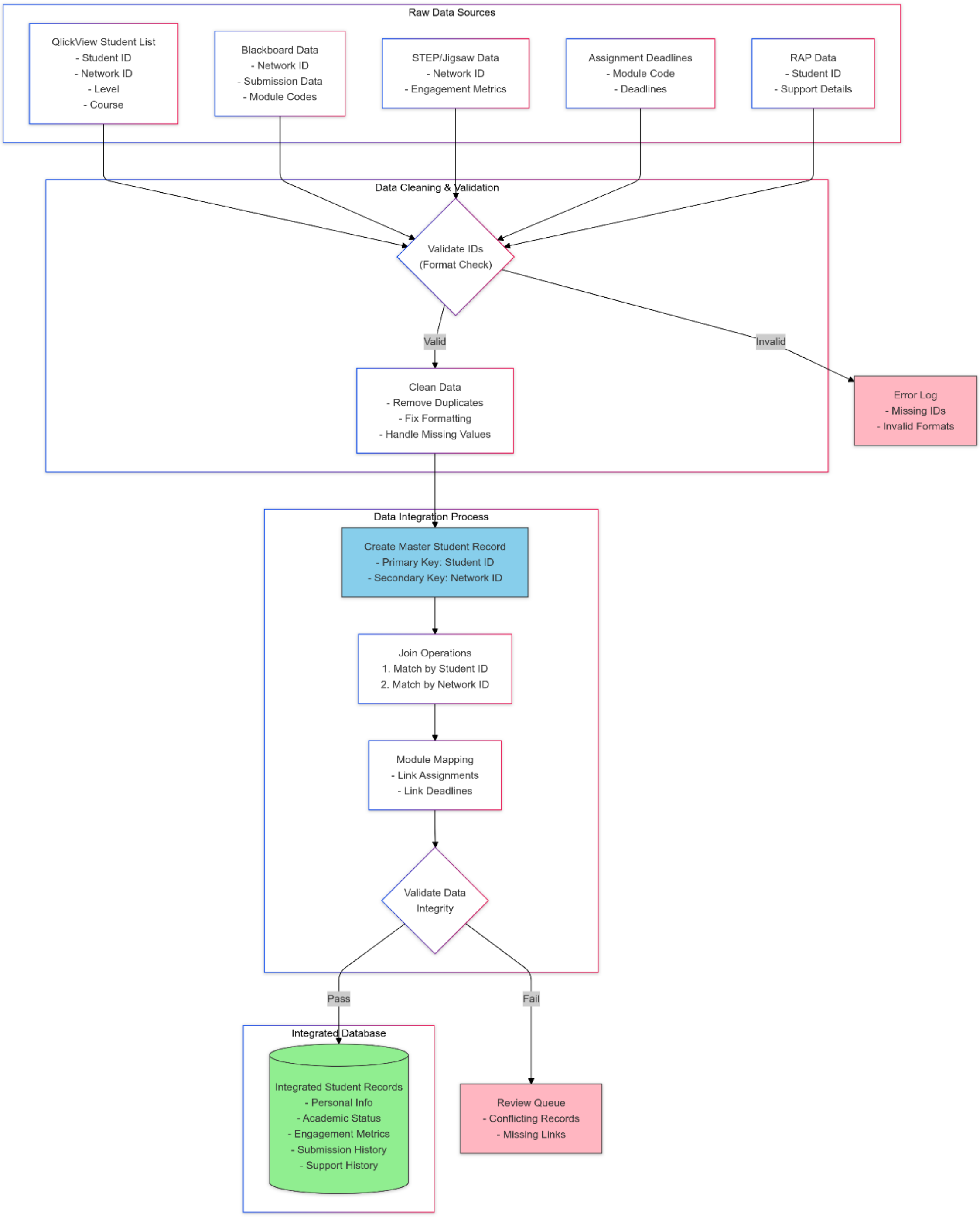


Figure 4. Data Clean, Integration and Entrance Automation

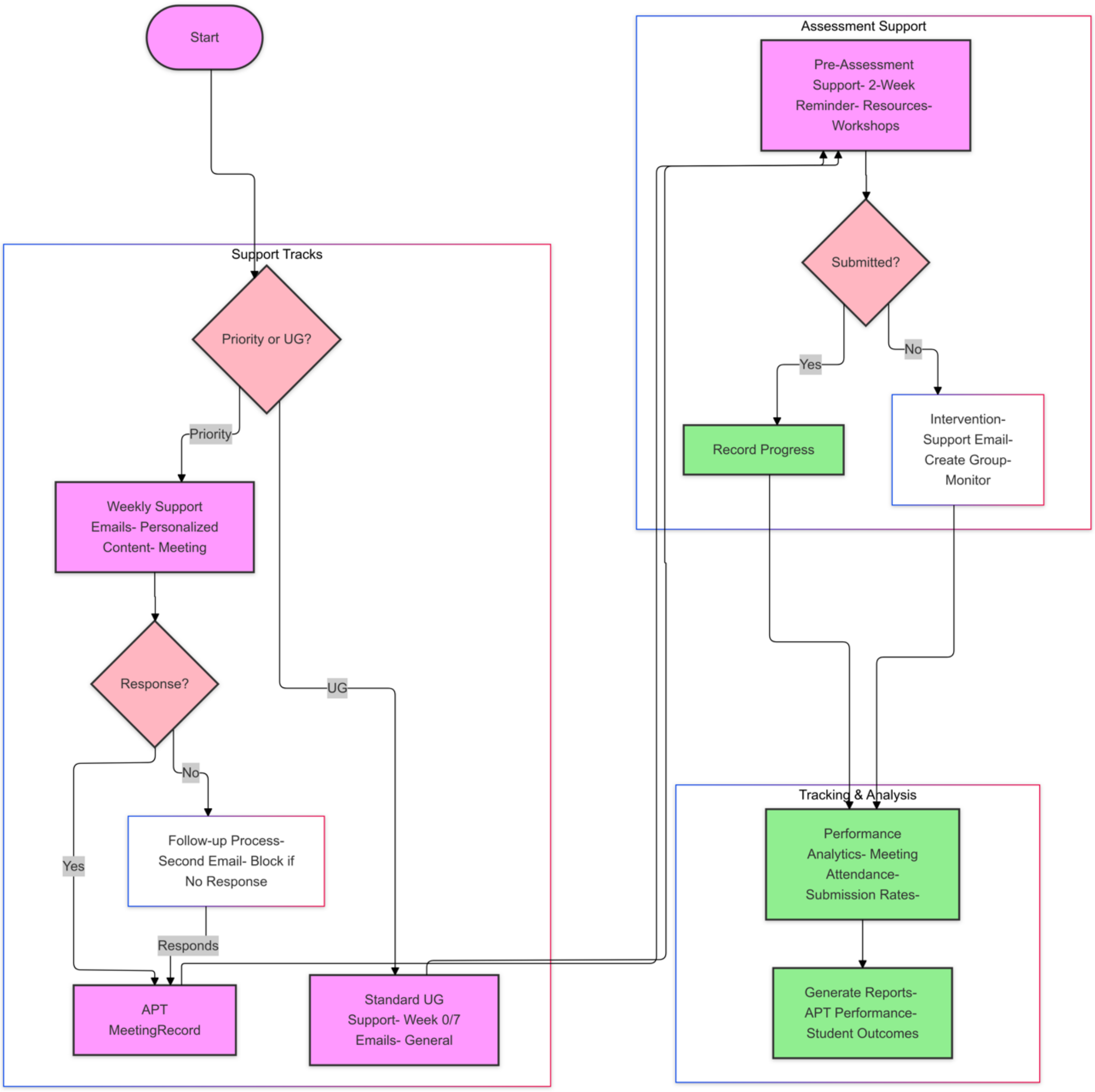

Start
Support Tracks
Priority or UG?
Priority
Weekly Support Emails- Personalized Content- Meeting
Response?
No
Follow-up Process- Second Email- Block if No Response
Yes
Responds
APT MeetingRecord
UG
Standard UG Support- Week 0/7 Emails- General
Assessment Support
Pre-Assessment Support- 2-Week Reminder- Resources- Workshops
Submitted?
Yes
No
Record Progress
Intervention- Support Email- Create Group- Monitor
Tracking & Analysis
Performance Analytics- Meeting Attendance- Submission Rates-
Generate Reports- APT Performance- Student Outcomes

Figure 5. APT Support Automation

2.4 Future Plan

We will adopt the Agile methodology for system development, which emphasises flexibility, collaboration, and stakeholder feedback throughout the development process. Agile methodologies are particularly well-suited for projects requiring iterative improvements and rapid responses to changing requirements, making them ideal for our system development efforts (Ardo and Gaber; 2022). According to Nicolaas (2018), agile allows teams to break down projects into smaller, manageable units, enabling continuous improvement and adaptive planning, which is critical in our context. The methodology will not only streamline the development process but also ensure that the end product is responsive to the needs of all stakeholders involved, fostering a more efficient academic environment.

During the next step, we will focus on prototype development. Our activities will include designing and developing a comprehensive data pipeline to consolidate data from Blackboard and QlikView for automation, alongside creating a draft prototype that incorporates these automation features. The deliverable for this phase will be a system prototype built on a predictive model, Microsoft Power Automate, and integrated datasets, ensuring that the system is aligned with user needs and organisational goals.

After the initial development, we will conduct pilot testing of the prototype with a selected group of APTs. This phase will involve gathering feedback from both the APT team and the broader UK academic community to identify necessary adjustments. Deliverables will include a focus group session for prototype testing, facilitating direct user engagement and insights into the system's functionality and usability.

Following this, we will refine the prototype based on the feedback obtained during the pilot phase, leading to an application prototype that effectively reflects these refinements. This iterative feedback loop is a cornerstone of the Agile methodology, as it fosters a culture of continuous improvement. Finally, we will compile a comprehensive final report documenting outcomes, lessons learned, and recommendations for future work. This report will also include training tutorials developed in collaboration with the APT team, designed to integrate automation and the Microsoft ecosystem into organizational practices. By doing so, we aim to enhance productivity and process efficiency, using the APT case study as a reference point.

3.0 Conclusion

This paper presented a research project currently underway, initiated in October 2024, aimed at enhancing organisational efficiency and productivity by using academic intervention as a pilot case study. We believe that this initiative has the potential to significantly improve student retention, satisfaction, and academic performance in the long term. By aligning with the University's strategic objectives, it is poised to make a meaningful impact not only on the institution but also on the broader landscape of HEIs.

Furthermore, we believe that the success of this project can open doors to additional external funding opportunities with both industrial partners and researchers. This includes business-led Innovate UK funding and potential Knowledge Transfer Partnerships (KTPs) aimed at enhancing productivity and efficiency in organisations with multiple stakeholders and system integration, such as AI solutions intended to improve productivity in key sectors. Additionally, research funding could be sought to explore advanced predictive models and expand data-driven frameworks for student support.

References

Ardo, A., Bass, J., & Gaber, T. (2022). *Towards secure agile software development process: a practice-based model*. Institute of Electrical and Electronics Engineers.

Brás, J., Pereira, R., & Moro, S. (2023). Intelligent Process Automation and Business Continuity: Areas for Future Research. *Information (Basel)*, *14*(2), 122-. Available at: https://doi.org/10.3390/info14020122

Chen, Y. and Han, K. (2024). Driving Student Success through a Data-Driven Approach in Higher Education, *In: the Proceeding of UKAIS*, Kent, UK, 25th -26th April 2024.

Chen, Y., Han, K. and Doolan, M.A. (2024). Generative AI assisted Life-long Learning in Higher Education: A Case Study of Coding Learning for Business Students at Salford Business School. *Book Chapter of: IFNTF Symposathon.*

Dumas, M. (2018). *Fundamentals of Business Process Management* (2nd ed. 2018.). Springer Nature. Available at: https://doi.org/10.1007/978-3-662-56509-4

Goedde, K. (2024). *Introducing the Microsoft Ecosystem for the Modern Worker.* Available at: Introducing the Microsoft Ecosystem For the Modern Worker | Velosio

Innovate UK (2024). Available at: Innovate UK – UKRI

Nicolaas, D. (2018). *Scrum for teams: a guide by practical example* (First edition.). Business Expert Press.

Does digitalization ameliorate government tax administrations?

Evidence from corporate tax avoidance

Guanming He

Durham Business School, Durham University

Durham, the United Kingdom; DH1 3LB

Email: guanming.he@durham.ac.uk

Zhichao Li
University of Exeter Business School,
Rennes Drive, Exeter, the United Kingdom; EX4 4PU
Email: z.li10@exeter.ac.uk

Dongxiao Shen

Durham Business School, Durham University

Durham, the United Kingdom; DH1 3LB

Email: dongxiao.shen@durham.ac.uk

*We appreciate the helpful comments and suggestions from Jonathan Batten, Craig Depken and conference/workshop participants at the 5th annual conference of the association of British Chinese Professors (ABCP) (which confers an outstanding early career research award on this paper), 3rd Marketing, Innovation and Digital Transformation Global Conference, 62nd World Continuous Audit & Reporting Symposium WCARS and 2nd British Accounting & Finance Association (BAFA) Accounting Information Systems & Emerging Technologies (AISET) Special Interest Group conference, the 2nd annual International Conference on Empirical Economics (ICEE) at PSU-Altoona, 2nd Ethical Finance, Accountability and Governance Symposium for sustainability, 15th Japanese Accounting Review conference, and 2024 International Doctoral Consortium at the Durham University Business School.

Does digitalization ameliorate government tax administration? Evidence from corporate tax avoidance

Abstract:

Improving tax administration is particularly important for developing countries, where corporate tax misconduct is a severer issue. We focus on the Golden Tax Project IV (GTP IV) in China, which emphasizes the adoption of advanced digital technologies to tax administration, and investigate its impact on corporate tax avoidance. We find that GTP IV curbs corporate tax avoidance, with a more pronounced effect on firms that feature a relatively higher propensity to avoid taxes (e.g., firms with higher risk and greater financial constraint). Our study highlights the effectiveness of digitalized tax administration in fostering a more equitable and efficient tax system.

Keywords: digitalization; tax avoidance

JEL classification codes: H26; M00

1. Introduction

Digital technologies are increasingly adopted in a wide range of areas. In recent years, a number of governments have also started utilizing various advanced digital technologies to streamline and enhance their administrative functions (Mergel *et al*., 2019). This digitalization is particularly important in the tax administration of developing countries, where corporate tax avoidance is prevalent (e.g., Chen *et al.*, 2021; Na and Yan, 2022; Tang *et al.*, 2017). Firms in general have incentives to avoid taxes for more internal funds and higher profits (Desai and Dharmapala, 2006; Edwards *et al*., 2016). Such tax avoidance, however, threatens the fairness of tax burden distribution and significantly undermines the government's fiscal capacity to provide public services (Slemrod and Yitzhaki, 2002). The application of advanced digital technologies offers a potential solution to curb corporate tax avoidance via improving tax administration.

In this study, we examine the effectiveness of such digitalization through the lens of a significant initiative in China — the Golden Tax Project IV (known as GTP IV).[1] Launched in December 2021 across three pilot provinces in China (see Appendix 1), GTP IV builds on the unified online tax information platform established by GTP III but introduces more advanced technologies. In specific, GTP III aimed at improving tax compliance through better invoice matching and verification. It established an information system, which covers all types of taxes, to unify and integrate the national and local taxation data. It is conducive to curbing corporate illegal practices such as fake invoices, as it ensures taxes paid by firms are accurately substantiated by actual invoices documented within the system.[2] Therefore, Phase III focused

[1] The Golden Tax Project commenced in 1994 and has gone through three distinct phases. Phase I aimed to verify the authenticity of value-added tax (VAT) invoices. Phase II expanded the focus to the management of issuance, declaration, and verification of VAT invoices. Phase III employed advanced information technologies to unify a nationwide online tax information system and emphasized tax management through electronic invoicing.

[2] Managers could issue fake invoices to claim tax-deductible expenses that were never incurred. By limiting and eliminating fake invoices, GTP III could mitigate the overstatement of tax-deductible expenses by firms.

primarily on enhancing the capabilities of tax administration at the collection level and may lack effectiveness in identifying and restraining corporate tax avoidance that can be realized via real activities manipulation, among others.

GTP IV utilizes more sophisticated digital technologies that go beyond the basic digitalization laid by the previous phase. It employs big data analytics, artificial intelligence (AI), cloud computing and blockchain not just for enhancing data collection but also for enabling real-time analysis and cross-departmental information sharing. On one hand, GTP IV applies big data analytics, AI and cloud computing to achieve systematical, automatic analysis of all tax-related information and detect potential corporate non-compliance risks. For instance, GTP IV introduces a "smart" tax system powered by the foregoing techniques to scrutinize industry-specific benchmarks, which are closely linked to a firm's reported financial numbers. Any significant deviation from the established norms could trigger an automatic alert within the new system, allowing the tax authorities to detect corporate tax avoidance more effectively than ever before. As such, we expect that tax authorities could improve their ability in tax-risk management by effectively reducing the costs and inefficiency of monitoring corporate tax misconduct. On the other hand, GTP IV also uses blockchain to promptly share tax-related information across different government departments and institutions, enhancing the traceability of data flows and supporting the real-time verification of business transactions. This not only reduces information asymmetry, but also improves the overall efficiency of tax monitoring by providing a reliable and unalterable audit trail. Consequently, corporate tax avoidance would be restrained (Atwood *et al.*, 2012; Kerr, 2019).

However, contrasting research counters the aforementioned view by showing that the enhanced information environment due to digitalization might provide firms with additional opportunities to minimize their tax liabilities (e.g., Zhou *et al.*, 2022). This finding casts doubt on the effectiveness of digitalized tax administration. From another perspective, the deployment of advanced digital technologies is not without risks and challenges. Issues such as technological obsolescence, privacy concerns and cybersecurity risks make it difficult to realize effective digital monitoring of potential tax misconduct (Anthopoulos *et al*., 2016; Gupta *et al*., 2024). Furthermore, the considerable time, substantial financial investments, learning curves and high uncertainties in applying digital technologies (Luo, 2022) may also prevent the digitalized tax administration from meeting its intended objectives. Therefore, it is unclear whether the digitalized tax administration is effective in curbing corporate tax avoidance.

To address this open question, we empirically analyze whether and how these innovative technologies as applied in GTP IV impact corporate tax avoidance. Although previous studies have documented the positive economic consequences of earlier phases of GTP to firms (e.g., Li *et al.*, 2020; He and Yi, 2023), our study is the first to explore how the integration of more advanced techniques as in GTP IV, which are complex and widely used in all taxation processes, influences corporate tax conduct. Our difference-in-differences regression analysis reveals that firms located in GTP IV pilot areas exhibit less tax avoidance than those in areas where GTP IV has not been implemented, suggesting that the digitalization being applied in tax administration helps deter corporate tax avoidance. Our moderation analysis further reveals that such a deterrence effect is more prominent in firms that are prone to avoid taxes, such as high-risk firms, small firms, financially constrained firms and firms with high financial opacity. In all, our study contributes to the discourse on tax enforcement and corporate tax avoidance, adding to the literature on the benefits of sophisticated infrastructure in this regard (e.g., D'Avino, 2023; Svetlozarova Nikolova, 2023; Zídková *et al*., 2024). Particularly, we shed light on how governmental digitalization, as a new form of advanced infrastructure, enhances tax enforcement practices. Our results hold implications for tax authorities worldwide in their pursuit of digitalization.

2. Data

Our initial sample covers Chinese firms listed on the Shanghai or Shenzhen Stock Exchange. To ensure a clear comparative analysis subject little to the time-series confounding effect of COVID-19, we focus on the years 2021 (the pre-GTP IV sample period) and 2022 (the post-GTP IV sample period). We do not include the year 2020 due to the significant influence of the pandemic on corporate behaviors in China, nor the year 2023 when the impact of the pandemic diminished substantially as a consequence of the lifted quarantine requirements by the Chinese government. We obtain data from the Chinese Stock Market and Accounting Research (CSMAR) database, and construct the variables as defined in Appendix 2; they are all winsorized at 1% and 99% for the multivariate test. The sample selection procedure is described in Table 1.

3. Research design and results

3.1. Baseline regression analysis

We use the following difference-in-differences (DID) regression model to test the effect of GTP IV on corporate tax avoidance:

$$DD\_BTD_{i,t} = \alpha_0 + \alpha_1 Treat_i + \alpha_2 Treat_i \times Post_t + Controls + \varepsilon \qquad (1)$$

The dependent variable, $DD\_BTD_{i,t}$, is the residual book-tax difference estimated from the firm-fixed-effects regression model developed by Desai and Dharmapala (2006). A higher value of DD_BTD represents a higher level of tax avoidance by the firm. The treatment indicator, $Treat_i$, equals 1 for the treatment firms and 0 for the control firms. The treatment firms are those located in the three pilot provincial regions where GTP IV was implemented during the sample period. The control firms are those located in other provinces, and exclude those located in Sichuan (Xiamen), where GTP IV was implemented in October (November) 2022, for a precise comparative analysis. The time indicator, $Post_t$, equals 1 (0) for the post-GTP IV (pre-

GTP IV) period. The coefficient of the interaction term, $Treat_i \times Post_t$, captures the change in corporate tax avoidance by the treatment firms between the pre- and post-GTP IV period, relative to that by the control firms. We include a battery of control variables alongside year dummies, industry dummies and region dummies in the regression. $Post_t$ is not included because it is multicollinear with year dummies. The inclusion of region dummies is aimed at controlling for potential hidden variations of treatment that are associated with plausibly differential tax enforcements by local governments across regions.

To reduce sample selection bias associated with the plausible non-random selection of pilot regions by the government, we use the propensity-score matching (PSM) approach to match each treatment firm, without replacement, with a control firm, of which the propensity score is closest to that of the treatment firm. The propensity scores are estimated using a logit regression, in which the binary variable (*Treat*) is regressed on a set of covariates as to the firm's fundamental characteristics, including firm size (*SIZE*), return on assets (*ROA*), book-to-market ratio (*BM*), financial leverage (*LEV*), firm risk (*Std_return*) and board independence (*Indp*).

Panel A of Table 2 shows the results for the univariate test of covariate balance in the post-matched sample. All the mean differences in covariates between the treatment and control groups are not statistically significant, while the standardized bias is less than 5% for all covariates. We also run the logit regression based on the post-matched samples to check the covariate balance. As shown in Panel B of Table 2, the coefficients for all covariates are not statistically significant, suggesting that our post-matched sample achieves a covariate balance. To further check the effectiveness of our matching, we conduct a test of common support in PSM. The results are presented in Figure 1. As seen in Figure 1-a, there is a notable difference in propensity scores between the treatment group and control group before the matching. In contrast, Figure 1-b reveals that the distribution trends of the treatment group and the control group become overlapping post matching. These results suggest that our sample matching effectively simulates the condition of randomization of observations between the treatment group and control group.

We use the post-matched sample for the difference-in-differences (DID) regression analysis. Panel A of Table 3 reports the descriptive statistics of regressors. Their correlation coefficients are presented in Panel B. Table 4 reports our DID regression results. The coefficient of the interaction term is negative (-0.009) and statistically significant at the 1% level (*t*-stat.=2.61). A one-standard-deviation increase in *Treat*Post* leads to a decrease of *DD_BTD* by 0.004, which accounts for 22.66% of the mean of *DD_BTD* and is thus economically significant. These results suggest that digitalization in tax administration, as applied in GTP IV, curbs corporate tax avoidance. Noticeably, the coefficient for *Treat* is not statistically significant, denoting that there is no difference in our outcome variable, *DD_BTD*, between the treatment and control firms before the enforcement of GTP IV. As such, the nonsignificant coefficient of *Treat* not only lends support to the parallel trend assumption for our DID regression estimation, but also refutes a plausible alternative explanation for our DID results – that in anticipation of implementation of GTP IV, the treatment firms avoid taxes more aggressively in advance of the regulatory event, relative to the control firms.

3.2. Placebo test

To lend further credence to the treatment effect of digitalized tax administration on corporate tax avoidance, we perform a placebo test. In specific, we create a pseudo-treatment sample by randomly selecting a number of firms, equivalent to the number of treatment firms, from the control group. Utilizing the same propensity-score-matching approach as did previously, we match the pseudo-treatment sample with the control sample, and then use the matched sample to run the DID regression model (1). We repeat this process 1,000 times. Figure 2 shows that the coefficients for the placebo DID estimator follow a normal distribution and are concentrated around zero. Most of the coefficients have p-values greater than 0.05 and are located to the right of the vertical dash line which indicates the magnitude of our baseline DID coefficient (i.e., -0.009). As such, the effect of digitalized tax administration on corporate tax avoidance is nullified after the randomization and placebo processes. These placebo results lend support to the stable unit treatment value assumption underlying our DID regression estimation, and in reverse, substantiate the treatment effect implied by our baseline DID results.

3.3. Moderation analysis

We further test whether our baseline regression results vary by financial constraint, state ownership, business risk, financial opacity and firm size. To this end, we divide our post-matched sample into two subsamples based on the binary variable for state ownership (*SOE*) and the medians of continuous variables for financial constraint (*SA_index*), financial opacity (*Opacity*), business risk (*Std_return*) and firm size (*SIZE*), respectively. Then we run the baseline DID regression for each set of subsamples. Table 5 reports the regression results. Digitalization in tax administration restrains corporate tax avoidance to a larger extent for non-state-owned firms, smaller firms, and firms with tighter financial constraints, higher business risk or greater financial opacity. These firms are featured by a higher tendency to avoid income taxes (e.g., Bradshaw *et al.*, 2019; Edwards *et al.*, 2016; Kerr, 2019). Therefore, the results of the moderation analysis further underscore the effectiveness of digitalized tax administration in constraining corporate tax avoidance.

4. Conclusion

We provide robust evidence to suggest that digitalization enhances the effectiveness of tax administration in curbing corporate tax avoidance. Our findings encourage tax authorities to embrace advanced digital technologies properly into tax systems, as employing digital tools facilitates precise detection of, and timely interventions to, corporate tax avoidance. Besides, our research highlights the need to tailor digitalized tax systems to firms that are prone to engage in tax avoidance. For instance, non-state-owned firms, smaller firms and those with greater financial opacity or constraints tend to avoid taxes more aggressively. By customizing digital interventions to target these firms with a high risk of non-tax compliance, tax authorities could reinforce the impact of their regulatory efforts. Lastly, digitalization can also help improve international collaboration on tax issues by simplifying the process of sharing information between countries. This improvement is particularly effective in tackling tax avoidance by multinational companies which often minimize their tax obligations through strategies such as profit shifting or exploiting differences in national tax rules.

References

Anthopoulos, L., Reddick, C. G., Giannakidou, I. and Mavridis, N. (2016). "Why e-government projects fail? An analysis of the Healthcare. gov website", *Government information quarterly*, Vol. 33 NO. 1, pp. 161–173. https://doi.org/10.1016/j.giq.2015.07.003.

Atwood, T. J., Drake, M. S., Myers, J. N. and Myers, L. A. (2012), "Home Country Tax System Characteristics and Corporate Tax Avoidance: International Evidence", *The Accounting Review*, Vol. 87 No.6, pp. 1831–1860. https://doi.org/10.2308/accr-50222.

Bradshaw, M., Liao, G. and Ma, M. (Shuai). (2019), "Agency costs and tax planning when the government is a major Shareholder", *Journal of Accounting and Economics*, Vol. 67 No.2, pp. 255–277. https://doi.org/10.1016/j.jacceco.2018.10.002.

Chen, H., Tang, S., Wu, D. and Yang, D. (2021), "The political dynamics of corporate tax avoidance: The Chinese experience", *The Accounting Review*, Vol. 96 No.5, pp. 157-180. https://doi.org/10.2308/TAR-2017-0601.

D'Avino, C. (2023), "Counteracting offshore tax evasion: Evidence from the foreign account tax compliance act", *International Review of Law and Economics*, Vol. 73, 106126. https://doi.org/10.1016/j.irle.2023.106126.

Desai, M. A. and Dharmapala, D. (2006), "Corporate tax avoidance and high-powered incentives", *Journal of Financial Economics*, Vol. 79 No. 1, pp. 145–179. https://doi.org/10.1016/j.jfineco.2005.02.002.

Edwards, A., Schwab, C. and Shevlin, T. (2016), "Financial constraints and cash tax savings", *The Accounting Review*, Vol. 91 No.3, pp. 859–881. https://doi.org/10.2308/accr-51282.

Gupta, P., Hooda, A., Jeyaraj, A., Seddon, J. J. and Dwivedi, Y. K. (2024). "Trust, risk, privacy and security in e-Government use: Insights from a MASEM analysis", *Information Systems Frontiers*, pp. 1-17. https://doi.org/10.1007/s10796-024-10497-8.

He, Y. and Yi, Y. (2023), "Digitalization of tax administration and corporate performance: Evidence from China", *International Review of Financial Analysis*, Vol. 90, 102859. https://doi.org/10.1016/j.irfa.2023.102859.

Kerr, J. N. (2019), "Transparency, Information Shocks, and Tax Avoidance", *Contemporary Accounting Research*, Vol. 36 No. 2, pp. 1146–1183. https://doi.org/10.1111/1911-3846.12449.

Li, J., Wang, X. and Wu, Y. (2020), "Can government improve tax compliance by adopting advanced information technology? Evidence from the Golden Tax Project III in China", *Economic Modelling*, Vol. 93, pp. 384-397. https://doi.org/10.1016/j.econmod.2020.08.009.

Luo, Y. (2022), "A general framework of digitization risks in international business", *Journal of International Business Studies*, Vol. 53 No. 2, pp. 344–361. https://doi.org/10.1057/s41267-021-00448-9.

Mergel, I., Edelmann, N. and Haug, N. (2019). "Defining digital transformation: Results from expert interviews", *Government information quarterly*, Vol. 36 NO. 4, 101385. https://doi.org/10.1016/j.giq.2019.06.002.

Na, K. and Yan, W. (2022), "Languages and corporate tax avoidance", *Review of Accounting Studies*, Vol. 27, pp. 148–184. https://doi.org/10.1007/s11142-021-09596-7.

Slemrod, J., and Yitzhaki, S. (2002). "Tax avoidance, evasion, and administration", *Handbook of public economics*, Elsevier, Vol. 3, pp. 1423-1470. https://doi.org/10.1016/S1573-4420(02)80026-X.

Svetlozarova Nikolova, B. (2023), "Modernization of Approaches in Tax Control", In: Tax Audit and Taxation in the Paradigm of Sustainable Development, *Contributions to Management Science*. Springer, Cham. https://doi.org/10.1007/978-3-031-32126-9_9.

Tang, T., Mo, P.L.L. and Chan, K.H. (2017), "Tax collector or tax avoider? An investigation of intergovernmental agency conflicts", *The Accounting Review*, Vol. 92 No.2, pp. 247-270. https://doi.org/10.2308/accr-51526.

Zhou, S., Zhou, P. and Ji, H. (2022), "Can digital transformation alleviate corporate tax stickiness: The mediation effect of tax avoidance", *Technological Forecasting and Social Change*, Vol. 184*1*, 122028. https://doi.org/10.1016/j.techfore.2022.122028.

Zídková, H., Arltová, M. and Josková, K. (2024), "Does the level of e-government affect value-added tax collection? A study conducted among the European Union Member States", *Policy & Internet*, https://dx.doi.org/10.1002/poi3.389.

Appendix 1: Pilot provincial regions or cities for GTP IV from 2021 to 2022

Date	Pilot provincial regions or cities
01.12.2021	Inner Mongolia, Shanghai, Guangdong (except Shenzhen)
28.10.2022	Sichuan
30.11.2022	Xiamen city

Appendix 2: Definitions of variables

Dependent variables in the baseline regression analysis	
DD_BTD	The residual book-tax difference estimated from the following regression: $$BTD_{i,t} = \alpha_1 TACC_{i,t} + \mu_i + \varepsilon_{i,t}$$ where *BTD* is the total book-tax difference, computed as pre-tax financial income minus taxable income scaled by the lagged total assets. Taxable income is calculated as income tax expense minus deferred tax expense, divided by the nominal tax rate. *TACC* is total accruals scaled by the lagged total assets. The total accruals are calculated as the operating profit minus operating cash flows. A higher *DD_BTD* represents a higher level of tax avoidance.
Key variables that compose the DID estimator	
Treat	1 if a firm is a treatment firm, and 0 if a firm is a control firm. Treatment firms are defined as those located in cities or provinces where GTP IV (i.e., Golden Tax Project IV) was implemented; the control firms are those that are not subject to the reform during the sample period.
Post	1 if a firm is in the post-GTP IV period (i.e., 2022), and 0 if a firm is in the pre-GTP IV period (i.e., 2021).
Moderating variables	
SA_index	Per Hadlock and Pierce (2010), *SA_index* is defined as follows: $$SA\ index = -0.737 \times size + 0.043 \times size^2 - 0.040 \times age$$ where *size* is the natural logarithm of a firm's book value of total assets, and *age* is the number of years for which the firm has been listed.
SOE	1 if a firm's largest shareholder is a central or local government or a government-controlled enterprise, and 0 otherwise.
Opacity	The absolute value of abnormal accruals of a firm in a fiscal year, which is estimated using the cross-sectional version of modified Jones model with at least 20 firm-year observations required for each industry-year.
Std_return	The annualized standard deviation of a firm's daily stock returns over a fiscal year.
SIZE	The natural logarithm of a firm's market value of total equity at the end of a fiscal year.
Matching covariates	
SIZE	The natural logarithm of a firm's market value of total equity at the end of a fiscal year.
ROA	Earnings before interests and taxes over a fiscal year, divided by the total assets, at the end of the year.
BM	The book value of equity, divided by the market value of equity, at the end of a fiscal year.

LEV	The total debt divided by the total assets at the end of a fiscal year.
Std_return	The annualized standard deviation of a firm's daily stock returns over a fiscal year.
Indp	The number of independent directors as a fraction of the total directors on the board of a firm at the end of a year.
Control variables	
SIZE	The natural logarithm of a firm's market value of total equity at the end of a fiscal year.
ROA	Earnings before interests and taxes over a fiscal year, divided by the total assets, at the end of the year.
LEV	The total debt divided by the total assets at the end of a fiscal year.
BM	The book value of equity, divided by the market value of equity, at the end of a fiscal year.
NOL	The net operating losses. Since Chinese firms do not report the tax benefits from net operating losses in the balance sheet, we use a continuous variable, equal to the natural logarithm of 1 plus the absolute value of accumulated pre-tax losses (in billions) reported in the last five years, to proxy for the net operating losses. *NOL* equals 0 if the accumulated pre-tax losses are positive.
Cash	The cash and cash equivalents divided by the total assets at the end of a fiscal year.
PPE	The fixed assets divided by the total assets at the end of a fiscal year.
Abs_DA	The absolute value of abnormal accruals of a firm in a fiscal year, which is estimated using the cross-sectional version of modified Jones model with at least 20 firm-year observations required for each industry-year.
Indp	The number of independent directors as a fraction of the total directors on the board of a firm at the end of a fiscal year.
Institution	The shares held by all institutional investors of a firm, divided by the total shares outstanding by the firm, at the end of a fiscal year.
Duality	1 if CEO and the chairman of board of directors are the same person, and 0 otherwise.
SOE	1 if a firm's largest shareholder is a central or local government or a government-controlled enterprise, and 0 otherwise.
BIG4	1 if a firm's financial statement is audited by one of the big-4 auditors, and 0 otherwise.
Std_return	The annualized standard deviation of a firm's daily stock returns over a fiscal year.

Table 1: Sample selection procedure

Initial firm-year observations which cover companies listed on the Shenzhen or Shanghai Stock Exchange for the period 2021-2022	10,270
Less: observations in the financial industry	(257)
Less: observations with missing values in tax expense or with negative pre-tax income	(2,308)
Less: observations of which the transaction status is “special treatment”, “suspension from trading” or “particular transfer”	(75)
Less: firms located in regions (i.e., Sichuan and Xiamen) that implemented GTP IV in 2022	(382)
Less: observations that are missing in the values of regressors used in the baseline regression analysis	(2,133)
Firm-year observations (unique firms) available for the propensity-score matching	5,115 (2,810)
Firm-year observations (unique firms) after the propensity-score matching	1,846 (1,319)

Table 2: Propensity-score matching

Panel A: Univariate tests of covariate balance

Variables	Matching statuses	No. of firm-years	Mean for treatment firms	Mean for control firms	Standardized bias	t-stat.
SIZE	Unmatched sample	5,115	23.296	23.213	7.1	1.97**
	Matched sample	1,846	23.296	23.207	-1.0	-0.21
ROA	Unmatched sample	5,115	0.067	0.069	-3.7	-1.03
	Matched sample	1,846	0.067	0.066	1.8	0.40
BM	Unmatched sample	5,115	0.577	0.573	1.6	0.44
	Matched sample	1,846	0.577	0.577	0.0	0.00
LEV	Unmatched sample	5,115	0.409	0.399	5.6	1.55
	Matched sample	1,846	0.409	0.417	-4.4	-0.93
Indp	Unmatched sample	5,115	0.383	0.379	8.3	2.33**
	Matched sample	1,846	0.383	0.384	-1.2	-0.25
Std_return	Unmatched sample	5,115	0.474	0.482	-3.7	-1.05
	Matched sample	1,846	0.474	0.468	3.0	0.70

Notes: This table reports the descriptive statistics of matching covariates for the treatment sample and the control sample. The results of the two-sample tests of mean differences, and the results of the standardized bias, for the covariates are provided for the pre-matched sample and post-matched sample, respectively. All the covariates are defined in Appendix 1. ***, **, * denote the two-tailed statistical significance at the 1%, 5%, and 10% levels, respectively.

Panel B: Multivariate tests of covariate balance

Variables	*Treat* (1)
SIZE	0.009
	(0.15)
ROA	0.471
	(0.37)
BM	0.168
	(0.63)
LEV	-0.371
	(-0.92)
Std_return	0.244
	(0.88)
Indp	-0.280
	(-0.27)
Constant	-0.573
	(-0.33)
Year-fixed effects	included
Industry-fixed effects	included
Observations	1,846
Pseudo R^2	0.004

Notes: This table reports the results of the logistic regression run based on the post-matched sample and aimed at comparing firm characteristics between the treatment firms and control firms after the matching. All the matching covariates are defined in Appendix 1. Year dummies and industry dummies are included in the regressions, but their results are not reported for the sake of simplicity. *t*-statistics in parentheses are based on the robust standard errors clustered by firm. ***, **, * represent the two-tailed statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 3: Univariate statistics

Panel A: Descriptive statistics

Variables	N	Mean	Min.	25%	Median	75%	Max.	Std.
DD_BTD	1,846	0.016	-0.097	0.013	0.013	0.038	0.211	0.047
Treat	1,846	0.500	0.000	0.500	0.500	1.000	1.000	0.500
Post	1,846	0.478	0.000	0.000	0.000	1.000	1.000	0.500
SIZE	1,846	23.302	21.391	23.081	23.081	23.990	26.830	1.198
ROA	1,846	0.066	0.001	0.054	0.054	0.087	0.270	0.049
LEV	1,846	0.413	0.068	0.408	0.408	0.548	0.824	0.185
BM	1,846	0.577	0.098	0.532	0.532	0.778	1.285	0.285
NOL	1,846	1.960	0.000	0.000	0.000	0.000	21.433	5.827
Cash	1,846	0.172	0.014	0.144	0.144	0.226	0.667	0.120
PPE	1,846	0.179	0.001	0.151	0.151	0.255	0.642	0.139
Abs_DA	1,846	0.050	0.000	0.037	0.037	0.069	0.238	0.047
Indp	1,846	0.384	0.333	0.375	0.375	0.429	0.571	0.055
Institution	1,846	65.545	0.417	52.938	52.938	75.062	544.862	67.318
Duality	1,846	0.348	0.000	0.000	0.000	1.000	1.000	0.476
SOE	1,846	0.308	0.000	0.000	0.000	1.000	1.000	0.462
BIG4	1,846	0.082	0.000	0.000	0.000	0.000	1.000	0.274
Std_return	1,846	0.467	0.195	0.457	0.457	0.538	1.415	0.147

Notes: This table tabulates the descriptive statistics of variables used for the difference-in-differences regression analysis. The sample consists of a post-matched sample of 1,846 firm-years and covers the years 2021-2022. All the variables are defined in Appendix 1. All the continuous variables are winsorized at the 1% and 99% levels, respectively.

Panel B: Correlation matrix

Variables	*DD_BTD*	*SIZE*	*ROA*	*LEV*	*BM*	*NOL*	*Cash*	*PPE*
DD_BTD	1.000							
SIZE	0.002	1.000						
ROA	0.380***	0.170***	1.000					
LEV	-0.237***	0.439***	-0.285***	1.000				
BM	-0.281***	0.191***	-0.351***	0.440***	1.000			
NOL	-0.081***	-0.097***	-0.116***	0.072***	-0.023	1.000		
Cash	0.029	-0.123***	0.146***	-0.353***	-0.258***	-0.068***	1.000	
PPE	-0.114***	-0.005	0.057**	-0.031	0.071***	0.005	-0.278***	1.000
Abs_DA	0.258***	-0.001	0.181***	0.006	-0.124***	-0.014	0.009	-0.146***
Indp	0.008	0.048**	0.057**	0.000	-0.041*	-0.015	-0.010	-0.094***
Institution	0.040*	0.215***	0.109***	0.051**	-0.038*	-0.115***	0.074***	0.050**
Duality	0.077***	-0.132***	0.015	-0.137***	-0.236***	-0.028	0.036	-0.077***
SOE	-0.116***	0.299***	-0.143***	0.312***	0.463***	-0.012	-0.063***	0.084***
BIG4	-0.031	0.355***	0.006	0.101***	0.114***	-0.021	-0.076***	0.004
Std_return	0.154***	-0.069***	0.152***	-0.083***	-0.437***	0.037	0.080***	-0.011

Variables	*Abs_DA*	*Indp*	*Institution*	*Duality*	*SOE*	*BIG4*	*Std_return*
Abs_DA	1.000						
Indp	0.008	1.000					
Institution	0.010	-0.026	1.000				
Duality	0.034	0.127***	-0.041*	1.000			
SOE	-0.061***	-0.056**	0.130***	-0.340***	1.000		
BIG4	-0.038*	0.025	0.155***	-0.060***	0.096***	1.000	
Std_return	0.133***	-0.035	0.175***	0.118***	-0.249***	-0.111***	1.000

Notes: This table presents the results for the Spearman correlations. The correlation matrix involves the variables used for the difference-in-differences regression analysis. The sample consists of a post-matched sample of 1,846 firm-years and covers the years 2021-2022. All the variables are defined in Appendix 1. ***, **, * represent the two-tailed statistical significance at the 1%, 5%, and 10% levels, respectively, for the correlation coefficients.

Table 4: Baseline difference-in-differences regression analysis

Variables	*DD_BTD* (1)
*Treat*Post*	-0.009***
	(-2.61)
Treat	0.004
	(0.80)
SIZE	0.002
	(1.14)
ROA	0.275***
	(7.18)
LEV	-0.043***
	(-5.28)
BM	-0.015**
	(-2.58)
NOL	-0.000
	(-1.52)
Cash	-0.063***
	(-5.85)
PPE	-0.057***
	(-6.25)
Abs_DA	0.172***
	(4.59)
Indp	-0.022
	(-1.20)
Institution	0.000
	(0.47)
Duality	0.003
	(1.25)
SOE	0.007***
	(2.70)
BIG4	-0.004
	(-0.83)
Std_return	0.011
	(1.43)
Constant	0.025
	(0.78)
Year-fixed effects	Included
Industry-fixed effects	Included
Region-fixed effects	Included
Observations	1,846
Adj.R^2	0.250

Notes: This table reports the results of DID regression analysis of the effect of the implementation of GTP IV on corporate tax avoidance. The sample period ranges from 2021 to 2022. All the variables are defined in Appendix 1. Year dummies, industry dummies, and region dummies (i.e., dummies as to provinces as well as centrally administered cities by the Chinese government) are included in the regression, but their results are not reported for the sake of brevity. *t*-statistics in parentheses are based on robust standard errors clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 5: Moderation analyses

Variables	*DD_BTD*									
	Firms with tighter financial constraints	Firms with fewer financial constraints	Non-SOE firms	SOE firms	Firms with higher financial opacity	Firms with lower financial opacity	Firms with higher risk	Firms with lower risk	Larger firms	Smaller firms
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
*Treat*Post*	-0.012**	-0.005	-0.015***	0.000	-0.015**	-0.005	-0.018***	-0.001	-0.005	-0.014***
	(-2.30)	(-1.00)	(-3.38)	(0.06)	(-2.34)	(-1.21)	(-3.21)	(-0.30)	(-1.03)	(-2.67)
Controls	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Constant	-0.046	0.099*	0.022	0.075	0.007	0.054	-0.021	0.023	0.025	0.063
	(-1.06)	(1.94)	(0.53)	(1.45)	(0.14)	(1.43)	(-0.40)	(0.57)	(0.50)	(0.79)
Year-fixed effects	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Industry-fixed effects	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Region-fixed effects	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
Observations	949	897	1,277	569	945	901	914	932	957	889
Adj.R^2	0.244	0.259	0.255	0.270	0.287	0.141	0.299	0.183	0.231	0.268

Notes: This table reports the results from the moderation analyses of the baseline regression results. The sample period covers the years 2021-2022. All the variables are defined in Appendix 1. The difference-in-differences regressions are run separately on the higher-*SA_index* (non-*SOE*, higher-*Opacity*, higher-*Std_return*, or lower-*SIZE*) subsample and the lower-*SA_index* (*SOE*, lower-*Opacity*, lower-*Std_return*, or higher-*SIZE*) subsample, respectively. The control variables, year dummies, industry dummies, and region dummies (i.e., dummies as to provinces as well as centrally administered cities by the Chinese government) are included in the regression, but their results are not reported for the sake of brevity. *t*-statistics in parentheses are based on robust standard errors clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Figure 1: Kernel density distribution of propensity scores

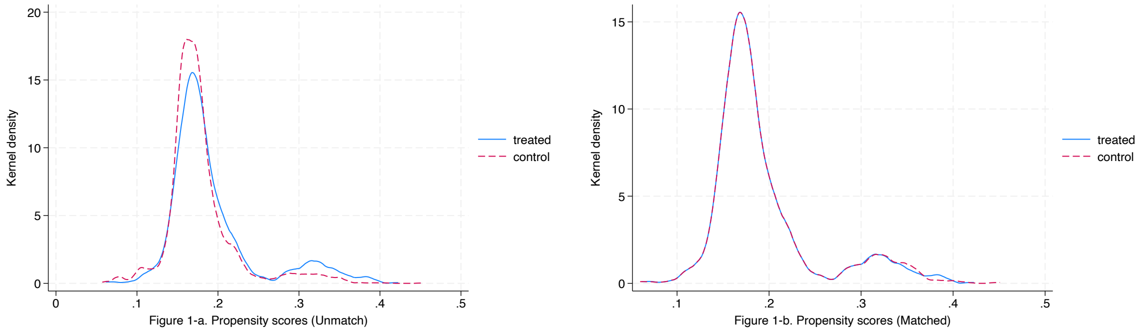
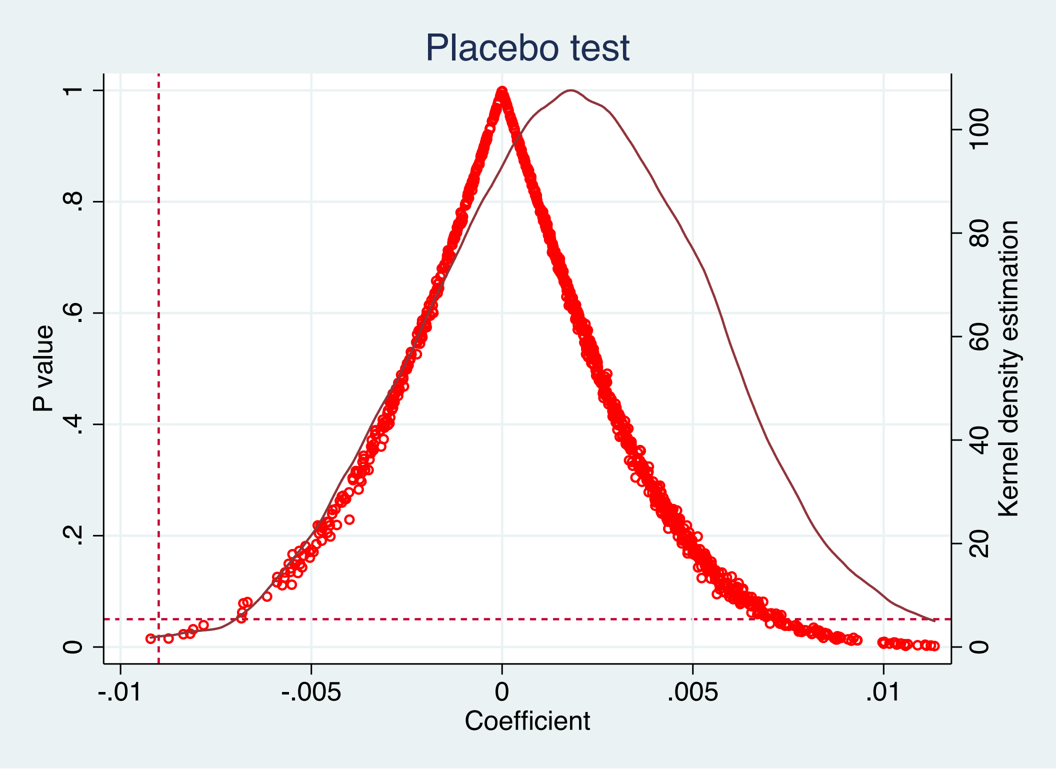


Figure 1-a. Propensity scores (Unmatch)

Figure 1-b. Propensity scores (Matched)

Notes: Figure 1 shows the kernel density distributions of propensity scores for the treatment group and control group before and after the propensity-score matching. The x-axis represents the propensity scores; the y-axis represents the probability density. Figure 1-a (b) displays the distribution of propensity scores before (after) the matching. The treated firms are defined as those located in areas where GTP IV was implemented during the sample period of 2021-2022; the control firms are those not subject to the policy during the sample period. The solid (dashed) curves represent the distribution of propensity scores for the treated (control) firms. Each treated firm is matched with a control firm using the closest propensity score without replacement.

Figure 2: Distribution of the 1,000 coefficient estimates in a placebo test



Notes: The X-axis indicates the coefficients for the interaction term that are estimated based on the pseudo sample. The left Y-axis indicates the *p*-values of the coefficients. The right Y-axis indicates the kernel density of the estimated coefficients. The red dots (solid curve) represent(s) the kernel density (*p*-values) corresponding to the estimated coefficients. The vertical dashed line represents the estimated coefficient of the interaction term in the baseline DID regression analysis. The horizontal dashed line represents the significance level of *p*=0.05 for a DID estimator.

Anthropomorphic Generative AI chatbots for enhancing customer engagement, experience and recommendation

Abhishek Behl (Keele Business School; Keele University, UK), Aman Kumar (IIM Ranchi), Amit Shankar (IIM Vishakapatnam, India) and Payal Kapoor (Management Development Institute Gurgaon, India)

Completed Research

Abstract (around 150 words)

This research focuses on developing and testing a conceptual model that explores customer behavioural responses (engagement, experience, recommendation) towards generative AI-enabled chatbots. It highlights the significant influence of anthropomorphic characteristics in enhancing perceptions of competence and warmth, further enhancing perceived authenticity. Additionally, the study investigates how the need for social interactions moderates these relationships. The study utilized a self-administered questionnaire distributed on Prolific Academic to gather data from 282 eligible participants worldwide. The study uses a structural equation modelling approach to answer the research questions. The findings reveal that anthropomorphic characteristics of generative AI-enabled chatbots are positively associated with perceived competence. Moreover, the findings show that generative AI-enabled chatbots' perceived competence and warmth are significantly associated with perceived authenticity. Furthermore, the results highlight that the perceived authenticity of generative AI-enabled chatbots is positively associated with customer engagement, experience and recommendation. Finally, the results illustrate that the need for social interaction significantly moderates the impact of anthropomorphic characteristics on perceived warmth.

Keywords: Artificial intelligence, Generative AI, Anthropomorphism, Social response theory, Chatbots

1.0 Introduction

The emergence of artificial intelligence (AI) has significantly reshaped the landscape of business performance tools (Hermann & Puntoni, 2024). Information and Communication Technology (ICT) has facilitated the integration of AI, transforming these tools into increasingly sophisticated systems (Bahoo et al., 2023). These AI-powered systems move beyond traditional automation, exhibiting capabilities akin to multicellular thinking through machine learning algorithms (Nazir et al., 2023). Machine learning enables these systems to analyze data, identify patterns, and adapt processes by mimicking human cognitive functions like problem-solving, planning, and learning (Nazir et al., 2023). The impact of AI extends beyond individual tools, revolutionizing core functionalities across various business sectors, including retail (Fu et al., 2023), hospitality (Wang & Uysal, 2024), tourism (Kong et al., 2023), banking (Rahman et al., 2023), healthcare (Dicuonzo et al., 2023), education (Celik, 2023), and manufacturing (Dey, 2023). One particularly impactful subfield of AI is generative AI, which utilizes AI's capabilities to create entirely new content. Generative AI refers to a category of artificial intelligence systems designed to create new content, such as text, images, music, or code, that resembles human-generated output (Cui et al., 2024; Dwivedi et al., 2023). These systems use algorithms and models, particularly deep learning techniques like neural networks, to generate data that mimics the patterns and structures found in existing datasets (Sætra, 2023).

Generative AI is poised to bring transformative and disruptive changes across all business sectors. According to a report by McKinsey & Company (2024), generative AI could contribute an astounding $2.6 to $4.4 trillion annually to the global economy, with the most significant impacts expected in customer operations, marketing and sales, software engineering, and research and development (R&D). The current capabilities of generative AI and other advanced technologies can potentially automate work activities that currently consume 60 to 70 percent of employees' time (McKinsey & Company, 2023). Moreover, generative AI has the potential to drive labour productivity growth by 0.1 to 0.6 percent annually through 2040, contingent upon the pace of technology adoption and the effective redeployment of worker time into other productive activities (McKinsey & Company, 2023). The transformative potential of generative AI is vast, promising to enhance efficiency and productivity and fundamentally reshape how businesses operate and compete in the global market.

Generative AI stands at the forefront of a revolution in consumer marketing by offering unparalleled capabilities for hyper-personalization, efficiency, and customer engagement (Harkness et al., 2023). While the technology is nascent, early applications in marketing tasks like content generation, customer segmentation, and product development suggest a disruptive potential (Jaboob et al., 2024; Chakraborty et al., 2024). As businesses begin to harness the transformative power of generative AI, it becomes evident that this technology can significantly elevate marketing practices by driving substantial operational efficiencies and competitive advantages (Sohn et al., 2021). However, the full potential of generative AI in consumer marketing remains largely untapped, warranting further research. Despite the transformative potential of generative AI in consumer marketing, research on understanding consumer behaviour towards its usage remains in a very nascent stage (Gupta et al., 2024). Current studies primarily focus on generative AI's technological capabilities and operational efficiencies from a conceptual perspective (Hermann & Puntoni, 2024; Dwivedi et al., 2023; Sætra, 2023), leaving a significant gap in the literature towards comprehending how consumers perceive and interact with generative AI-enabled chatbots. In the context of technology-human interaction, the need for social interaction emerges as a crucial factor (Belanche et al., 2021). As individuals increasingly rely on these technologies for communication and engagement, understanding how the desire for social connection impacts their interactions becomes essential (Flavián et al., 2024). Therefore, this study considers the need for social interaction as a moderator. Notably, the existing literature has largely overlooked this aspect, focusing primarily on technical capabilities and user satisfaction without delving into the social dimensions of these interactions. This study aims to address these research gaps by exploring how generative AI-specific characteristics influence customer behavioural responses. The research questions for this study are given below:

RQ1: How do anthropomorphic characteristics of generative AI chatbots influence customers' perception of competence and warmth?

RQ2: How do customers' perceptions of competence and warmth towards generative AI chatbots influence perceived authenticity?

RQ3: How does the perceived authenticity of generative AI chatbots influence customer behavioural responses (engagement, experience, recommendation)?

RQ4: How do customers' perceptions of competence and warmth towards generative AI chatbots vary at different levels in need of social interaction?

This study leverages Social Response Theory (SRT) (Moon, 2000) to elucidate the factors influencing consumer behavioural responses (engagement, experience, recommendation) towards generative AI chatbots and their anthropomorphic characteristics. SRT provides a robust theoretical framework due to its established capacity to explain how individuals interact and respond to technology (Nguyen et al., 2023; Huang & Lin, 2011). By applying SRT, we aim to understand how anthropomorphic elements within generative AI chatbots influence customer behaviour. By examining how anthropomorphic elements influence social responses, this study aims to elucidate the impact of perceived competence and warmth on the effectiveness of generative AI chatbots. Thus, social response theory is instrumental in this research, contributing to the emerging literature on generative AI in consumer marketing.

This study makes a significant contribution to the growing body of literature on generative AI and social response theory. This study explores how anthropomorphic characteristics in generative AI-enabled chatbots influence customer behavioural responses, which deepens our understanding of how users perceive and interact with generative AI systems. For organizations, the findings of this research offer actionable insights into how generative AI-enabled chatbots can be designed and deployed to optimize customer engagement and experience. Further, this research helps organizations leverage AI to improve customer experience and drive positive behavioural outcomes.

2. Literature review

2.1. Generative AI

The adoption of generative AI is growing rapidly among consumers and businesses, finding applications across various functional areas, including sales and marketing (Kshetri et al., 2023; Dwivedi et al., 2023). A range of Gen AI tools, including Bard, ChatGPT, Jasper, and others, are now employed to develop advertising content (text, images, and videos), craft digital marketing strategies, create chatbot solutions, generate blog posts, and even design sales training programs (Ooi et al., 2023). Generative AI leverages algorithms trained on vast amounts of data (Dwivedi et al., 2023; Sætra, 2023). These algorithms function by predicting the most likely subsequent word or pixel in a sequence, enabling them to produce entirely new creative content (Jaboob et al., 2024). Driven by the potential for improved efficiency and customer experience, a growing number of brands are deploying generative AI-enabled chatbots for various customer service tasks (Cui et al., 2024). These chatbots can handle a range of interactions, including resolving complaints, assisting with product selection, and providing after-sales support (Ferraro et al., 2024). A comprehensive review of the extant literature reveals a paucity of research directly investigating customer behavioural responses towards generative AI-enabled chatbots (Ooi et al., 2023). Moreover, recent literature on generative AI highlighted the need to explore and examine the applications of generative AI in various contexts, including marketing (Huy et al., 2024; Ooi et al., 2023). Despite this extensive body of research on AI-enabled chatbots, there is limited focus on generative AI-enabled chatbots, which are capable of dynamically generating personalized content and engaging in more sophisticated, contextually aware conversations. Therefore, this study aims to fill this literature gap and enrich the emerging literature on generative AI.

2.2. Social response theory (SRT)

Social response theory posits that when individuals encounter technology exhibiting human-like characteristics, they tend to respond with social behaviours and attribute human qualities to the technology (Nguyen et al., 2023; Moon, 2000). Consequently, the social norms governing interpersonal interactions become relevant and applicable to human-technology interactions (Huang & Lee, 2022; Pérez-Vega et al., 2018). The underlying cause of these social responses lies in the unconscious thought processes of humans, triggered by contextual social cues (Premathilake & li, 2024). These cues invoke established social scripts and expected behaviours based on past interactions (Huang & Lin, 2011). These social scripts, often simplified versions of how we navigate social interactions, lead users to respond to the technology with social behaviours (Premathilake & li, 2024; Li & Li, 2014). SRT posits a direct correlation between a chatbot's anthropomorphism and the likelihood of users attributing human characteristics and social rules to it (Nguyen et al., 2023; Huang & Lee, 2022; Huang & Lin, 2011). This theory is highly relevant for examining behavioral responses toward generative AI-enabled chatbots, as these bots mimic human conversational abilities and foster interactions resembling human communication. With advanced language capabilities, generative AI chatbots blur the lines between human and machine, prompting users to attribute personality, emotions, and intentions to them. This study builds upon SRT to investigate how user behavioural intentions towards generative AI-enabled chatbots are influenced by their perceived level of anthropomorphic characteristics. In other words, the more human-like a generative AI-enabled chatbot appears or behaves, the greater the tendency for users to perceive it as a social actor and respond accordingly. Consequently, SRT not only explains users' natural inclination to treat AI as social entities but also provides a lens through which to study how these interactions influence broader user behaviours and decisions in various contexts, making it an ideal theory for understanding human responses to generative AI-enabled chatbots. Therefore, this study proposes a conceptual model to examine customer behavioural responses—namely engagement, experience, and recommendation—towards generative AI-enabled chatbots. At the core of this framework is the hypothesis that anthropomorphic characteristics of chatbots significantly enhance perceptions of competence and warmth. We argue that these enhanced perceptions improve the perceived authenticity of user interactions with chatbots. Furthermore, we propose that perceived authenticity positively influences customer behavioural responses, thereby fostering greater engagement, enriching user experiences, and encouraging favourable recommendations. Additionally, we propose that the need for social interactions is a crucial moderator.

3. Hypotheses development

3.1. Anthropomorphic characteristics

The phenomenon of attributing human characteristics to non-human entities is termed anthropomorphism (Nguyen et al., 2023; Roy & Naidoo, 2021). It encompasses a wide spectrum of human-like qualities, ranging from physical appearance to complex mental states like reasoning, moral judgment, and emotional experience (Pelau et al., 2021). Research suggests a natural human tendency to perceive human attributes in non-human entities, particularly when imbued with human-like features (Premathilake & Li, 2024). This fundamental premise is supported by findings that human attributes in chatbots trigger perceptions of humanness (Belanche et al., 2021). Therefore, customers tend to interact with these chatbots as they would with humans, exhibiting social behaviours towards them (Roy & Naidoo, 2021). Moreover, research indicates that anthropomorphic characteristics enhance human-computer interactions, increase customers' perceptions of chatbot perceived competence and warmth, and ultimately improve user satisfaction with the interaction process (Nguyen et al., 2023; Cheng, 2022). Therefore, when users perceive generative AI-enabled chatbots as possessing these anthropomorphic attributes, they are more likely to view them as competent problem solvers and empathetic companions.

H1a: *Anthropomorphic characteristics of generative AI-enabled chatbot are significantly associated with generative AI-enabled chatbot perceived competence.*

H1b: *Anthropomorphic characteristics of generative AI-enabled chatbot are significantly associated with generative AI-enabled chatbot perceived warmth.*

3.2. Perceived competence

Perceived competence pertains to an individual's belief that the other party possesses the ability to benefit or harm them through the execution of negative or positive intentions (Cheng et al., 2022). This perception of competence is a functional expectation of users, encompassing the intelligence, efficiency, functionality, proficiency, and skill perceived in an intelligent technology (Harris-Watson et al., 2023; Belanche et al., 2021). Customers anticipate generative AI-enabled chatbots to accurately diagnose their inquiries and deliver valuable information through high-quality interactions that are seamless and comprehensive (Bai et al., 2024). Chatbots that demonstrate high prediction accuracy and a successful rate of user assistance are more likely to be perceived as highly competent (Yang et al., 2024). Moreover, consumers perceive these chatbots as credible when they exhibit strong problem-solving abilities and customization skills (Dwivedi et al., 2023). Existing research underscores the critical role of competence in promoting trust in chatbots (Dwivedi et al., 2023; Cheng et al., 2022). Anthropomorphic traits such as efficient task performance and accurate information delivery enhance perceptions of competence (Nguyen et al., 2023). Users are more likely to regard a chatbot as competent when it effectively mimics human conversation and demonstrates a deep understanding of context and nuance. This perceived authenticity arises from the chatbot's ability to provide relevant and coherent responses, thereby creating an engaging and believable interaction. Therefore, competence in generative AI-enabled chatbots may enhance the perceived authenticity of their interactions.

H2: *Generative AI-enabled chatbot perceived competence is significantly associated with the perceived authenticity of the generative AI-enabled chatbot.*

3.3. Perceived warmth

Perceived warmth encompasses emotional attributes such as empathy, understanding, and friendliness (Bai et al., 2024; Yang et al., 2024). It is primarily linked to emotional rather than cognitive responses (Peng et al., 2022). When a task necessitates a high degree of warmth, an effective server must demonstrate emotional intelligence and empathy (Cheng et al., 2022). Despite the increasing capability of intelligent technologies to interact with customers and transform their experiences, customers often still prefer human workers over AI when tasks involve significant emotional engagement (Peng et al., 2022; Belanche et al., 2021). Perceived warmth in generative AI-enabled chatbots can promote social proximity between users and the technology during service interactions, thereby enhancing the authenticity of the chatbot and building stronger connections. Therefore, anthropomorphic traits such as human-like voices and empathetic interactions promote warmth in the context of generative AI-enabled chatbots. Furthermore, the warmth generated by these traits enhances the perceived authenticity of the chatbot, deepening customer positive behavioural intentions.

***H3:** Generative AI-enabled chatbot perceived warmth is significantly associated with the perceived authenticity of the generative AI-enabled chatbot.*

3.4. Perceived Authenticity

Authenticity can be conceptualized as *"the real thing"* (Nguyen et al., 2023). In the context of chatbots, authenticity refers to a customer's ability to engage with the chatbot naturally (Rese et al., 2020). However, algorithms are often perceived as less authentic than human counterparts (Renier et al., 2021). Anthropomorphic cues, such as language and visual elements, can signal a chatbot's authenticity to customers, enhancing engagement and eliciting social responses (Blut et al., 2021). Also, when interacting with chatbots exhibiting higher levels of anthropomorphism, customers are likely to perceive them as more authentic (Nguyen et al., 2023). Research has shown that the social presence of an anthropomorphic chatbot increases its perceived authenticity and promotes behavioural intentions such as engagement, experience and recommendation (Dwivedi et al., 2023; Nguyen et al., 2023). Engagement refers to the degree of involvement, interaction, and attention a customer invests in their interaction with a chatbot (Nguyen et al., 2023; Jiang et al., 2022). Experience encompasses the overall impression and satisfaction derived from the interaction with the chatbot, including factors such as usability, efficiency, and emotional resonance (Dwivedi et al., 2023; Kushwah et al., 2021). Recommendation intention pertains to the likelihood of a customer endorsing or advocating the chatbot to others based on their positive experience (Dwivedi et al., 2023). When users perceive a chatbot as authentic, they are more likely to engage with it actively, leading to deeper and more meaningful interactions. This sense of authenticity enhances the overall customer experience, as users feel understood and valued, resulting in increased satisfaction and loyalty. Furthermore, a strong perception of authenticity often translates into positive word-of-mouth, where satisfied users are more inclined to recommend the chatbot to others. Therefore, perceived authenticity plays a crucial role in influencing customer behavioural intentions (Lee & Kim, 2024; Rese et al., 2020), including engagement, experience, and recommendation.

H4: *Perceived authenticity of generative AI-enabled chatbot is significantly associated with* ***a)*** *customer engagement,* ***b)*** *customer experience,* ***c)*** *customer recommendation.*

3.5. Moderating role of the need for social interaction

This research proposes that the need for social interaction acts as a moderating factor, amplifying the influence of anthropomorphic characteristics on perceived generative AI competence and warmth. The need for social interaction is defined as the degree to which consumers value human interaction during service encounters (Tuguinay et al., 2022). Previous investigations within service robotics have elucidated that due to service robots' emulation of human-like behaviour, social interaction assumes particular salience in elucidating consumer responses to such technological advancements (Belanche et al., 2021). Consequently, the need for social interaction may be a key contingency variable within the proposed conceptual framework (Flavián et al. 2024). Specifically, individuals exhibiting a pronounced need for social interaction prioritize human contact in service interactions (Flavián et al., 2024; Tuguinay et al., 2022). Consequently, these specific consumer cohorts exhibit a diminished inclination towards embracing technological substitutes for human engagement, attributable to the absence of intrinsic motivation (Belanche et al., 2021). Therefore, anthropomorphic attributes inherent in generative AI-enabled chatbots fulfil consumers' social requisites and enhance their interpersonal engagements by engendering a perception akin to interacting with another individual. Consequently, consumers exhibiting a heightened inclination for social interaction will manifest greater sensitivity to enhancements in the anthropomorphic attributes of generative AI-enabled chatbots. This heightened sensitivity is attributed to the convergence of their technology-mediated service encounters towards a semblance of traditional human service provision, which aligns closely with their pronounced preferences. In essence, the affirmative impacts of anthropomorphic attributes of generative AI-enabled chatbots on perceptions of chatbot competence and warmth will be accentuated among consumers with an elevated need for social interaction.

H5a: *The need for social interaction moderates the association between anthropomorphic characteristics of generative AI-enabled chatbot and generative AI-enabled chatbot perceived competence.*

H5b: *The need for social interaction moderates the association between anthropomorphic characteristics of generative AI-enabled chatbot and generative AI-enabled chatbot perceived warmth.*

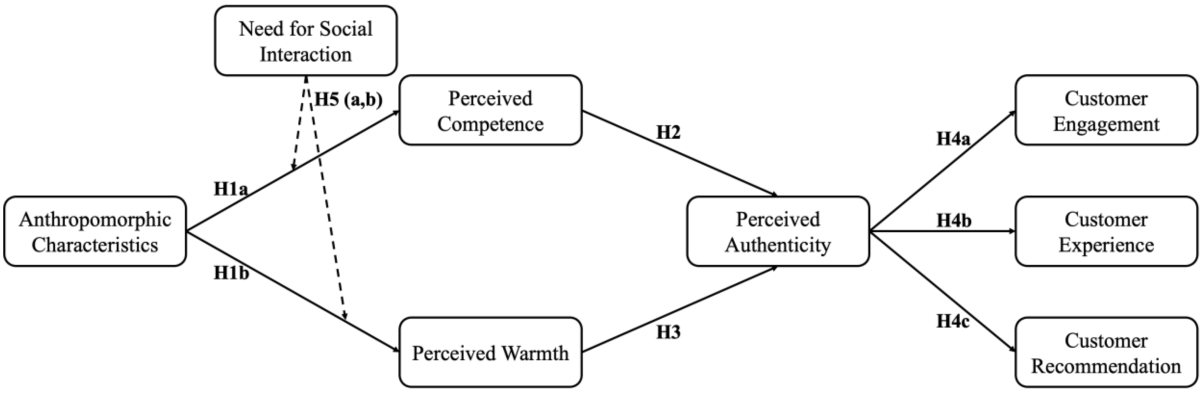


Figure 1: Conceptual framework

4. Research methods

4.1. Measures development and data collection

The proposed conceptual framework is shown in Figure 1. The questionnaire was developed using items obtained from previous literature. However, we ensured that the items suited the context of the study. Further, the variables and the sources of items used to measure the constructs. Appendix 1 contains the items used to measure the constructs. The questionnaire was sent to subject matter experts to verify the content and framing of the items. Additionally, a pilot study was conducted using 50 respondents. Minor changes to the questionnaire were made after receiving feedback from the pilot study and expert group. To assess our conceptual model, we utilized a cross-sectional survey method. We gathered data from customers from across the globe who had used generative AI-enabled chatbots, ensuring a diverse and representative sample. The data collection included respondents from various countries, such as the United States, Canada, the United Kingdom, Germany, Australia, India, and Brazil. The Prolific Academic platform was used to collect the data from the respondents. Out of 300 potential responses, 282 successfully passed attention and screening questions. Among the participants, 54.3% identified as female and 45.7% as male, reflecting a balanced gender representation. The age distribution revealed that 22.0% were aged 18 to 24 years, 38.3% fell within the 25 to 34 year range, and 39.7% were 35 years and older, indicating a predominantly young to middle-aged demographic.

4.2. Common method bias (CMB)

To mitigate CMB in our data collection, we applied procedural measures and statistical controls. Furthermore, we executed Harman’s single-factor test to evaluate CMB within our study. Results revealed that CMB accounted for below the 50% threshold (Podsakoff et al., 2003), signifying negligible influence within our findings. Furthermore, we included attention-check items throughout the survey to identify and exclude participants who may not have been fully engaged or attentive during the response process. Consequently, we can infer that participants engaged attentively and responded considerately to the survey questionnaire.

Table 1: Measurement model

Variables and Items	**Factor loading**	**Cronbach's alpha**	**Composite reliability**	**Average variance extracted**
Anthropomorphic Characteristics (Pelau et al., 2021)		0.93	0.93	0.69
AC1	0.80			
AC2	0.84			
AC3	0.84			
AC4	0.80			
AC5	0.88			
AC6	0.82			

Perceived Competence (Dwivedi et al., 2023)		0.88	0.87	0.64
CO1	0.84			
CO2	0.81			
CO3	0.81			
CO4	0.72			
Perceived Warmth (Dwivedi et al., 2023)		0.95	0.95	0.82
WA1	0.87			
WA2	0.94			
WA3	0.92			
WA4	0.91			
Perceived Authenticity (Nguyen et al., 2023)		0.83	0.77	0.52
PA1	0.81			
PA2	0.69			
PA3	0.65			
Customer Engagement (Nguyen et al., 2023)		0.90	0.91	0.77
ENG1	0.78			
ENG2	0.96			
ENG3	0.88			
Customer Experience (Dwivedi et al., 2023)		0.94	0.94	0.79
CEX1	0.89			
CEX2	0.90			
CEX3	0.88			
CEX4	0.88			
Customer Recommendation (Dwivedi et al., 2023)		0.83	0.84	0.63
CUR1	0.88			
CUR2	0.82			
CUR3	0.66			
Need for Social Interaction (Flavián et al., 2024)		0.93	0.93	0.76
INT1	0.96			
INT2	0.93			
INT3	0.81			
INT4	0.77			

Table 2: Discriminant validity (HTMT)

Variables	AC	CO	WA	PA	ENG	CEX	CUR	INT
Anthropomorphic Characteristics (AC)								
Perceived Competence (CO)	0.65							
Perceived Warmth (WA)	0.09	0.13						
Perceived Authenticity (PA)	0.63	0.79	0.22					
Customer Engagement (ENG)	0.61	0.60	0.13	0.67				
Customer Experience (CEX)	0.55	0.65	0.21	0.67	0.62			
Customer Recommendation (CUR)	0.66	0.84	0.10	0.80	0.57	0.71		
Need for Social Interaction (INT)	0.63	0.81	0.13	0.85	0.59	0.67	0.87	

5. Results

5.1. Measurement model

According to Hair et al. (2017), we assessed the measurement model by evaluating the reliability, convergent, and discriminant validity of the latent constructs. Confirmatory Factor Analysis (CFA) was conducted to scrutinize the reliability and validity of the study constructs. To assess reliability, Cronbach’s alpha values were computed for all constructs, each surpassing 0.7, affirming their reliability (Hair et al., 2017). Convergent and discriminant validity were established by examining the average variance extracted (AVE) values (all above 0.5) and composite reliability (CR) values (all above 0.7). Discriminant validity was evaluated using heterotrait-monotrait (HTMT) analysis (Henseler et al., 2015). The values presented in Table 2 all fall below the 0.9 cut-off value, suggesting that the constructs are distinct and capture unique variance.

Table 3: Path analysis

Paths	β	SE	T
Anthropomorphic Characteristics → Perceived Competence	0.686***	0.063	10.898
Anthropomorphic Characteristics → Perceived Warmth	-0.078ns	0.072	1.240
Perceived Competence → Perceived Authenticity	0.929***	0.078	14.642
Perceived Warmth → Perceived Authenticity	0.120**	0.041	3.159
Perceived Authenticity → Customer Engagement	0.670***	0.047	10.194
Perceived Authenticity → Customer Experience	0.755***	0.052	12.647
Perceived Authenticity → Customer Recommendation	0.927***	0.056	15.500
Note: *** denotes p<0.001; ** denotes p< 0.01; ns denotes non-significant			

5.2. Hypotheses testing

The outcomes of the path analysis, as presented in Table 3, suggest that the anthropomorphic characteristics (β=0.686***) of generative AI-enabled chatbots are positively associated with perceived competence. Therefore, supporting H1a. Moreover, the findings show that the perceived competence (β=0.929***) and perceived warmth (β=0.120***) of generative AI-enabled chatbots are significantly associated with perceived authenticity. Therefore, H2 and H3 are supported. Furthermore, the results highlight that the perceived authenticity of generative AI-enabled chatbots is positively associated with customer engagement (β=0.670***), experience (β=0.755***) and recommendation (β=0.927***). Therefore, H4a, H4b, and H4c are supported. However, the effect of anthropomorphic characteristics (β=-0.078ns) of generative AI-enabled chatbots on perceived warmth was not significant. Hence, H1b is rejected. The R^2 values for perceived competence, perceived warmth, perceived authenticity, customer engagement, customer experience and customer recommendation are 0.47, 0.06, 0.86, 0.44, 0.56 and 0.86, respectively.

Table 4: Moderation analysis

Moderating need for social interaction	Effect	SE	LLCI	ULCI	Moderation
Anthropomorphic Characteristics → Perceived Competence	-0.007	0.028	-0.06	0.046	No
Anthropomorphic Characteristics → Perceived Warmth	0.153	0.066	0.012	0.273	Yes

Table 5: Moderation at low and high levels

Moderating need for social interaction	Level	Effect	SE	T
Anthropomorphic Characteristics → Perceived Warmth	Low	-0.381***	0.106	3.562
	Medium	-0.222*	0.091	2.425
	High	-0.064ns	0.121	0.531
Note: *** denotes p<0.001; * denotes p< 0.05; ns denotes non-significant				

5.3. Moderation analysis

Model 1 in the Process Macro was used to assess the moderation hypotheses (Hayes, 2013). The results presented in Table 4 and Table 5 illustrate the impact of anthropomorphic characteristics on perceived warmth is significantly moderated by the need for social interaction, thereby supporting H5b. However, the influence of anthropomorphic characteristics on perceived competence is not moderated by the need for social interaction, thereby not supporting H5a.

6. Discussion

This study investigated how adding human-like characteristics to generative AI-enabled chatbots (anthropomorphism) affects how users perceive their competence and warmth. Moreover, this study investigates the customer's behavioural responses (engagement, experience, recommendation) towards these generative AI-enabled chatbots. The findings of this study reveal that anthropomorphic characteristics of generative AI-enabled chatbots are positively associated with competence. The findings are in line with prior literature in other contexts (Nguyen et al., 2023; Cheng, 2022). One probable reason for the finding is rooted in the inherent human tendency to attribute human-like qualities to entities displaying human traits. When generative AI-enabled chatbots exhibit anthropomorphic characteristics—such as human-like language, tone, and behaviour—users are more likely to perceive these chatbots as intelligent and capable. This perception stems from the familiarity and comfort humans generally feel when interacting with entities that resemble other humans. The anthropomorphic design cues can thus bridge the gap between humans and machines, making the chatbot appear more competent in understanding and responding to user queries.

Furthermore, the findings show that the perceived competence and warmth of generative AI-enabled chatbots are significantly associated with perceived authenticity, which is supported by previous studies in other contexts (Nguyen et al., 2023; Cheng et al., 2022; Belanche et al., 2021). One of the probable reasons for this finding is that perceived authenticity in interactions with generative AI-enabled chatbots hinges on the chatbot's ability to convincingly replicate human-like behaviour and responses. When chatbots exhibit high levels of competence, they can accurately understand and respond to user queries, thereby creating interactions that feel genuine and reliable. This accuracy and relevance in responses enhance the user's perception that the chatbot is behaving authentically, as it demonstrates an understanding of the user's needs and contexts similar to that of a human service provider.

Further, consistent with the literature (Lee & Kim, 2024; Nguyen et al., 2023), our findings revealed that the perceived authenticity of generative AI-enabled chatbots is positively associated with customer engagement, experience and recommendation. One potential explanation for the positive association is trust and rapport building. When users perceive a chatbot as authentic, they are more likely to feel comfortable and confident in interacting with it. This comfort level leads to more frequent and deeper interactions, as users are assured that the chatbot can provide genuine and helpful responses, thus increasing overall engagement. Similarly, authentic interactions with chatbots that exhibit human-like characteristics can make the interaction more enjoyable and satisfying. When customers feel that their interactions are authentic, they are more likely to view the experience positively, as it resembles the personalized service they would receive from a human agent. Moreover, perceived authenticity can enhance the likelihood of customer recommendations. When customers have positive and authentic interactions with chatbots, they are more likely to share their experiences with others. Customers who perceive their interactions as genuine are more likely to advocate for the brand and recommend it to others.

However, contrary to the prior literature in other contexts (Cheng et al., 2022), the effect of anthropomorphic characteristics of generative AI-enabled chatbots on warmth was not significant. One of the probable reasons for this finding is that anthropomorphic characteristics alone may not be sufficient to convey warmth in interactions with generative AI-enabled chatbots. Warmth often requires not just human-like features but also genuine empathy and appropriate emotional responses. While anthropomorphic characteristics can make a chatbot appear more relatable, they do not inherently ensure that it responds with the emotional intelligence and empathy that users associate with warmth. Users might perceive these characteristics as superficial without accompanying empathetic behaviour and understanding, thus diminishing their impact on perceived warmth.

Further, in line with prior literature (Flavián et al., 2024; Belanche et al., 2021), results revealed that the need for social interaction significantly moderates the impact of anthropomorphic characteristics on warmth. However, the influence of anthropomorphic characteristics on competence was not moderated by the need for social interaction. One of the probable reasons for the finding is that individuals with a high need for social interaction place greater value on emotionally rich and engaging interactions. These users are more likely to appreciate and respond positively to the human-like traits of chatbots because these traits fulfil their desire for social connection and emotional engagement. On the other hand, the finding that the influence of anthropomorphic characteristics on competence was not moderated by the need for social interaction can be explained by the universal importance of task-oriented performance across all user groups. Competence in chatbots is generally assessed based on the ability to understand and accurately respond to queries, provide relevant information, and perform tasks efficiently. These attributes are critical for all users, regardless of their need for social interaction.

7. Implications

7.1. Theoretical implications

The outcomes of our study offer theoretical advancements within the consumer behaviour literature on new technology and innovation adoption. The theoretical implications of our study extend beyond the immediate context of generative AI-enabled chatbots, contributing to broader discussions about the nature of user interactions with technology. Our contributions stem from introducing a novel conceptual framework that generates fresh insights within a distinct context, offering new perspectives and findings. This study contributes to our understanding of human-computer interaction by shedding light on the interplay between anthropomorphism, perceived authenticity, and user experience in the context of generative AI-enabled chatbots. Further, this study reinforces social response theory by demonstrating that human-like characteristics in AI-enabled chatbots elicit social responses from users, similar to interactions with other humans. This study expands the theoretical framework of social response theory by highlighting the differential effects of anthropomorphic characteristics on competence and warmth, moderated by individual user traits (need for social interactions). It emphasizes the necessity for personalized generative AI-enabled interactions that cater to the diverse social needs of users, thereby offering a deeper understanding of how anthropomorphism and user-specific factors influence the efficacy of generative AI-enabled interactions. This theoretical advancement can guide future research and development in the field of human-computer interaction, particularly in designing generative AI systems that promote meaningful and satisfying customer experiences.

7.2. Practical implications

This study yields several practical implications for customers seeking to use generative AI-enabled chatbots. The findings of this study highlighted that anthropomorphic characteristics of generative AI-enabled chatbots significantly influence generative AI-enabled chatbots' competence. To enhance the influence of anthropomorphic characteristics, brands should invest in advanced natural language processing technologies to ensure chatbots can understand and respond to user queries in a natural, contextually appropriate manner. Further, brands should design chatbots to provide personalized responses that reflect an understanding of individual user preferences and histories.

Furthermore, the findings show that the competence and warmth of generative AI-enabled chatbots are significantly associated with perceived authenticity. To enhance the significance of the association, organizations and brands should integrate emotional intelligence into chatbots so they can recognize and respond appropriately to users' emotions. Also, generative AI platforms should implement machine learning algorithms that allow chatbots to learn from interactions and continuously improve their performance. Further, they should train chatbots to provide empathetic and supportive responses, making interactions feel more personal and engaging.

Moreover, the findings show that the perceived authenticity of generative AI-enabled chatbots is positively associated with customer engagement, experience and recommendation. To enhance the significance of the perceived authenticity of generative AI-enabled chatbots on customer engagement, experience, and recommendation, organizations and brands should implement strategies that focus on consistent branding, advanced AI capabilities, and promoting trust through transparency. Incorporating engaging features and proactive assistance can boost user engagement, while seamless integration across touchpoints and collecting user feedback will enhance the overall experience. Positive interaction outcomes should be prioritized to encourage recommendations, supported by tracking key performance indicators and conducting regular audits. Finally, the findings highlighted that the need for social interaction significantly moderates the impact of anthropomorphic characteristics on warmth. Generative AI-enabled platforms should provide users with options to customize the chatbot's behaviour and level of anthropomorphic characteristics based on their social interaction preferences. Moreover, generative AI organizations should conduct periodic surveys to understand users' preferences regarding chatbot interaction styles and adjust settings accordingly.

8. Limitations and future research

Similar to other research studies, our study also faces several limitations. Primarily constrained by resources, this paper relies solely on cross-sectional data. Future researchers could broaden their scope by incorporating data across a longitudinal period to facilitate a more comprehensive analysis of adoption intention. The study relied on self-reported measures of perceived competence, warmth, and authenticity. Future research could incorporate physiological measures or behavioural tracking to gain a more objective understanding of user responses. Moreover, the study focused on a generative AI-enabled chatbot from a very generic perspective. Future research should explore how these findings translate to different chatbot functionalities and industries (e.g., healthcare vs e-commerce vs education). This study also informs future researchers to explore the nuances of user autonomy in human-computer interactions, highlighting how individual preferences and experiences can redefine traditional models of technology adoption. Ultimately, this study lays the groundwork for exploring new avenues of research focused on the emotional and psychological dimensions of technology interaction, providing a richer understanding of how users navigate their relationships with increasingly sophisticated AI systems.

References

Bahoo, S., Cucculelli, M., and Qamar, D. (2023), "Artificial intelligence and corporate innovation: A review and research agenda", Technological Forecasting and Social Change, Vol. 188, pp. 122264.

Bai, S., Yu, D., Han, C., Yang, M., Gupta, B. B., Arya, V., ... and Zhao, J. (2024), "Warmth trumps competence? Uncovering the influence of multimodal AI anthropomorphic interaction experience on intelligent service evaluation: Insights from the high-evoked automated social presence", Technological Forecasting and Social Change, Vol. 204, pp. 123395.

Belanche, D., Casaló, L. V., Schepers, J., and Flavián, C. (2021), "Examining the effects of robots' physical appearance, warmth, and competence in frontline services: The Humanness-Value-Loyalty model", Psychology & Marketing, Vol. 38 No. 12, pp. 2357-2376.

Blut, M., Wang, C., Wünderlich, N. V., and Brock, C. (2021), "Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI", Journal of the Academy of Marketing Science, Vol. 49, pp. 632-658.

Celik, I. (2023), "Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education", Computers in Human Behavior, Vol. 138, pp. 107468.

Chakraborty, D., Kar, A. K., Patre, S., and Gupta, S. (2024), "Enhancing trust in online grocery shopping through generative AI chatbots", Journal of Business Research, Vol. 180, pp. 114737.

Cheng, L. K. (2022). The effects of smartphone assistants' anthropomorphism on consumers' psychological ownership and perceived competence of smartphone assistants. Journal of Consumer Behaviour, 21(2), 427-442.

Cheng, X., Zhang, X., Cohen, J., and Mou, J. (2022), "Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms", Information Processing & Management, Vol. 59 No. 3, pp. 102940.

Cui, Y. G., van Esch, P., and Phelan, S. (2024), "How to build a competitive advantage for your brand using generative AI", Business Horizons, Vol. ahead-of-print No. ahead-of-print.

Dey, P. K., Chowdhury, S., Abadie, A., Vann Yaroson, E., and Sarkar, S. (2023), "Artificial intelligence-driven supply chain resilience in Vietnamese manufacturing small-and medium-sized enterprises", International Journal of Production Research, pp. 1-40.

Dicuonzo, G., Donofrio, F., Fusco, A., and Shini, M. (2023), "Healthcare system: Moving forward with artificial intelligence", Technovation, Vol. 120, pp. 102510.

Dwivedi, Y. K., Balakrishnan, J., Baabdullah, A. M., and Das, R. (2023), "Do chatbots establish "humanness" in the customer purchase journey? An investigation through explanatory sequential design", Psychology & Marketing, Vol. 40 No. 11, pp. 2244-2271.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023), ""So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy", International Journal of Information Management, Vol. 71, pp. 102642.

Ferraro, C., Demsar, V., Sands, S., Restrepo, M., and Campbell, C. (2024), "The paradoxes of generative AI-enabled customer service: A guide for managers", Business Horizons, Vol. ahead-of-print No. ahead-of-print.

Flavián, C., Belk, R. W., Belanche, D., and Casaló, L. V. (2024), "Automated social presence in AI: Avoiding consumer psychological tensions to improve service value", Journal of Business Research, Vol. 175, pp. 114545.

Fu, H.-P., Chang, T.-H., Lin, S.-W., Teng, Y.-H. and Huang, Y.-Z. (2023), "Evaluation and adoption of artificial intelligence in the retail industry", International Journal of Retail & Distribution Management, Vol. 51 No. 6, pp. 773-790.

Gupta, R., Nair, K., Mishra, M., Ibrahim, B., and Bhardwaj, S. (2024), "Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda", International Journal of Information Management Data Insights, Vol. 4 No. 1, pp. 100232.

Hair, J. F., Celsi, M. W., Ortinau, D. J., and Bush, R. P. (2017), "Essentials of Marketing Research", McGraw-Hill.

Harkness, L., Robinson, K., Stein, E., and Wu, W. (2023), "How generative AI can boost consumer marketing", McKinsey & Company, Available at: https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/how-generative-ai-can-boost-consumer-marketing#/

Harris-Watson, A. M., Larson, L. E., Lauharatanahirun, N., DeChurch, L. A., and Contractor, N. S. (2023), "Social perception in Human-AI teams: Warmth and competence predict receptivity to AI teammates", Computers in Human Behavior, Vol. 145, pp. 107765.

Hayes, A.F. (2013), "Mediation, moderation, and conditional process analysis. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-based Approach", Guilford Publications, New York, Vol. 1, pp. 20.

Henseler, J., Ringle, C. M., and Sarstedt, M. (2015), "A new criterion for assessing discriminant validity in variance-based structural equation modeling", Journal of the Academy of Marketing Science, Vol. 43, pp. 115-135.

Hermann, E., and Puntoni, S. (2024), "Artificial intelligence and consumer behavior: From predictive to generative AI", Journal of Business Research, Vol. 180, pp. 114720.

Huang, J. W., and Lin, C. P. (2011), "To stick or not to stick: The social response theory in the development of continuance intention from organizational cross-level perspective", Computers in Human Behavior, Vol. 27 No. 5, pp. 1963-1973.

Huang, S. Y., and Lee, C. J. (2022), "Predicting continuance intention to fintech chatbot", Computers in Human Behavior, Vol. 129, pp. 107027.

Huy, L. V., Nguyen, H. T., Vo-Thanh, T., Thinh, N. H. T., and Thi Thu Dung, T. (2024), "Generative AI, Why, How, and Outcomes: A User Adoption Study", AIS Transactions on Human-Computer Interaction, Vol. 16 No. 1, pp. 1-27.

Jaboob, M., Hazaimeh, M., and Al-Ansi, A. M. (2024), "Integration of Generative AI Techniques and Applications in Student Behavior and Cognitive Achievement in Arab Higher Education", International Journal of Human–Computer Interaction, pp. 1-14.

Jiang, H., Cheng, Y., Yang, J., and Gao, S. (2022), "AI-powered chatbot communication with customers: Dialogic interactions, satisfaction, engagement, and customer behavior", Computers in Human Behavior, Vol. 134, pp. 107329.

Kong, H., Wang, K., Qiu, X., Cheung, C. and Bu, N. (2023), "30 years of artificial intelligence (AI) research relating to the hospitality and tourism industry", International Journal of Contemporary Hospitality Management, Vol. 35 No. 6, pp. 2157-2177.

Kshetri, N., Dwivedi, Y. K., Davenport, T. H., and Panteli, N. (2023), "Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda", International Journal of Information Management, pp. 102716.

Kushwaha, A. K., Kumar, P., and Kar, A. K. (2021), "What impacts customer experience for B2B enterprises on using AI-enabled chatbots? Insights from Big data analytics", Industrial Marketing Management, Vol. 98, pp. 207-221.

Lee, G., and Kim, H. Y. (2024), "Human vs. AI: The battle for authenticity in fashion design and consumer response", Journal of Retailing and Consumer Services, Vol. 77, pp. 103690.

Li, Z., and Li, C. (2014), "Twitter as a social actor: How consumers evaluate brands differently on Twitter based on relationship norms", Computers in Human Behavior, Vol. 39, pp. 187-196.

McKinsey & Company (2024), "How generative AI is disrupting distribution", Available at: https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/distribution-blog/how-generative-ai-is-disrupting-distribution

McKinsey & Company (2024), "The economic potential of generative AI: The next productivity frontier", Available at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction

Moon, Y. (2000), "Intimate exchanges: Using computers to elicit self-disclosure from consumers", Journal of Consumer Research, Vol. 26 No. 4, pp. 323-339.

Nazir, S., Khadim, S., Asadullah, M. A., and Syed, N. (2023), "Exploring the influence of artificial intelligence technology on consumer repurchase intention: The mediation and moderation approach", Technology in Society, Vol. 72, pp. 102190.

Nguyen, M., Casper Ferm, L. E., Quach, S., Pontes, N., and Thaichon, P. (2023), "Chatbots in frontline services and customer experience: An anthropomorphism perspective", Psychology & Marketing, Vol. 40 No. 11, pp. 2201-2225.

Ooi, K. B., Tan, G. W. H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., ... and Wong, L. W. (2023), "The potential of generative artificial intelligence across disciplines: Perspectives and future directions", Journal of Computer Information Systems, pp. 1-32.

Pelau, C., Dabija, D. C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Computers in Human Behavior, 122, 106855.

Peng, C., van Doorn, J., Eggers, F., and Wieringa, J. E. (2022), "The effect of required warmth on consumer acceptance of artificial intelligence in service: The moderating role of AI-human collaboration", International Journal of Information Management, Vol. 66, pp. 102533.

Pérez-Vega, R., Taheri, B., Farrington, T., and O'Gorman, K. (2018), "On being attractive, social and visually appealing in social media: The effects of anthropomorphic tourism brands on Facebook fan pages", Tourism Management, Vol. 66, pp. 339-347.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., and Podsakoff, N. P. (2003), "Common method biases in behavioral research: a critical review of the literature and recommended remedies", Journal of Applied Psychology, Vol. 88 No. 5, pp. 879.

Premathilake, G. W., and Li, H. (2024), "Users' responses to humanoid social robots: A social response view", Telematics and Informatics, pp. 102146.

Rahman, M., Ming, T.H., Baigh, T.A. and Sarker, M. (2023), "Adoption of artificial intelligence in banking services: an empirical analysis", International Journal of Emerging Markets, Vol. 18 No. 10, pp. 4270-4300.

Renier, L. A., Mast, M. S., and Bekbergenova, A. (2021), "To err is human, not algorithmic–Robust reactions to erring algorithms", Computers in Human Behavior, Vol. 124, pp. 106879.

Rese, A., Ganster, L., and Baier, D. (2020), "Chatbots in retailers' customer communication: How to measure their acceptance?", Journal of Retailing and Consumer Services, Vol. 56, pp. 102176.

Roy, R., & Naidoo, V. (2021). Enhancing chatbot effectiveness: The role of anthropomorphic conversational styles and time orientation. Journal of Business Research, 126, 23-34.

Sætra, H. S. (2023), "Generative AI: Here to stay, but for good?", Technology in Society, Vol. 75, pp. 102372.

Sohn, K., Sung, C.E., Koo, G. and Kwon, O. (2021), "Artificial intelligence in the fashion industry: consumer responses to generative adversarial network (GAN) technology", International Journal of Retail & Distribution Management, Vol. 49 No. 1, pp. 61-80.

Tuguinay, J. A., Prentice, C., and Moyle, B. (2022), "The influence of customer experience with automated games and social interaction on customer engagement and loyalty in casinos", Journal of Retailing and Consumer Services, Vol. 64, pp. 102830.

Wang, Y.-C. and Uysal, M. (2024), "Artificial intelligence-assisted mindfulness in tourism, hospitality, and events", International Journal of Contemporary Hospitality Management, Vol. 36 No. 4, pp. 1262-1278.

Yang, S., Xie, W., Chen, Y., Li, Y., and Jiang, H. (2024), "Warmth or competence? Understanding voice shopping intentions from Human-AI interaction perspective", Electronic Commerce Research, pp. 1-30.

Appendix 1: Variable items with literature sources

Variables and items
Anthropomorphic Characteristics (Pelau et al., 2021)
AC1: I believe that generative AI has its own intentions
AC2: I believe that generative AI has its own decision-making power
AC3: I believe that generative AI has consciousness
AC4: I believe that generative AI is creative and has its own imagination
AC5: I believe that generative AI has a human touch
AC6: I believe that generative AI is trustworthy
Competence (Dwivedi et al., 2023)
CO1: I perceive that generative AI is intelligent
CO2: I perceive that generative AI is skillfull
CO3: I perceive that generative AI is capable
CO4: I perceive that generative AI is efficient
Warmth (Dwivedi et al., 2023)
WA1: I perceive that generative AI cares about me while interacting
WA2: I perceive that generative AI cares is kind to me
WA3: I perceive that generative AI is friendly to me during the conversation
WA4: I perceive that, the conversation with the generative AI is warm
Perceived Authenticity (Nguyen et al., 2023)
PA1: I would feel that the generative AI is natural
PA2: I would feel that the generative AI is organic
PA3: I would feel that the generative AI is real
Customer Engagement (Nguyen et al., 2023)
ENG1: I feel motivated to learn more about generative AI
ENG2: I encourage friends and relatives to use generative AI that offers this online chat support
ENG3: I consider generative AI that offers this online chat support to be my first choice when buying products
Customer Experience (Dwivedi et al., 2023)
CEX1: The interaction through generative AI is more appealing
CEX2: The generative AI query results are returned promptly
CEX3: The interaction with generative AI is more personalized
CEX4: The generative AI results are always accurate
Customer Recommendation (Dwivedi et al., 2023)
CUR1: I will recommend others to use generative AI
CUR2: I will say positive things to others about using generative AI
CUR3: I will encourage friends and relatives to use generative AI
Need for Social Interaction (Flavián et al., 2024)
INT1: Human contact in services provision makes the process enjoyable for the consumer
INT2: I like interacting with the person who provides the service
INT3: Personal attention by the service employee is very important to me
INT4: It bothers me to use a generative AI when I could talk to a person instead

A Systematic Review of IS Literature in Dementia Care: An Application of the NASSS Framework

Zhengyang Feng
School of Computer Science
University of Technology Sydney
Ultimo, NSW, Australia
Zhengyang.feng@student.uts.edu.au

Jayan Chirayath Kurian
School of Computer Science
University of Technology Sydney
Ultimo, NSW, Australia
jayanchirayathkurian@uts.edu.au

Mukesh Prasad
School of Computer Science
University of Technology Sydney
Ultimo, NSW, Australia
mukesh.prasad@uts.edu.au

Nimish Biloria
School of Architecture
University of Technology Sydney
Ultimo, NSW, Australia
nimish.biloria@uts.edu.au

Priya Saravanakumar
School of Nursing and Midwifery
University of Technology Sydney
Ultimo, NSW, Australia
priya.saravanakumar@uts.edu.au

Research In progress

Abstract

The application of technologies in healthcare offers substantial benefits for stakeholders. Due to the recent advances in emerging technologies, these benefits are further improved for individuals experiencing chronic conditions, especially people living with dementia. Healthcare professionals, especially caregivers, could be assisted by such technologies in the monitoring and recording of health conditions of patients. However, the trend and type of technologies applied in dementia care and a comprehensive analysis of such technologies in dementia care are limited. Therefore, this study uses the PRISMA approach to examine previous studies, identify gaps, and outline research directions. In addition, the NASSS (non-adoption, abandonment, scale-up, spread and sustainability) framework was used as a tool for analysing healthcare technologies in dementia care. From the analysis, we identified five themes to support future research in dementia care. The key challenges associated with technologies in dementia care are ensuring interoperability, understanding economic and ethical implications, and aligning with patient and caregiver needs. In future studies, expanding the literature review could provide valuable insights into the significance of all the seven domains of the NASSS framework.

Keywords: Technology, dementia care, caregivers, NASSS framework

1. Introduction

In recent years, Australia has been grappling with an alarming surge in dementia cases, which is the second leading cause of death nationwide and will become the primary cause soon (Australian Institute of Health and Welfare, 2022). The scale of the issue is overwhelming, with over 421,000 Australians currently living with dementia, whereas the number is projected to be more than double by 2054 without significant medical breakthroughs (Dementia Australia, 2023). Distressingly, this trend has also extended to the younger demographics, with nearly 29,000 individuals experiencing early onset dementia, the figure is expected to climb substantially over the coming decades (Dementia Australia, 2024). The burden extends beyond those directly affected, with an estimated 1.6 million Australians involved in dementia care, underscoring the widespread societal impact of the condition (Dementia Australia, 2024). Furthermore, a substantial number of individuals living with dementia are residing within communities, presenting unique challenges for support and care services. Within aged care facilities, cognitive impairment is prevalent, affecting over two-thirds of residents, highlighting the critical need for emerging technologies to address the challenges of dementia care and support in Australia (Dementia Australia, 2024).

In addressing the complex challenges posed by dementia in Australia, Information Technologies play a crucial role in providing innovative solutions to support people living with dementia, their caregivers, and healthcare professionals (Bhargava & Baths, 2022). The rising prevalence of dementia has sparked a technological revolution in dementia care, with a focus on both patient-centric and caretaker-centric solutions. The patient-centric solution contains various technologies, such as wearable devices, virtual reality (VR), augmented reality (AR), interactive games and robots. Additionally, embedded sensors and health monitoring systems could be the major examples of caretaker-centric solutions (Bhargava & Baths, 2022).

Integrating Information Technologies in dementia care offers numerous benefits such as enhancing patient independence, safety, social engagement, mood, and overall quality of life (Bhargava & Baths, 2022). These technologies facilitate activities of daily living, encouraging healthy aging and reducing the burden on caregivers (Lee-Cheong et al., 2022). Health monitoring capabilities enable timely intervention by professional and informal caregivers, potentially improving health outcomes and reducing healthcare costs. There are several challenges that have been discussed in the context of ethical practices and legal frameworks that safeguard the rights of people living with dementia (Lee-Cheong et al., 2022; Vollmer Dahlke & Ory, 2020). The major challenges are ensuring patient privacy, confidentiality, and security, which might decrease the confidence and trust of the public in these innovative technologies (Bhargava & Baths, 2022; Frisardi et al., 2022; Lee-Cheong et al., 2022). For instance, individuals and caregivers must be informed about the data collection protocols and should have the option to restrict the collection of sensitive information (Lee-Cheong et al., 2022).

Wearables and smart devices, offer an innovative approach to tracking and managing lifestyle habits, including physical activity, nutrition, sleep, and stress levels. Thorpe et al. (2019) suggest that it would be promising and feasible to adopt smartphones and smartwatches in the cognitive rehabilitation for people with early-stage dementia. The participants in the study were motivated to increase their activity by tracking their step count, along with substantial improvements in mobility and mood, reducing patient anxiety and caregiver burden. These devices provide actionable insights, allowing patients and caretakers to make informed decisions about the type of healthcare interventions and lifestyle adjustments.

Technologies like VR (Virtual Reality) and AR (Augmented Reality) aid in cognitive rehabilitation, while interactive games and robots enhance independence and daily living activities for people living with dementia. Matsangidou et al. (2023) suggest that VR has significant potential as a therapeutic tool in dementia care, especially in reducing behavioral and psychological symptoms of people living with dementia. They have created a VR system that effectively meets the needs of people with mild to severe dementia in long-term care facilities. Their contribution was in the design, implementation, and utilization of VR technology in dementia care. Although wearable AR headsets are still considered emerging and less user-friendly for aged people, Dickinson et al. (2023) found that AR is more tolerable than VR, with fewer side effects, whereas its reduced immersion is considered beneficial for older people. However, due to concerns about the ease of use, privacy, and data collection practices, AR has still not been widely utilized and implemented in dementia care.

Safety and security are paramount for people living with severe dementia or those living alone. Technologies, such as embedded sensors, might help monitor home activities and alert caregivers on emergencies. Anwar et al. (2023) have created a novel, low-complexity wearable sensing system using RF (radio frequency) sensors for the early detection of conditions associated with vascular dementia. Health monitoring systems utilise sensors to track physiological data, offer real-time alerts and provide effective management plans. Enshaeifar et al. (2020) have developed a digital platform for remote healthcare monitoring and support in dementia care. The key feature of this platform is a clinical interface, which could be accessed by a monitoring team, which allows the care team to review in-home activities and physiological data securely, while maintaining the privacy of people living with dementia.

One of the limitations of the above studies is the lack of theoretical approaches used in analysis. Hence, the NASSS framework, developed by Greenhalgh et al (2017), has been chosen in this review to analyse technologies used by stakeholders in dementia care. This framework is a critical tool in health informatics for analysing and predicting the success or failure of healthcare technologies in real-world settings. NASSS stands for Non-adoption, Abandonment, Scale-up (e.g. local demonstration), Spread (e.g. application used in a new setting), and Sustainability (e.g. long-term use of an application), and it is structured around 7 domains. These domains are Condition and Illness, Technology, Value Proposition, Adopter System, Organization(s), Wider System, and Embedding and Adaptation Over Time. The "Condition and Illness" domain assesses the complexity of the medical condition and the technology's ability to meet healthcare requirements. The "Technology" domain examines characteristics, usability, and knowledge into existing systems, along with its adaptability to different contexts. The "Value Proposition" considers the perceived benefits among stakeholders, such as technology value and demand-side value to determine whether they are compelling enough for adoption. The "Adopter System" focuses on the preparation and role of stakeholders and potential resistance to the technology. "Organization(s)" evaluates the implementing organization's readiness, leadership support, financial health, and openness to innovation. The "Wider System" focuses on external influences like policies and regulations that affect the technology's application and implementation. Finally, "Embedding and Adaptation Over Time" assesses how the technology evolves, and adapts to changing healthcare needs, requiring ongoing updates, and training to remain effective. By providing a comprehensive view of these factors, the NASSS Framework offers valuable insights into why certain healthcare technologies succeed or fail, guiding the design and implementation of effective health interventions.

Despite the growing adoption of technologies in dementia care, there remains a significant gap in the systematic evaluation of these technologies from a holistic, multi-stakeholder perspective. Existing literature tends to focus on individual technological solutions without addressing usability and effectiveness of these technologies. In addition, they often lack theoretical grounding, limiting the generalizability of findings. The NASSS framework, a widely recognized model for evaluating healthcare technologies, has been underutilized in dementia care research, particularly in assessing the multi-dimensional factors that influence technology adoption. This study aims to bridge these gaps by applying a rigorous systematic review using PRISMA and evaluating dementia care technologies through the lens of the NASSS framework to provide a structured understanding of their adoption, challenges, and future research directions.

Specifically, the objectives of this study are:

- Identify and categorize different technologies used in dementia care and their primary functions.
- Evaluate the challenges associated with these technologies in terms of usability, integration, and safety considerations.
- Analyse stakeholder perspectives, including caregivers, healthcare professionals, and people living with dementia, regarding technology adoption.
- Identify research gaps and propose future research directions to enhance the development and implementation of dementia care technologies.

Based on the above discussion, it is evident that there is an imminent need for research and development in the field of technologies for dementia care. Such initiatives will address the existing challenges and fully realize the potential of technologies in improving the lives of those affected by dementia. Therefore, this systematic review will shed insights into the intersection of technology and dementia care using the lens of NASSS framework. The PRISMA approach used to select relevant articles for this study is outlined in the next section.

2. Methods

2.1. Searching String

The literature search was conducted in the first six months of 2024. The following keywords and logical operations were used in the search: (dementia) AND ((care) OR ("aged care") OR ("elderly care")) AND ((technolog*) OR (application*) OR (system*) OR ("assistive technolog*") OR ("digital tool*") OR ("smart system*") OR ("healthcare technolog*")). The search was restricted to publications from 2020 to 2024 and the PRISMA method was used to structure the search. The decision to focus on studies published between 2020 and 2024 is based on the rapid advancements in technology within dementia care over the past five years. The field has seen significant developments in digital health solutions, and artificial intelligence, which were not as prevalent in earlier literature. Additionally, the COVID-19 pandemic accelerated the utilization of remote monitoring and assistive technologies.

2.2. Study Selection

This research has an Information Systems focus and hence initial literature was limited to information systems databases - AIS eLibrary, IEEE, and Scopus. Firstly, 955 articles were retrieved which was restricted to journal and conference papers. 193 items were excluded for the following reasons, including non-academic sources (e.g., webpages, and opinion pieces) or studies that did not focus on technologies. The title and abstract screening have further excluded 677 articles since they lacked direct relevance to dementia or assistive technologies or focused on unrelated fields such as general aged care. Finally, a full-text examination eliminated another 68 papers, leaving 17 articles which were focused on the use of technology in dementia care.

However, these 17 articles were not enough for transverse and longitudinal analysis. Therefore, the literature search was extended into the medical database – PubMed. A 4-step PRISMA process shown in figure 1 was adopted to gather relevant articles. The initial phase of the review involved retrieving 7,601 articles from PubMed, focusing primarily on journal articles and conference papers with a particular focus on technology in dementia care. This targeted selection led to the exclusion of 4,026 papers that included webpages, book chapters, and reviews.

Subsequent screening involved a title examination of the remaining 3,575 articles which then ascertained their direct relevance to the application of information technology or digital technologies in dementia care. The articles which merely mentioned dementia or technologies without a direct link or discussed unrelated topics such as healthy ageing was discarded, which resulted in removing another 3,469 articles. The final phase involved an in-depth review of the abstracts and full texts of the 106 articles that met the initial criteria. This rigorous criterion which focused on the unique and direct application of technologies in dementia care resulted in the further exclusion of 25 articles. In the final stage, 81 articles were selected for analysis. In total, 98 articles were selected from the four (i.e. AIS eLibrary, IEEE, Scopus and PubMed) databases. The PRISMA steps followed to select articles from the PubMed database is outlined below.

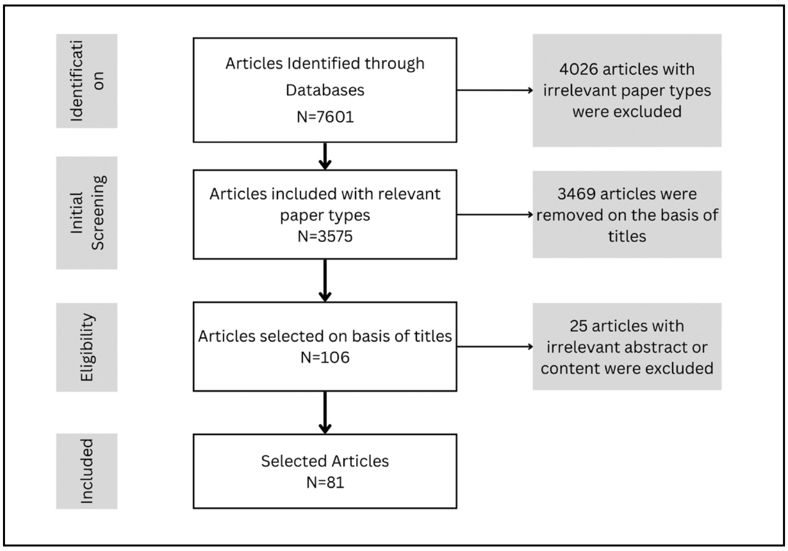

Identificati on
Articles Identified through Databases N=7601
4026 articles with irrelevant paper types were excluded
Initial Screening
Articles included with relevant paper types N=3575
3469 articles were removed on the basis of titles
Eligibility
Articles selected on basis of titles N=106
25 articles with irrelevant abstract or content were excluded
Included
Selected Articles N=81

Figure 1 PRISMA steps

2.3. Thematic Analysis

To ensure a systematic thematic analysis on the adoption of various technologies on dementia care, Braun & Clarke's (2006) thematic analysis approach was followed. First, an in-depth review of the 17 selected papers from the information systems databases (AIS eLibrary, IEEE, and Scopus) was conducted to extract key findings related to technology adoption in dementia care. Next, initial coding was performed, identifying key characteristics, technological requirements, and adoption factors relevant to the NASSS framework. These codes were then clustered into broader themes, reflecting emerging challenges, value propositions, and stakeholder perspectives. The identified themes were iteratively refined to ensure internal coherence and alignment with the research objectives. Finally, the themes were synthesized into a cohesive narrative, highlighting key research gaps and future directions. The themes emerged from the literature, ensuring a data-driven approach, with the NASSS framework guiding their alignment with broader adoption and sustainability challenges in dementia care.

3. Findings

3.1. Distribution over time

As illustrated in Figure 2, studies related to technologies used in dementia care have uncovered specific trends, which reflect the evolving nature of research in this field. There was a significant upward trend in publication numbers from 2020 to 2022, which reached its peak in 2022 with 29 publications. However, the number of publications declined in 2023. It must be noted that only papers published in the first half of 2024 were considered in this review.

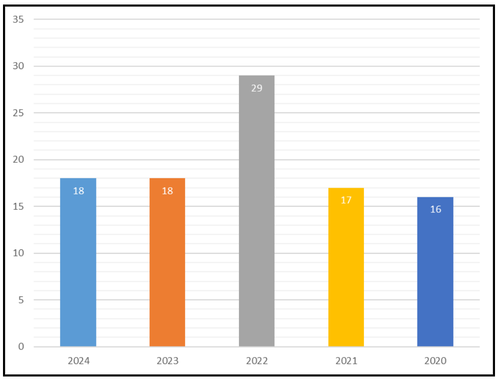


Figure 2 Number of papers across years

3.2. Distribution over technologies

In terms of technologies, there are 11 different technology groups defined by the Aged Care Research & Industry Innovation Australia (ARIIA, 2024). Research on various technologies used in dementia care highlights a diverse landscape as shown in Figure 3. Sensors and Monitoring Technology topped the list with 21 publications which represent 22% of the total number of publications selected in this study. Robots, with 13 publications (14%), reflected substantial interest with respect to the use of advanced technologies. There were 12 publications (12%) captured with a focus on Augmented or Virtual Reality. Assistive Technologies and Wearable Technologies, each had 10 publications (10%), and their focus was on people living with dementia. Studies on Telehealth was found in 9 publications (9%). Both Artificial Intelligence, Smartphones and Mobile Apps, with 7 publications (7%) each, was relevant with respect to health management. On the other hand, there were also publications focusing on Smart Homes, Care Management Systems, and Social Engagement Technologies, which were under 10% in total. This diverse spread of technologies illustrates a multi-faceted approach to dementia care, emphasizing the need to integrate innovative technologies to enhance patient outcomes and to support caregivers.

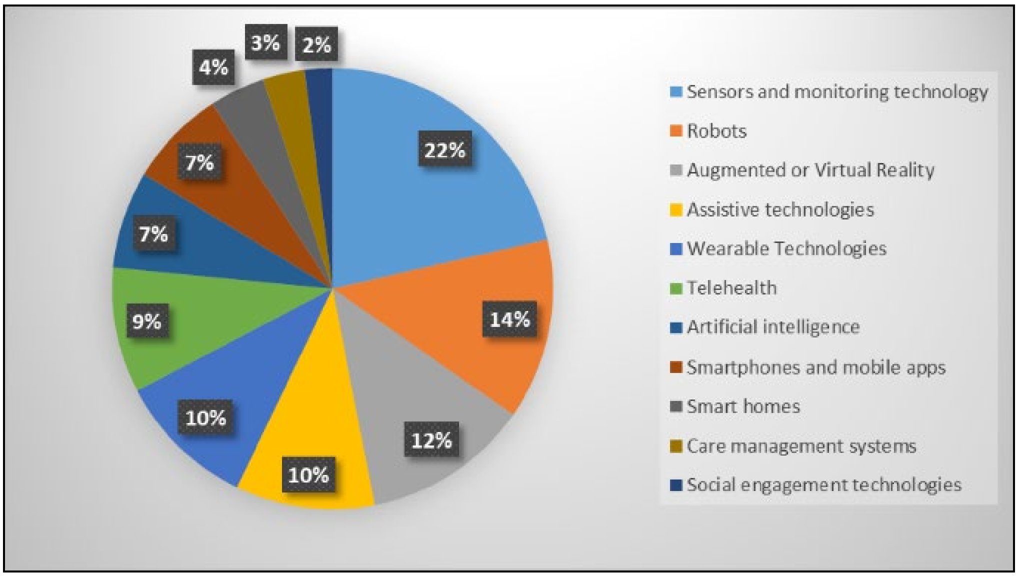


Figure 3 Number of papers across technologies

3.3. Research trend over time for various technologies

The distribution of technologies over time in Figure 4 exhibits patterns of various dementia care technologies over time, which reflects the evolving priorities and advancements within this field. It is evident that research on dementia care technologies experienced an increase in publications from 2020 to 2022. The number of publications was the highest in 2022, which could indicate a period of intense innovation and development. Specifically, Sensors and Monitoring Technology reached the highest number of publications in 2022 before significantly declining, suggesting a technological readiness level from design to implementation. Research in the Robotic field similarly maintained a consistent interest which could indicate ongoing innovations after the COVID-19 pandemic. Augmented or Virtual Reality has shown a relatively consistent and a gradual upward trend, reflecting its importance for cognitive and therapeutic purposes. There is a cyclical pattern examined in assistive technologies over time, and the fluctuation could possibly be affected by external factor such as cost, and government investment, mentioned in WHO report (World Health Organization, 2022). Overall, these trends suggest that while certain technologies have reached a maturation point, others are emerging in dementia care research, driven by ongoing technological advancements and changing healthcare needs.

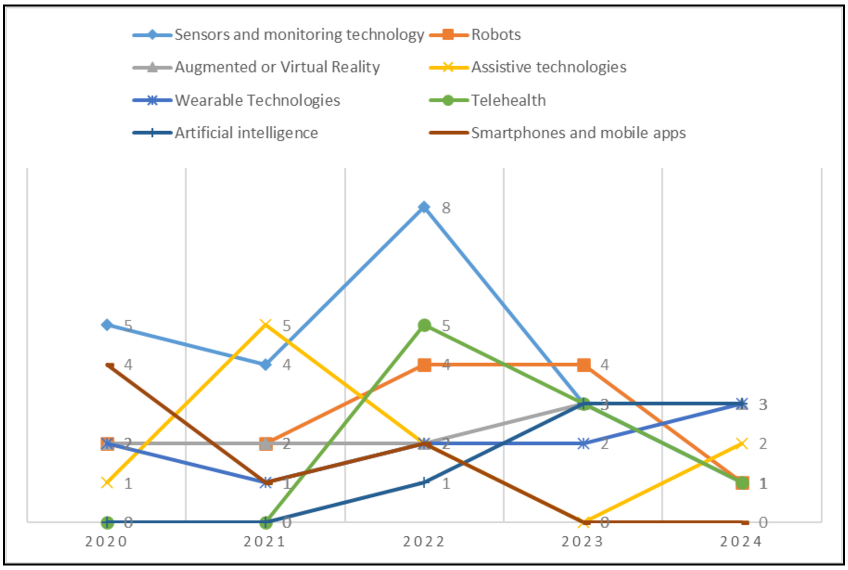


Figure 4 Number of Papers across technologies over time

3.4. Distribution over research methods

In terms of distribution over research methods, an article has been counted multiple times if it has adopted multiple methods. In that case, articles may appear in multiple categories. Among the Information Systems studies, 17 articles utilized a mixed methods approach, showcasing a preference for integrating both qualitative and quantitative methods. Among the articles, 14 were qualitative, emphasizing the importance of understanding the experiences and perspectives of those affected by dementia. 12 articles employed randomized controlled trials, and 11 adopted quantitative methods, reflecting an emphasis on rigorous, empirical evaluations of technologies. Additionally, 11 articles focused on design and implementation research, highlighting the development and enhancement of practical solutions. Eight articles used case study methodologies to investigate specific instances, while 7 articles included experimental research, testing hypotheses under controlled conditions. Pilot studies appeared in 5 articles, indicating the exploration of new interventions on a smaller scale. Prospective observational studies and cross-sectional surveys were the least common, each featured in 2 articles, suggesting a reduced dependence on observational data. Overall, the distribution of research methods underscores a need for multifaceted approach in advancing dementia care technologies, combining empirical rigor with experiential insights.

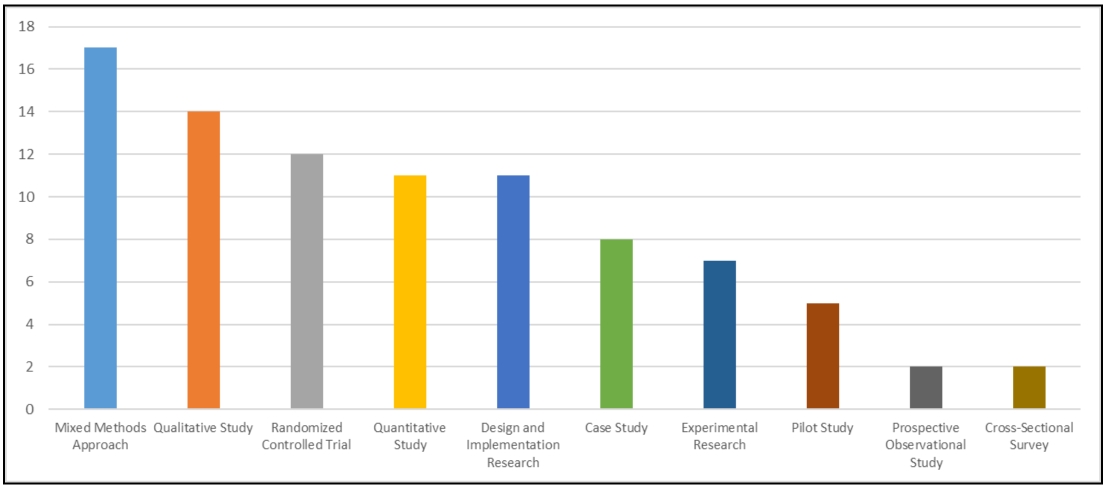


Figure 5 Number of papers across research methods

3.5. Theoretical evaluation based on the NASSS framework.

The NASSS framework was used as a theoretical assessment tool for evaluating the 17 Information Systems studies on dementia care technologies which include sensors and monitoring technology, virtual reality, augmented reality, smart homes, and wearable devices. There are 7 domains in the NASSS framework. Domain 1 (Condition and Illness – 1A,1B) of the NASSS framework examines the nature, characteristics, and impacts of dementia, assessing the extent of details provided about the illness. Domain 2 (Technology – 2A, 2B,2C,2D) evaluates the key features, associated knowledge, required support, and supply model of the technology, from basic to highly integrated levels. Domain 3 (Value Proposition – 3A,3B) considers the business case and value from both developers and user perspectives, focusing on desirability, efficacy, safety, and cost-effectiveness. Domain 4 (Adopters – 4A,4B,4C) analyses the perspectives of various stakeholders, including patients, caregivers, and healthcare providers, assessing their readiness and willingness to adopt the technology. Domain 5 (Organization – 5A,5B,5C,5D,5E) examines the suitability of a technology within existing organizational structures, considering workflow integration and resource allocation. Domain 6 (Wider System – 6A) examines the broader system context, including regulatory, policy, and market factors that influence the adoption and diffusion of the technology. Lastly, the Domain 7 (Embedding and Adaptation Over Time – 7A,7B) assesses the sustainability and adaptability of the technology over time. In the absence of extensive theory-based reviews in the adoption of technologies in dementia care, this comprehensive evaluation aids in understanding the adoption and spread of dementia care technologies in Information

Systems research. The domain questions (i.e. 1A to 7B) used in this study are adopted from the work of Greenhalgh et al., 2017). The analysis using the NASSS framework is outlined below in Table 1.

Articles	**Research Samples**	**Technology**	**NASSS Assessment High match (H), Basic match (B)**
Ambika et al. (2023)	N/A	Wearable Technologies	H: 1A,3A,3B; B: 1B, 2A, 2B, 2C, 4A
Appel et al. (2022)	Veterans with cognitive impairment	Virtual Reality	H: 2C, 3B; B: 1A, 1B, 2A, 2B, 3A, 4B, 4C, 7A
Fixl et al. (2021)	Healthy people from Poland and Austria	Assistive technologies	H: 1A, 2A, 2B, 3B; B: 1B, 2C, 3A, 4A
Gaikwad et al. (2023)	Aged People Suffering from Dementia	Sensors and monitoring technology	H: 3A, 3B; B: 1A, 1B, 2A, 2B, 2C, 6
Garcia-Constantino et al. (2021)	N/A	Smart homes	H: 2B; B: 1A, 1B, 2C, 3B, 4A,4B
García-Requejo et al. (2022)	N/A	Sensors and monitoring technology	H: 1A, 2A, 3A, 3B; B: 2B, 4A
Hamilton et al. (2021)	N/A	Augmented Reality	H: 2C, 3A; B: 1B, 2A, 2B, 3B, 4A, 4B
John et al. (2023)	Healthy users	Virtual Reality	H: 1A, 2A, 2C, 3A, 3B; B: 2B, 4A, 4B
Kocher et al. (2022)	Healthy users	Virtual Reality	H: 3B; B: 1A, 2A, 2B, 3A
Matsangidou et al. (2022)	People living with dementia, medical and paramedical personnel.	Sensors and monitoring technology	H: 1A, 2B, 3B; B: 1B, 2A, 2C, 3A, 4A, 4B
Megalingam et al. (2022)	People living with dementia and mentally unstable dementia patients	Sensors and monitoring technology	H: 1A, 2A, 2C, 3A, 3B; B: 1B, 2B, 4A, 4B
Moreno and Martínez (2023)	People with Acquired Brain Injury (ABI)	Robots	H: 2B, 3B; B: 1B, 2A, 3A, 4A, 4B
Pandhi andre Tiwari (2022)	N/A	Smartphones and mobile apps	H: 1A, 2A, 3A; B: 1B, 2B, 2C,
Perimal-Lewis et al. (2020)	Caregivers, Research Staff, Business Partners, older adults.	Smartphones and mobile apps	H: 1A, 3B; B: 1B, 2B, 3A,
Pratama et al. (2020)	People living with dementia or Alzheimer	Sensors and monitoring	H: 2A, 3A, 3B; B: 1A, 1B, 2B,

		technology	4A, 6
Schweiger and Wolff (2023)	N/A	Robots	H: 2B; B: 1A, 2A, 3A, 3B
Tarbert and Singhatat (2023)	Elderly people	Wearable Technologies	H: 2A, 3B; B: 1B, 2B, 3A

Table 1 Summary of reviewed articles

The theoretical assessment of various dementia care technologies, based on the NASSS framework, reveals significant insights into their alignment with the key domains. The Domain 1 (Condition and Illness) examines the nature of dementia, focusing on its characteristics and major impacts rather than specific comorbidities. Several studies, such as those by Appel et al. (2022) and Gaikwad et al. (2023), address these aspects by involving participants with cognitive impairments and dementia, providing detailed information on the chronic disease. In Domain 2 (Technology), the assessment encompasses key features, associated knowledge, support required for effective use, and the supply model. Technologies like Sensors and Monitoring (Gaikwad et al., 2023; Megalingam et al., 2022) and Virtual Reality (John et al., 2023; Kocher et al., 2022) were evaluated highly for their innovative integration and specific knowledge, often achieving level 2 in features and knowledge (2A, 2B). These technologies also demonstrate comprehensive support structures (2C). The Domain 3 (Value Proposition) evaluates both the business case and user demand. Technologies such as Wearable Technologies (Ambika et al., 2023; Tarbert and Singhatat, 2023) and Robots (Moreno and Martínez, 2023; Schweiger and Wolff, 2023) exhibit strong business models and high desirability, efficacy, and safety, often reaching level 2 in both developer's business case (3A) and demand-side value (3B). This robust alignment across NASSS domains underscores the potential effectiveness and market viability of these technologies in addressing the complex needs of dementia care.

Based on the analysis using the NASSS framework, researchers often focused on the Questions - 2A and 3B, as illustrated in Table 1. This indicates a primary interest in the key characteristics and demand-side value of technologies for dementia care. Key characteristics serve as the foundation of innovative technology, while understanding the demand-side value highlights the potential for technology utilisation, particularly in dementia care. By emphasizing these aspects, researchers can better assess and enhance the adoption and effectiveness of these healthcare technologies in dementia care.

3.6. Themes from Analysis

The components of the NASSS framework were used to analyse the overarching themes evident across the literature. The following 5 themes have been identified.

Theme 1: **Technology Optimization**. The analysis of technologies in dementia care reveals a significant focus on enhancing care outcomes through digital tools. A key focus is on the integration and prioritisation of monitoring and interaction technologies, with 22% of studies dedicated to sensors and monitoring devices. This underscores their role in continuous care and real-time health monitoring for dementia patients. This aligns with the study's research objectives by addressing the primary functions of these technologies (Objective 1) and how these technologies can associate with safety considerations (Objective 2).

Robots have also received considerable attention for their potential therapeutic benefits through interactive engagement. Additionally, immersive technologies such as augmented and virtual reality were explored for cognitive engagement and symptom management. Supportive technologies, including assistive and wearable devices, emphasise empowering dementia patients to maintain independence and safety in daily life activities. The growing importance of telehealth underscores the relevance of remote care delivery, particularly during situations like the COVID-19 pandemic.

However, research gaps have been found, including a need for comparative analysis to determine which technologies offer the most significant benefits. There is a lack of emphasis on the long-term adoption and sustainability of these technologies, including maintenance costs and integration into existing healthcare management. Only limited studies have integrated multiple technologies into holistic care models which must be addressed in future research. Further research which examines the impact of these technologies on caregivers, addressing their burden, stress, and daily activities at the workplace is also necessary to mitigate caregiver shortage in the aged care sector.

Theme 2: **Technological Advancements in Elderly Care**. The analysis of dementia care technologies reveals an evolving landscape, with a surge in published works reaching its peak in 2022. This could be driven by technological advancements in recent years and the wide-spread implementation of healthcare technologies during the recent pandemic. However, a decline in 2023 suggests a consolidation phase that could be attributed to the implementation of technologies. The key trends include the maturation of sensors and monitoring technologies, continued interest in robotics due to social distancing and workforce shortages, and a growing focus on augmented and virtual reality technologies for cognitive and therapeutic applications. Despite these developments, there are still significant research gaps. There is a need to study the integration and interoperability of new technologies within the existing healthcare infrastructures and their interaction with other care processes in health care management. The economic and ethical implications of adopting such technologies, including cost-benefit analyses and considerations around continuous surveillance in robotics require more comprehensive research. Additionally, there is a lack of emphasis on patient and caregiver perspectives, with limited studies focusing on stakeholder feedback and experiences especially in dementia care. Furthermore, understanding user experiences is crucial for ensuring the usability and long-term acceptance of technologies in dementia care.

Theme 3: **Multidimensional Research Approaches**. The study of dementia care technologies showcases a comprehensive research approach, with a strong preference for mixed methods, reflecting a trend toward integrating qualitative and quantitative methods. Qualitative studies capture the lived experiences of those affected by dementia, essential for designing user-centered technologies. Empirical evaluations through randomized controlled trials and quantitative studies ensure the reliability of data and technology efficacy. However, significant research gaps remain. There are limited studies on the scalability of successful interventions, which is essential for transitioning technologies from pilot studies to full-scale deployment. Furthermore, the underutilisation of prospective observational studies and cross-sectional surveys limits valuable insights into the ongoing needs and impacts of technologies on broader population, which are essential for understanding technology adoption and usage in diverse cultural settings.

Theme 4: **Technology Characteristics**. The analysis of papers using the NASSS framework indicates a major focus on Domain 2A, which examines the key characteristics of technologies. This domain assesses the fundamental properties and functionalities of technologies, such as usability, reliability, and performance, which are crucial for their success in real-life care environments. Studies frequently emphasise these core attributes to ensure that technologies meet the specific needs of dementia care without adding complexity to users' lives. For instance, Fixl et al. (2021) highlighted the importance of seamless integration and robust data management in assistive technologies, addressing the critical components of Domain

2A. Similarly, John et al. (2023) explored virtual reality technologies to assess cognitive processing and dexterity, showcasing their potential therapeutic benefits for elderly people. However, there is a need for studies on the long-term usability and adaptability of technologies to understand how technologies perform over time and adapt to users' needs. Additionally, there exist a lack of cross-technology comparison studies to determine the most beneficial technology that can address the specific needs of patients and caregivers in dementia care. Addressing these gaps is essential for designing task-technology fit guidelines, ensuring ongoing and sustained effectiveness for stakeholders in dementia care.

Theme 5: **User Adoption impacted by Demand-side Value of Technologies**. The focus on Domain 3B within the NASSS framework underscores the importance of assessing the demand-side value of technologies in dementia care. Domain 3B examines how end-users perceive and value the technology in terms of desirability, efficacy, safety, and cost-effectiveness, which are critical factors for the adoption and sustained use of technology in real-life environments. Desirability involves the technology's appeal to users, particularly in terms of user-friendly design, accessibility, and the ability to meet the specific needs of targeted population, such as people living with dementia or older adults. Efficacy assesses the technology's effectiveness in achieving its intended purposes, such as improving communication, managing daily tasks, or ensuring patient safety. The safety aspect pertains to the technology's ability to operate without posing risks to users, crucial in settings involving vulnerable population. Cost-effectiveness considers both initial and ongoing operational costs, essential for scalability and accessibility. In relation to this domain, Perimal-Lewis et al. (2020) highlighted desirable design features in smartphones and mobile apps, such as large buttons and intuitive navigation, which are designed to cater the needs of older adults. The app’s efficacy was shown through simplified call functions and calendar reminders, aiding communication, and task management. Megalingam et al. (2022) emphasize the high desirability of sensor systems for dementia care due to its non-restrictive monitoring capabilities. The system's safety features prevent patients from wandering off, enhancing safety, and its low manufacturing cost makes it cost-effective for widespread adoption and use.

However, there is a need to conduct long-term impact assessments to understand how perceptions of desirability and efficacy evolve among stakeholders in elderly care. Other research gaps include comparative effectiveness and cost analysis, which could support stakeholders in decision-making. Additionally, detailed examination of safety protocols and standards, especially in unsupervised settings are yet to be addressed. Furthermore, the cultural and contextual adaptability of these technologies is underexplored, which could be critical for their acceptance and effectiveness across diverse user groups. If these gaps could be addressed, the implementation and sustained use of dementia care technologies could be significantly enhanced, ensuring positive outcome for stakeholders in elderly care.

4. Conclusion

The five themes identified in this study highlights significant technological advancements and ongoing challenges in dementia care. Emphasis on monitoring and interaction technologies, robots, and immersive technologies such as augmented and virtual reality underscores their potential for improving care outcomes. However, several research gaps were found, particularly with respect to long-term adoption, comparative effectiveness, and holistic integration into existing care models. Specific trends were evident in dementia care which was driven by technological advancements and global healthcare shifts. There is a pressing need for research on the interoperability of new technologies, their economic and ethical implications, in addition to patient and caregiver perspectives to ensure their usability and acceptance in real-life environments.

A multidimensional research approach, blending qualitative and quantitative methods, is essential for capturing these technologies' comprehensive impact. Despite robust empirical evaluations, research gaps in scalability, integration into existing health care systems, and the impact on wider demographic could lead to further observational studies and cross-sectional surveys. The review of papers using the NASSS framework focused primarily on the two domains - key technology characteristics and demand-side value which provided a structured approach to evaluate health care technologies. Addressing the research gaps identified in this study can enhance dementia care technologies, improve the quality of life of patients living with dementia and more significantly, reduce the burden on caregivers. Therefore, future studies in this field are crucial for advancing effective and sustainable dementia care solutions.

5. Limitations

This review paper has several limitations which must be considered in the future work. First, more databases could be added to extend the insights from this review. Second, to emphasize on information systems studies, the AIS eLibrary database search could be extended into the past decade instead of five years. Including the senior scholar's basket of journals could also be considered. Lastly, the NASSS framework was applied to 17 studies from the Information System databases, while additional articles were excluded due to their focus on clinical outcomes, or patient well-being rather than adoption and scalability. Future research should consider applying NASSS to the other 81 studies, integrating clinical and technological perspectives for a more comprehensive understanding of dementia care technologies. This could provide further details on the significance of the domains of NASSS framework for future dementia care solutions.

6. Acknowledgement

We would like to acknowledge NeuronsVR for their invaluable support in providing specialised therapy for Aged Care and individuals living with Dementia through Virtual Reality (VR). We are particularly grateful for their contribution in offering VR training to two members of our research team, which has greatly enhanced our capabilities and enriched our research efforts.

References

Aged Care Research & Industry Innovation Australia (2024). https://www.ariia.org.au/knowledge-implementation-hub/technology-in-aged-care/types-technology-aged-care

Ambika, R., Deekshitha, S., Keerthana, N., & Vandana, K. (2023). Implementation of Wearable Device for Monitoring Alzheimer's Patients. 2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES).

Anwar, U., Arslan, T., Hussain, A., Russ, T. C., & Lomax, P. (2023). Design and Evaluation of Wearable Multimodal RF Sensing System for Vascular Dementia Detection. *IEEE Transactions on Biomedical Circuits and Systems*.

Appel, L., Appel, E., Kisonas, E., Lewis, S., & Sheng, L. Q. (2022). Virtual Reality for Veteran Relaxation: Can VR Therapy Help Veterans Living With Dementia Who Exhibit Responsive Behaviors? *Frontiers in Virtual Reality*, *2*, 724020.

Australian Institute of Health and Welfare (2022). Dementia in Australia. AIHW, Australian Government. https://www.aihw.gov.au/reports/dementia/dementia-in-aus/contents/population-health-impacts-of-dementia/prevalence-of-dementia

Bhargava, Y., & Baths, V. (2022). Technology for dementia care: benefits, opportunities and concerns. *Journal of Global Health Reports*, *6*, e2022056.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Dementia Australia (2023). Dementia Prevalence Data 2024-2054. Australian Institute of Health and Welfare. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/ https://www.dementia.org.au/sites/default/files/2024-01/Prevalence-Data-2024-Updates-All-forms-of-dementia.pdf

Dementia Australia (2024, January 2024). *Key facts and statistics*. Dementia Australia. https://www.dementia.org.au/statistics

Dickinson, R., Kimball, J., Fahed, M., Chang, T., Sekhon, H., & Vahia, I. V. (2023). Augmented Reality (AR) in Dementia Care: Understanding its Scope and Defining its Potential. *The American Journal of Geriatric Psychiatry*, *31*(3), S132-S133.

Enshaeifar, S., Barnaghi, P., Skillman, S., Sharp, D., Nilforooshan, R., & Rostill, H. (2020). A digital platform for remote healthcare monitoring. Companion Proceedings of the Web Conference 2020,

Fixl, L., Parker, S., Starosta-Sztuczka, J., Mettouris, C., Yeratziotis, A., Koumou, S., Kaili, M., Papadopoulos, G. A., & Clarke, V. (2021). eSticky–An Advanced Remote Reminder System for People with Early Dementia. International Conference on ICT for Health, Accessibility and Wellbeing,

Frisardi, V., Soysal, P., & Shenkin, S. D. (2022). New horizons in digital innovation and technology in dementia: potential and possible pitfalls. *European geriatric medicine*, *13*(5), 1025-1027.

Gaikwad, V., Thopate, K., Chame, A., Jyoti, R., Arthamwar, V., & Khadde, M. (2023). A9G-based Dementia GPS Tracker. 2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC),

Garcia-Constantino, M., Orr, C., Synnott, J., Shewell, C., Ennis, A., Cleland, I., Nugent, C., Rafferty, J., Morrison, G., & Larkham, L. (2021). Design and implementation of a smart home in a box to monitor the wellbeing of residents with dementia in care homes. *Frontiers in Digital Health*, *3*, 798889.

García-Requejo, A., Pérez-Rubio, M., Villadangos, J., & Hernández, A. (2022). Indoor-Outdoor Tracking and Activity Monitoring System for Dementia Patients. 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA).

Greenhalgh, T., Wherton, J., Papoutsi, C., Lynch, J., Hughes, G., Hinder, S., ... & Shaw, S. (2017). Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of medical Internet research*, 19(11), e8775.

Hamilton, M. A., Beug, A. P., Hamilton, H. J., & Norton, W. J. (2021). Augmented reality technology for people living with dementia and their care partners. 2021 the 5th International Conference on Virtual and Augmented Reality Simulations,

John, B., Subramanian, R., & Kurian, J. C. (2023). Design and Evaluation of a Virtual Reality Game to Improve Physical and Cognitive Acuity.

Kocher, S., Safikhani, S., & Pirker, J. (2022). Work-In-Progress—Exploring the Feasibility of Using Hand Tracking in VR Application for Memory Training

Exercises. 2022 8th International Conference of the Immersive Learning Research Network (iLRN),

Lee-Cheong, S., Amanullah, S., & Jardine, M. (2022). New assistive technologies in dementia and mild cognitive impairment care: A PubMed review. *Asian Journal of Psychiatry*, *73*, 103135.

Matsangidou, M., Frangoudes, F., Solomou, T., Papayianni, E., & Pattichis, C. (2022). Free of walls: Participatory design of an out-world experience via virtual reality for dementia in-patients. Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization.

Matsangidou, M., Solomou, T., Frangoudes, F., Papayianni, E., & Pattichis, C. S. (2023). Offering Outworld Experiences to In-Patients With Dementia Through Virtual Reality: Mixed Methods Study. *JMIR aging*, *6*(1), e45799.

Megalingam, R. K., Kota, A. H., & Reddy, C. P. K. (2022). Indoor Tracking of Dementia Patients without GPS. 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM),

Moreno, L., & Martínez, P. (2023). Managing daily living activities for people with acquired brain injury using the DailyCare application. XXIII International Conference on Human Computer Interaction,

Pandhi, S., & Tiwari, R. (2022). Dementia Care: An Android Application for Assisting Dementia Patients. 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM),

Perimal-Lewis, L., Maeder, A., Gordon, S., & Tieman, J. (2020). A tablet-based memory enhancement application for older users: Design approach.

Pratama, E. R., Renaldi, F., Umbara, F. R., & Djamal, E. C. (2020). Geofencing technology in monitoring of geriatric patients suffering from dementia and alzheimer. 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE).

Schweiger, N., & Wolff, C. (2023). Robotic Support for Haptic Dementia Exercises. 2023 IEEE 11th International Conference on Serious Games and Applications for Health (SeGAH),

Tarbert, R. J., & Singhatat, W. (2023). Real world evidence of wearable smartbelt for mitigation of fall impact in older adult care. *IEEE journal of translational engineering in health and medicine*, *11*, 247-251.

Thorpe, J., Forchhammer, B. H., & Maier, A. M. (2019). Adapting mobile and wearable technology to provide support and monitoring in rehabilitation for dementia: feasibility case series. *JMIR formative research*, *3*(4), e12346.

Vollmer Dahlke, D., & Ory, M. G. (2020). Emerging issues of intelligent assistive technology use among people with dementia and their caregivers: A US Perspective. Frontiers in Public Health, 8, 191.

World Health Organization. (2022). *Global report on assistive technology*. WHO. https://www.who.int/publications/i/item/9789240049451

Revolutionising Education: Leveraging AI to Boost Student Engagement through Constructivist and Social Collaborative Learning - A Study of Perusall

Claire Li

Royal Holloway, University of London, United Kingdom

Xiangping Du

University of Hertfordshire, United Kingdom

Perry Xiao

London South Bank University, United Kingdom

Completed Research

Abstract

This study examines the integration of Constructivist and Social Learning Theories within AI-enhanced collaborative community learning, with a focus on the implementation of Perusall, a social reading and peer annotation platform. Constructivist theory foregrounds active knowledge construction, while Social Learning Theory emphasises learning through social interaction and observation. Their convergence supports pedagogical approaches that promote exploration, problem-solving, and collective reflection. Drawing on three years of data from a postgraduate module at a UK university, this research compares student engagement and performance before and after the adoption of Perusall. Findings indicate that the platform enhances learner engagement and academic outcomes by fostering interactive and collaborative learning environments. The study offers empirical evidence on the pedagogical efficacy of AI-driven tools, contributing to the theoretical discourse on technology-mediated learning and providing practical implications for educators. It highlights the transformative potential of AI in fostering more engaging, effective, and inclusive educational practices.

Keywords: AI; Perusall; machine learning; collaborative learning; community-based learning; student engagement and performance

1.0 Introduction

Artificial Intelligence (AI) has increasingly become a transformative force in higher education, contributing significantly to student learning through its ability to personalise, enhance, and streamline the learning experience (Saaida, 2023; Babu and Wooden, 2023). AI technologies, such as intelligent tutoring systems, adaptive learning platforms, and natural language processing tools, are designed to meet the diverse needs of students by providing tailored educational experiences (Chen, et al, 2020). These systems analyse vast amounts of data from student interactions to understand individual learning styles, strengths, and areas needing improvement. By doing so, AI can customise lesson plans, adjust the difficulty of tasks, and offer personalised feedback in real-time, ensuring that each student receives instruction that is best suited to their unique learning trajectory. Additionally, AI facilitates more efficient and effective communication and collaboration among students and between students and educators (Chen, et al. 2020). For instance, AI-powered chatbots, virtual assistants and peer supporting tools such as Perusall can provide instant support and answer queries, freeing up teachers to focus on more complex instructional tasks. Through these capabilities, AI not only enhances the learning experience by making it more engaging and responsive to individual needs but also helps to improve learning outcomes by fostering a more adaptive and supportive educational environment (Gligorea et al., 2023).

The integration of technology in education has transformed traditional learning paradigms, offering new opportunities to enhance student engagement and performance (Trowler, 2010). Among the various theoretical frameworks that guide educational practices, Constructivist Learning Theory and Social Learning Theory stand out for their emphasis on active, engaged learning processes (Mukhalalati and Taylor, 2019). Constructivist Learning Theory, rooted in the works of Jean Piaget and Lev Vygotsky, suggests that learners actively construct their own understanding through experiences and reflection (Gordon, 2009). This theory underscores the importance of personal engagement and the integration of prior knowledge in the learning process. On the other hand, Social Learning Theory, developed by Albert Bandura highlights the significance of social interactions, observational learning, and modelling in educational settings (Tran, 2013). This theory emphasises that learning occurs within a social context and is greatly influenced by observing the behaviour and attitudes of others.

Collaborative community learning represents an educational approach that synthesises the core principles of both Constructivist and Social Learning Theories (Blayone et al., 2017). By engaging students in group activities that promote exploration, problem-solving, and shared reflection, collaborative learning environments foster both individual cognitive development and social interaction (Kreijns, et al. 2003). These environments provide opportunities for students to actively participate in their learning, build on their prior experiences, and learn from their peers through observation and feedback. The complementarity of these theories suggests that combining individual and social aspects of learning can create a more dynamic and effective educational experience.

In recent years, AI has emerged as a powerful tool to further enhance collaborative community learning. AI technologies personalise learning experiences, facilitate communication, and provide data-driven insights into group dynamics. By leveraging AI, educators can create enriched learning environments that support the active engagement and social interaction essential to both Constructivist and Social Learning Theories. AI-mediated personalization ensures that each student is appropriately challenged, while AI-powered collaboration tools foster effective group work and communication. Additionally, AI-driven analytics provide valuable feedback to both students and educators, helping to optimize learning outcomes.

Perusall is an AI-enhanced social reading platform enables students to annotate, comment and interact with each other; it transforms course reading content into a social learning experience (Adams & Wilson, 2020). Perusall uses generative AI and machine learning to grade student engagement based on its own pre-trained model, including reading fully, active engagement time, commenting, upvoting, eliciting and getting responses. In addition, it provides individual student instant automated feedback for further improvement, and supplies educators detailed automated analytics and reports, i.e. 'student confusion report' indicating student confusion and misunderstanding as well as student activity and engagement with reading individually and collectively (Perusall, 2025). That means through Perusall educators can get continuous insights on students' understanding and engagement while students get instant feedback and motivation which in turn drives deeper understanding and meaningful engagement and attainment (Bharath, 2021). It is reported that students' engagement with reading on Persuall is over 90% compared to the usual 20%-30% engagement with normal assigned reading in general (Perusall, 2025). Miller (2018) and Bharath (2021) advocate that Perusall engages students in scale, facilities flipped learning, fosters a collaborative learning community where students are supported and motivated for learning.

However, besides the AI-enhanced features of Persuall and its role in enhancing student engagement and performance, there seems mere literature focusing on 758,855 overseas students studying in UK higher education institutions as of 2022/23 (Bolton, et al, 2024). International students account for 26% of all students at UK universities; within which 52% are postgraduate students (HESA, 2024). Their adoption of Persuall and the impact of Perusall on their engagement and academic performance remain undiscovered. With the trend of growing number of international students coming to study in the UK, it is vital to discover the answers to contribute to the theoretical paradigm and pedagogical practice.

This paper explores the integration of AI in collaborative community learning environments i.e., Perusall, examining how AI can enhance the principles of Constructivist and Social

Learning Theories to improve student engagement and performance. The following sections will delve into the theoretical underpinnings of these learning theories, the synergistic benefits of their complementarity, and the ways in which AI technologies can mediate and amplify these effects. By providing a comprehensive analysis of these elements, this study aims to contribute to the ongoing discourse on the transformative potential of AI in education.

This study contributes significantly to the theoretical application of both Constructivist Learning Theory and Social Learning Theory by demonstrating how AI technologies can be effectively integrated to enhance the learning processes posited by these frameworks. Constructivist theory, with its focus on active, experiential learning and personal knowledge construction, is augmented by AI's ability to provide personalised and adaptive learning experiences. By tailoring educational content to individual learners' needs and providing immediate, contextualised feedback, AI technologies support the constructivist emphasis on active engagement and deep, meaningful learning. This study illustrates how AI can create more interactive learning environments, thereby advancing the practical applications of Constructivist Learning Theory.

Similarly, the study expands Social Learning Theory by incorporating AI's capabilities to facilitate and enhance social interactions and observational learning within educational settings. AI-driven platforms can simulate social interactions, model behaviour through digital agents, and provide tools for collaborative learning that are more interactive and engaging. These technologies not only support the observation and imitation processes central to Social Learning Theory but also enhance the social dimension of learning by enabling more dynamic and effective communication and collaboration among students. This integration of AI into social learning contexts provides a modern extension to Social Learning Theory, showcasing its relevance and adaptability in contemporary educational landscapes.

The study also makes a substantial contribution to the literature on collaborative community learning by providing empirical evidence on the effectiveness of AI-enhanced collaborative environments. While the benefits of collaborative learning are well-documented, this study bridges a crucial gap by exploring how AI can further optimise these benefits. Through detailed analysis and practical examples, the research demonstrates how AI technologies can facilitate more efficient group dynamics, enhance student engagement, and improve overall learning outcomes. This enriches the existing body of knowledge by offering concrete strategies and insights into the implementation of AI in collaborative learning settings, thereby providing a valuable resource for educators and researchers alike.

This study addresses the practical implications of integrating AI into educational practices. It provides educators with actionable insights into how AI can be leveraged to support and enhance both individual and social aspects of learning. By showcasing successful use of Perusall as a case study, the research offers a pragmatic approach to adopting AI technologies, making it a valuable guide for practitioners aiming to improve student engagement and performance through innovative educational strategies.

Finally, the study sets the stage for future research by identifying key areas where further investigation is needed. It highlights the potential of AI to transform educational practices and calls for continued exploration into the long-term impacts of AI on learning outcomes, student engagement, and educational equity. By proposing a framework for integrating AI with Constructivist and Social Learning Theories, the study encourages future researchers to build on its findings and explore new ways to leverage technology to enhance learning. This forward-looking perspective not only underscores the study's relevance to current educational challenges but also its contribution to shaping the future direction of research in the field of educational technology.

2.0 Literature review

2.1 Constructivist learning theory and social learning theory

Constructivist Learning Theory is rooted in the idea that learners actively construct their own understanding and knowledge of the world through experiences and reflection (Bada and Olusegun, 2015). This theory, heavily influenced by the works of Jean Piaget and Lev Vygotsky, emphasises that learning is an active, contextualised process of constructing meaning rather than passively receiving information (Gash, 2014). Constructivist classrooms are typically learner-centred, where students are encouraged to explore, ask questions, and engage in problem-solving activities. Teachers in this framework act as facilitators, guiding students to discover principles for themselves and build on their prior knowledge through meaningful interactions and real-world applications.

Social Learning Theory, developed by Albert Bandura, posits that people learn within a social context and that learning occurs through observation, imitation, and modelling (Bandura and Walters, 1977). This theory suggests that behaviour is learned from the environment through the process of observational learning, where individuals watch others and then replicate their actions (Bandura and Walters, 1977). It highlights the role of reinforcement and punishment in learning and emphasises the impact of social influences such as family, peers, and media.

Both Constructivist Learning Theory and Social Learning Theory recognise the active role of the learner in the educational process. They emphasise that learning is not a passive absorption of information but involves active engagement and interaction with the environment. Both theories also acknowledge the importance of prior knowledge and experiences in shaping new learning. Constructivist theory focuses on how individuals build on their existing cognitive structures, while social learning theory emphasises the role of social context and observational experiences in learning. Furthermore, both theories advocate for environments that encourage exploration, interaction, and engagement to facilitate deeper understanding and retention of knowledge.

The primary differences between Constructivist Learning Theory and Social Learning Theory lie in their focus on the sources and mechanisms of learning. Constructivist Learning Theory emphasises *internal* cognitive processes, where learning is driven by the learner's active engagement and personal reflection. It focuses on how individuals construct knowledge through experiences and prior understanding, often in a self-directed and exploratory manner. In contrast, Social Learning Theory places greater emphasis on *external* social influences, highlighting the role of observation and imitation of others within a social context. It stresses the importance of social interactions, models, and the environment in shaping behaviour and learning outcomes.

Constructivist Learning Theory is often associated with hands-on, experiential learning activities that promote critical thinking and problem-solving skills. These activities are designed to be learner-centred, allowing individuals to explore and construct their own understanding. On the other hand, Social Learning Theory is closely linked with observational learning and the use of role models. It involves learning through watching others, understanding the consequences of actions, and then applying this knowledge in similar situations. Social Learning Theory also incorporates the concepts of reinforcement and punishment to explain how behaviours are acquired and maintained.

Constructivist Learning Theory and Social Learning Theory share commonalities that highlight the importance of active engagement and the role of prior knowledge. Constructivist Learning Theory focuses on the *internal* cognitive construction of knowledge through experience and reflection, whereas Social Learning Theory emphasises the *external* social influences and observational learning. Together, these theories provide a comprehensive understanding of how individuals learn, both through personal experiences and social interactions.

One key area of complementarity lies in the recognition that learning is both a personal and a social endeavour. Constructivist theory highlights the importance of learners actively making sense of information based on their prior knowledge and experiences, which can be enriched through social interactions as posited by Social Learning Theory. For instance, collaborative learning activities allow learners to engage in dialogue, share diverse perspectives, and co-construct understanding, integrating personal insights with social experiences. This integration aligns with Vygotsky's concept of social constructivism, which bridges the two theories by emphasising the social nature of cognitive development and the role of social interaction in scaffolding individual learning.

Moreover, the complementarity of these theories is evident in the practical applications of educational strategies that incorporate both individual and social dimensions of learning. In a constructivist classroom, learners might engage in hands-on activities, projects, and problem-solving tasks that encourage active exploration and reflection. When these activities are carried out in a collaborative setting, as suggested by Social Learning Theory, learners benefit from observing peers, receiving feedback, and engaging in discussions that enhance their understanding. The blend of individual and social learning experiences fosters a richer, more dynamic learning environment where learners can internalize and apply knowledge more effectively.

In essence, the integration of Constructivist Learning Theory and Social Learning Theory creates a synergistic approach that leverages the strengths of both perspectives. By acknowledging that learning is both an individual cognitive process and a socially mediated activity, educators can design instructional strategies that provide opportunities for active engagement, personal reflection, and meaningful social interaction. This comprehensive approach not only supports the development of individual knowledge and skills but also fosters a collaborative and supportive learning community.

2.2 AI and constructivist learning theory

Constructivist learning theory posits that learners actively construct their own understanding and knowledge through experiences and reflection (Mvududu and Thiel-Burgess, 2012). AI can enhance this process by providing personalised learning environments that adapt to the individual needs and prior knowledge of each learner. Adaptive learning platforms are prime examples of AI applications that support constructivist learning. These systems use data from learners' interactions to tailor instructional content, provide immediate feedback, thus facilitating a more personalised and self-directed learning experience. Moreover, AI can facilitate constructivist learning by enabling collaborative learning opportunities. Through natural language processing and machine learning algorithms, AI can analyse and understand the contributions of each learner in a collaborative setting, providing insights and feedback that help to scaffold the learning process. AI-powered collaboration tools can match learners with peers who have complementary knowledge and skills, fostering an environment where learners can engage in meaningful dialogue, share perspectives, and co-construct knowledge. By leveraging AI, educators can create dynamic and adaptive learning ecosystems that support the constructivist emphasis on active, experiential, and individualised learning.

2.3 AI and social learning theory

Social learning theory, developed by Albert Bandura, emphasises the importance of observing and modelling the behaviours, attitudes, and emotional reactions of others (Bandura and Walters, 1977). AI can play a significant role in enhancing social learning environments by facilitating observation and interaction through advanced technologies. For instance, AI-driven platforms can host virtual classrooms and accommodate peer annotations functions where learners can observe and interact with avatars or digital agents that demonstrate desired behaviours and skills. These digital agents can be programmed to exhibit a range of behaviours and provide immediate feedback, thereby serving as effective models for learners to emulate.

AI also enhances social learning through social media and online collaborative tools, which are increasingly used in educational settings. These platforms, powered by AI, can analyse interactions and provide insights into group dynamics, identify key influencers, and suggest interventions to promote positive social behaviours and collaboration. AI algorithms can monitor discussions and highlight contributions that align with learning objectives, ensuring that learners are exposed to high-quality content and diverse perspectives. Furthermore, AI can support the assessment of social learning by tracking participation, engagement, and the application of learned behaviours in real-world contexts, providing educators with valuable data to inform instructional strategies.

By integrating AI into social learning environments, educators can create more interactive, engaging, and effective learning experiences that align with the principles of social learning theory. AI not only facilitates the observation and imitation of behaviours but also enhances the social dimensions of learning by fostering collaboration, communication, and community building among learners.

Incorporating AI into both constructivist and social learning frameworks offers the potential to transform educational experiences. For constructivist learning, AI provides personalised, adaptive, and immersive learning environments that align with the learner's active role in constructing knowledge. In the context of social learning, AI facilitates observation, interaction, and collaboration, enhancing the social aspects of learning through advanced technologies. By leveraging the capabilities of AI, educators can create enriched learning environments that support and enhance the principles of both constructivist and social learning theories. Figure 1 below is our proposed framework showing the complementarity of constructivist and social learning theories and how AI may play a role in the association.

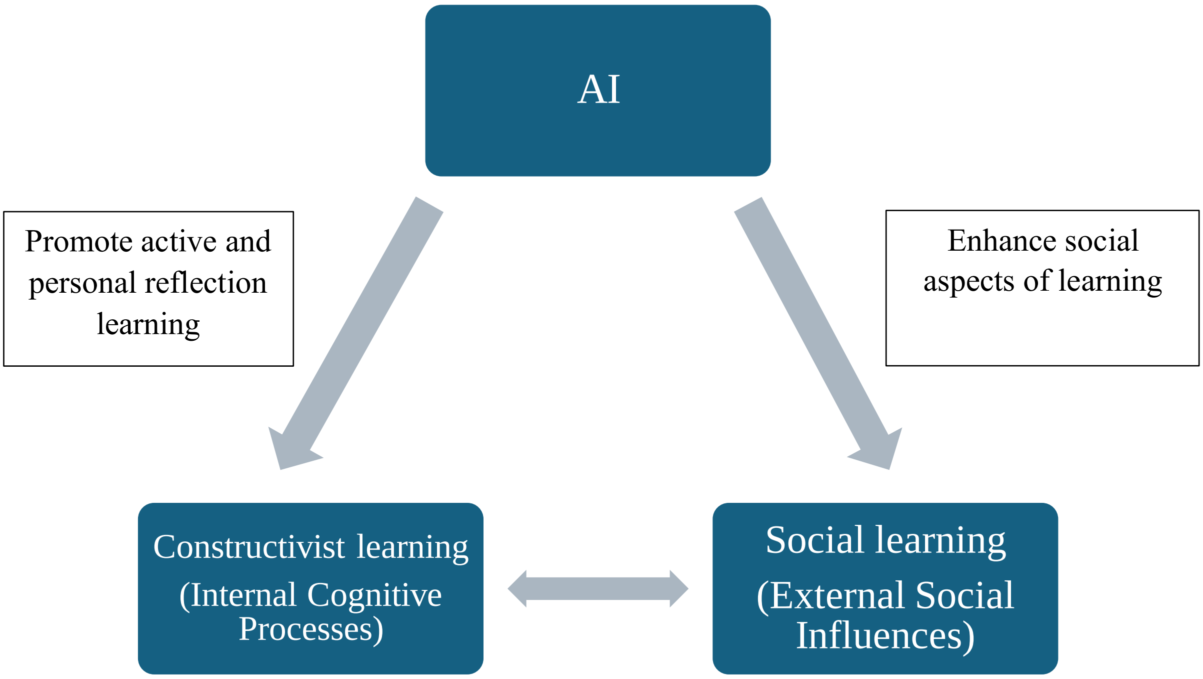

AI
Promote active and personal reflection learning
Enhance social aspects of learning
Constructivist learning (Internal Cognitive Processes)
Social learning (External Social Influences)

Figure 1. Framework – AI, constructivist and social learning

Figure 1 visually represents how AI can be integrated into both Constructivist and Social Learning Theories. AI enhances the internal cognitive processes emphasised by Constructivist Learning Theory by providing personalised, adaptive, and immersive learning experiences. Simultaneously, AI supports the external social influences emphasised by Social Learning Theory by facilitating observation, interaction, and collaboration, thus enriching the social dimensions of learning.

2.4 AI, collaborative community learning, student engagement and performance

Constructivist Learning Theory posits that learners actively construct their own understanding and knowledge through experiences and reflection. Within the context of collaborative community learning, this theory emphasises the importance of students engaging in group activities that promote exploration, questioning, and problem-solving. Collaborative learning environments allow students to actively participate in their learning process, building new knowledge based on their interactions with peers and the tasks they undertake. This active engagement is crucial for deep learning, as it encourages students to apply concepts in practical, real-world situations, thus enhancing their understanding and retention of the material. Additionally, the collaborative nature of these activities helps students to integrate their prior knowledge and experiences, leading to a richer, more nuanced understanding of the subject matter. The opportunity to reflect on their learning and receive feedback from peers further reinforces their cognitive development, making collaborative community learning a powerful application of constructivist principles.

Social Learning Theory, developed by Albert Bandura, emphasises the role of social interactions and observational learning in the educational process. In collaborative community learning settings, students benefit from observing the strategies and approaches used by their peers, which can model successful behaviours and techniques. This observational learning is particularly effective in reinforcing desired behaviours and correcting misconceptions. Social interactions within the group also provide immediate feedback, enabling students to learn from each other and improve their understanding in real-time. The sense of belonging and motivation that arises from being part of a supportive learning community further enhances student engagement. The social presence in collaborative learning settings encourages active participation and fosters a sense of accountability, as students are more likely to contribute meaningfully when they feel connected to their peers.

The complementarity of Constructivist Learning Theory and Social Learning Theory provides a comprehensive framework for understanding the benefits of collaborative community learning. Both theories recognise the active role of the learner and the importance of prior knowledge in the learning process. Constructivist theory focuses on how individuals construct knowledge through personal experiences and reflection, while Social Learning Theory highlights the significance of social interactions and the observation of others. Together, these theories suggest that collaborative learning environments, which combine personal exploration with social interaction, can significantly enhance student engagement and performance. By providing opportunities for students to actively engage with the material, reflect on their learning, and interact with peers, collaborative community learning environments create a dynamic and supportive atmosphere that promotes deep, meaningful learning.

AI can significantly enhance the impact of collaborative community learning by mediating the association between student engagement and performance through several mechanisms. AI-driven adaptive learning platforms can personalise the learning experience for each student, ensuring that all members of the group are working at an appropriate level of challenge. This personalisation helps maintain high levels of engagement and ensures meaningful contributions from all students. Intelligent tutoring systems provide personalised hints and feedback, keeping students on track and engaged in their learning.

AI also facilitates communication and collaboration within the learning community. Natural Language Processing tools can analyse and summarise group discussions, highlight key points, and translate languages in multilingual groups, making communication smoother and more effective. AI-powered collaboration tools can match students with complementary skills and knowledge, fostering productive and effective group work. Additionally, AI can provide valuable data-driven insights into group dynamics and individual contributions, helping educators to better support their students. Learning analytics can identify patterns in student interactions and suggest effective collaborative strategies, while real-time feedback systems monitor group activities and address issues as they arise.

Incorporating AI into collaborative community learning environments enhances both Constructivist and Social Learning Theory principles. AI-mediated personalisation, communication facilitation, and data-driven insights create enriched learning environments that maximise student engagement and performance. This synergy between human cognitive processes and advanced technology offers a comprehensive approach to modern education, leveraging the strengths of both internal cognitive construction and external social influences.

Based on the above discussion, this study proposes the following hypotheses.

Hypothesis 1. Collaborative learning improves student engagement and performance.

Hypothesis 2. AI-enahnced collaborative community learning platform, i.e. Perusall mediates the possibly positive association between collaborative learning and student engagement and performance.

3.0 Methodology

This secondary quantitative research adopts positivist philosophy, where the researchers value facts and consider knowledge gained through observations and measurement in an objective way. It involves the use of literature to develop hypotheses to be tested during the research process (Saunders et al, 2019). In this research, students' engagement with learning materials and their academic performance on a postgraduate course in a UK university were measured across three academic years, i.e. 2020-21, 2021-22 and 2022-23. Written permission to use the module's information and student's engagement and performance data were obtained from the designated data owner of the participating University. The ethics protocol number is BUS/SF/UH/06197.

This research firstly examines the student engagement with the module's weekly reading materials and then measures the student performance year by year from 2020 to 2022. The examination focuses on the role of the collaborative AI learning tool, i.e. Perusall in engaging students and facilitating collaborative social learning.

3.1. The experimental AI tool: Persuall

Perusall is a collaborative reading and annotation tool for community-based learning that is suitable for both in-person and online environments (Adams & Wilson, 2020; Bharath, 2021). It helps to foster collaborative reading and learning that allows for interaction between students, their peers, teaching resources, and tutors. Persuall also offers educators an organised tool to flip reading, and to monitor student engagement, achievement and progress. This is because Persuall provides a space where students can annotate (study or reflect), identify areas of low comprehension (ask for help), and interact with other readers (engage in social and peer learning) (Bharath, 2021). Miller (2018) and Bharath (2021) state, from their experience of using Persuall, students who undertake a reading prior to their seminar often arrive better prepared and with greater confidence to discuss the subject openly in class.

What is more, Perusall uses AI and machine learning algorithm to assesses the quality of annotations and students' engagement. It awards scores for the 'insightfulness' of their annotations by looking into the linguistic features of the text on any flipped reading or learning materials (Perusall, 2025). The score can be used as formative or summative assessment to encourage better student engagement and interaction with readings, as well as to prepare students for assessments that require application, analysis and critical evaluation of teaching resources. It is an effective tool for collaborative community learning (Ateh, 2024).

3.2. The context and data of the research

This research examines students' engagement and performance of a large UK postgraduate module where case studies were required to be read before class for understanding, discussion in class and application in their assessment. It spans over three years when different module design, delivery and assessment were adopted. Students in this module were from overseas, mostly India, Pakistan, Bagaladah and Nigeria, resembles the overall UK international student population (Bolton, et al 2024).

In 2020-21, there were 463 international students registered on the module and the module was delivered 100% online due to the pandemic. All learning materials were featured on the University's VLE, i.e. Canvas where students were expected to engage prior to online tutorials for discussion. Students' engagement with the materials were assessed by individual reflections. The student engagement data with the weekly materials was extracted via Canvas Analytics.

In 2021-22, the module had 778 students and adopted blended delivery with on-campus lectures and online tutorials. Persuall was then applied for students to engage with the weekly reading materials. While reading the materials in Persuall, students were automatically allocated to small groups and students were required to read and make annotations on the materials, write comments, view and provoke others' comments so they can communicate and learn from each other anonymously. They would then discuss their learning in the respective week's tutorials. It is worth noting that students' interaction and engagement with Persuall were captured using pre-set infrastructures as they read, annotate and comment. Meanwhile, a grade was also awarded to each student, using Perusall's AI algorithm, based on their individual contribution and engagement with reading material and collaborative learning. .

In 2022-23, 902 students were registered on the module and they were taught on-campus with no online or Persuall applied. The students were assessed byt reflections on group learning conversations took place during tutorials. Table 1 below summarise the three years' module design, delivery and assessment:

Table 1: Characteristics of the Module Design & Assessment

Academic Year	*Student No.*	*Delivery*	*Features*	*Assessed*
2020-21	463	Online	10 case studies uploaded on Canvas for students to engage. The case studies are discussed during online tutorials	Yes; 1 piece of written reflection on students' learning from 5 out of the 10 case studies. Manually marked by the tutorial tutors.
2021-22	778	Blended	5 case studies uploaded in Persuall where students need to make annotation and comments on the case studies, view and learn from each other	Yes; a minimum of 5 pieces of reflective comments and annotations recording student learning of each case study and/or from others' annotations and comments
2022-23	902	On-campus	5 pieces of reflection on weekly group learning conversations took place in tutorials	Yes, 5 x 250-word reflection on a weekly conversation amongst a group of students (learning set) during tutorials.

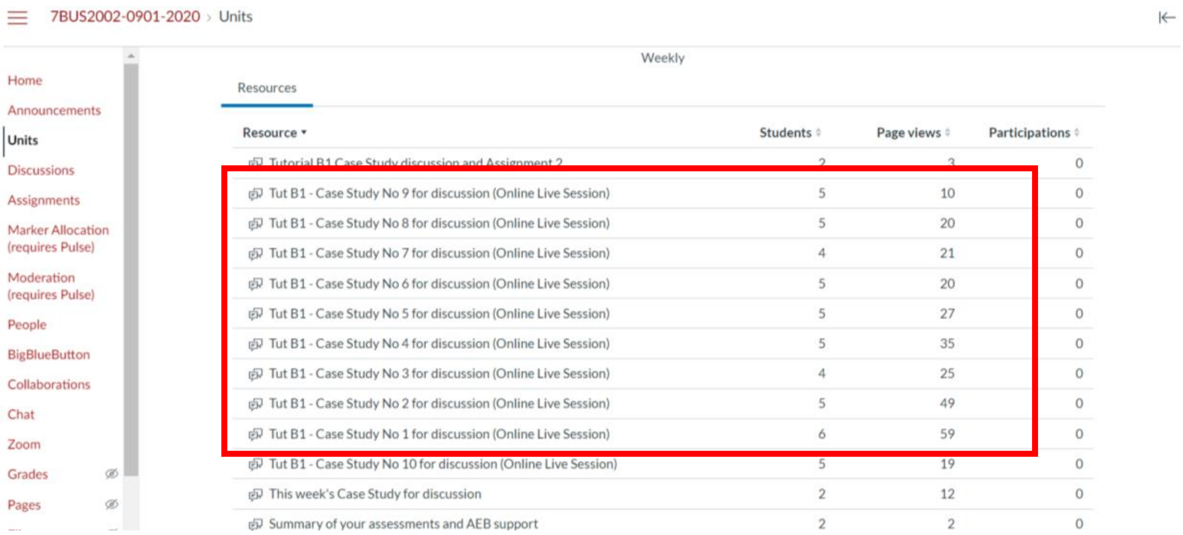
As shown above, the module was designed building in engagement with learning materials and associated with assessment to encourage engagement across the three academic years.

4.0 Findings and analysis

The research presents two sets of data consisting of 1) student engagement data with learning materials and 2) student academic performance data with grade breakdowns. The data will test hypothesis 1: Collaborative learning improves student engagement and performance. What is more, the research presents detailed machine learning regression analysis of Perusall engagement and examines the association between collaborative learning and students' engagement and performance as outlined in hypothesis 2.

4.1 Student engagement with Canvas

The postgraduate module was hosted on Canvas, the University's VLE where the students need to access the learning materials. Below figures (Figure 2, 3 and 4) demonstrate student engagement data with Canvas and the learning materials. 'Resources' refer to the Canvas materials students are expected to engage, such as case studies. 'Students' counts for unique students who have access a page at the time. 'Page views' counts of hits for a specific page and 'Participations' counts of times students have contributed to a specific page. Hence, 'Page reviews' should be considered when measuring student engagement with the learning materials.



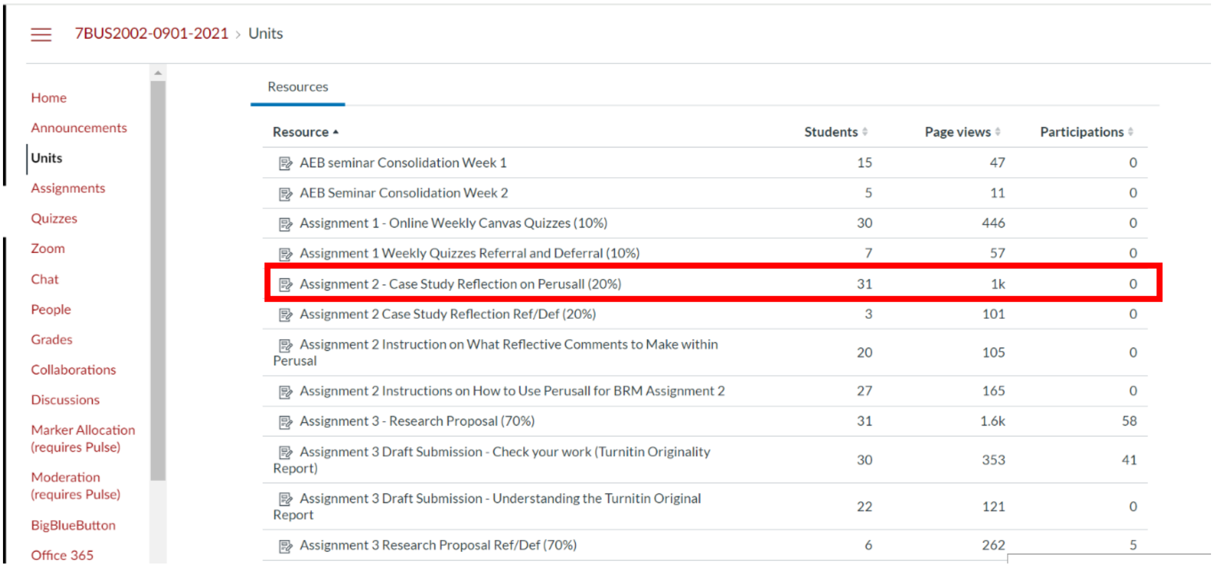
Resource ▾	Students	Page views	Participations
Tutorial B1 Case Study discussion and Assignment 2	2	3	0
Tut B1 - Case Study No 9 for discussion (Online Live Session)	5	10	0
Tut B1 - Case Study No 8 for discussion (Online Live Session)	5	20	0
Tut B1 - Case Study No 7 for discussion (Online Live Session)	4	21	0
Tut B1 - Case Study No 6 for discussion (Online Live Session)	5	20	0
Tut B1 - Case Study No 5 for discussion (Online Live Session)	5	27	0
Tut B1 - Case Study No 4 for discussion (Online Live Session)	5	35	0
Tut B1 - Case Study No 3 for discussion (Online Live Session)	4	25	0
Tut B1 - Case Study No 2 for discussion (Online Live Session)	5	49	0
Tut B1 - Case Study No 1 for discussion (Online Live Session)	6	59	0
Tut B1 - Case Study No 10 for discussion (Online Live Session)	5	19	0
This week's Case Study for discussion	2	12	0
Summary of your assessments and AEB support	2	2	0

Figure 2: 2020-21 Canvas Engagement Data with non-Assessed Weekly Reading Material

Figure 2 above illustrates there has been very limited engagement with the learning materials which was featured under 'Units' section and not directly related to assignment. It is worth noting that in 2020-21 amongst over 400 students, the number of students who accessed the 'Tut B1 – Case Study No X for discussion (Oline Liver Session) was very small which counts for 1% with an average page review of 29 hits. This may be because of the 1st year of pandemic when teaching was converted online, and students were struggling with online learning. Meanwhile, the case studies would also be shared by the tutors in tutorials, explained and discussed so not engaging with them beforehand were unlikely to disadvantage the students.

As the pandemic has eased, a more flexible method of delivery featuring blended on-campus and online was adopted in 2021-22. Learning from the low engagement in 2020-21, the innovative AI collaborative community reading tool, Perusall was adopted in the module to encourage collaborative and social learning. In this way, all reading materials were uploaded to Perusall for students to engage. Upon access Persuall, students would be randomly assigned to groups where they could see each other online, communicate with one another and view others' annotation and comment on them. Accordingly, students' engagement with the materials, including reading to the end, active reading, commenting and provoking would be assessed using Persuall's pre-designed AI model. Meanwhile, each individual student's performance is measured using Persuall's algorithm, supported by an automated just-in-time feedback for improvement. However, given Persuall was not integrated within Canvas, students need to access Canvas module site first before joining Perusall. Therefore, Persuall was featured as 'Assignment' in Canvas.

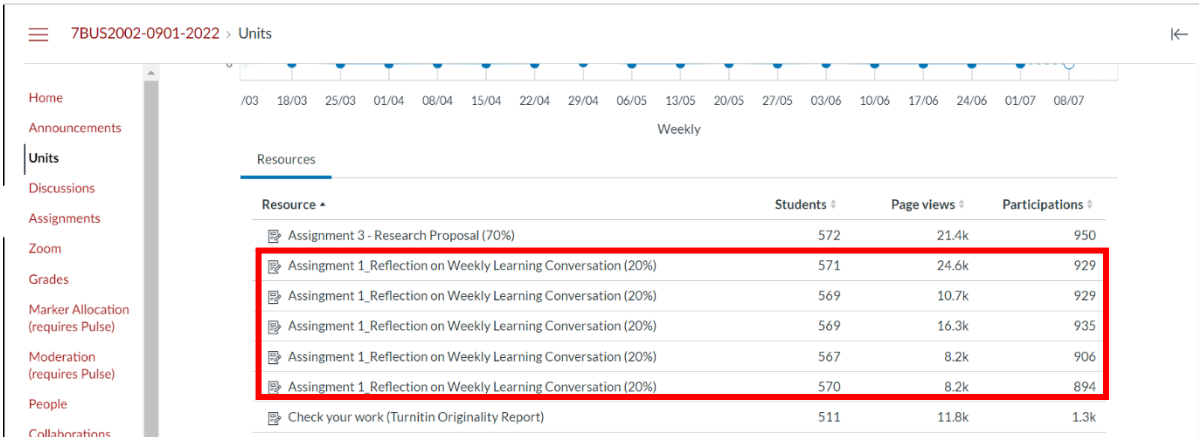
Figure 3 below demonstrates students' engagement with the "Assignment 2 – Case Study Reflection on Perusall". As it demonstrates that there are about 1k page views amongst 778 students who have accessed the assignment instructions and how to use Persuall for the assessment.



Resource ▴	Students	Page views	Participations
AEB seminar Consolidation Week 1	15	47	0
AEB Seminar Consolidation Week 2	5	11	0
Assignment 1 - Online Weekly Canvas Quizzes (10%)	30	446	0
Assignment 1 Weekly Quizzes Referral and Deferral (10%)	7	57	0
Assignment 2 - Case Study Reflection on Perusall (20%)	31	1k	0
Assignment 2 Case Study Reflection Ref/Def (20%)	3	101	0
Assignment 2 Instruction on What Reflective Comments to Make within Perusal	20	105	0
Assignment 2 Instructions on How to Use Perusall for BRM Assignment 2	27	165	0
Assignment 3 - Research Proposal (70%)	31	1.6k	58
Assignment 3 Draft Submission - Check your work (Turnitin Originality Report)	30	353	41
Assignment 3 Draft Submission - Understanding the Turnitin Original Report	22	121	0
Assignment 3 Research Proposal Ref/Def (70%)	6	262	5

Figure 3: 2021-22 Canvas Engagement Data with Perusall Assessment

In 2022-23, teaching converted to 100% on campus. However, due to University's licensing restriction, the module could no longer use Persuall but rely on Canvas. In order to mirror the collaborative community learning model like Perusall, the module adopted group learning conversations in tutorials where students could discuss the case studies and share their learning, followed by an individual reflection about the learning conversation submitted and assessed each week.



Resource ▴	Students	Page views	Participations
Assignment 3 - Research Proposal (70%)	572	21.4k	950
Assingment 1_Reflection on Weekly Learning Conversation (20%)	571	24.6k	929
Assingment 1_Reflection on Weekly Learning Conversation (20%)	569	10.7k	929
Assingment 1_Reflection on Weekly Learning Conversation (20%)	569	16.3k	935
Assingment 1_Reflection on Weekly Learning Conversation (20%)	567	8.2k	906
Assingment 1_Reflection on Weekly Learning Conversation (20%)	570	8.2k	894
Check your work (Turnitin Originality Report)	511	11.8k	1.3k

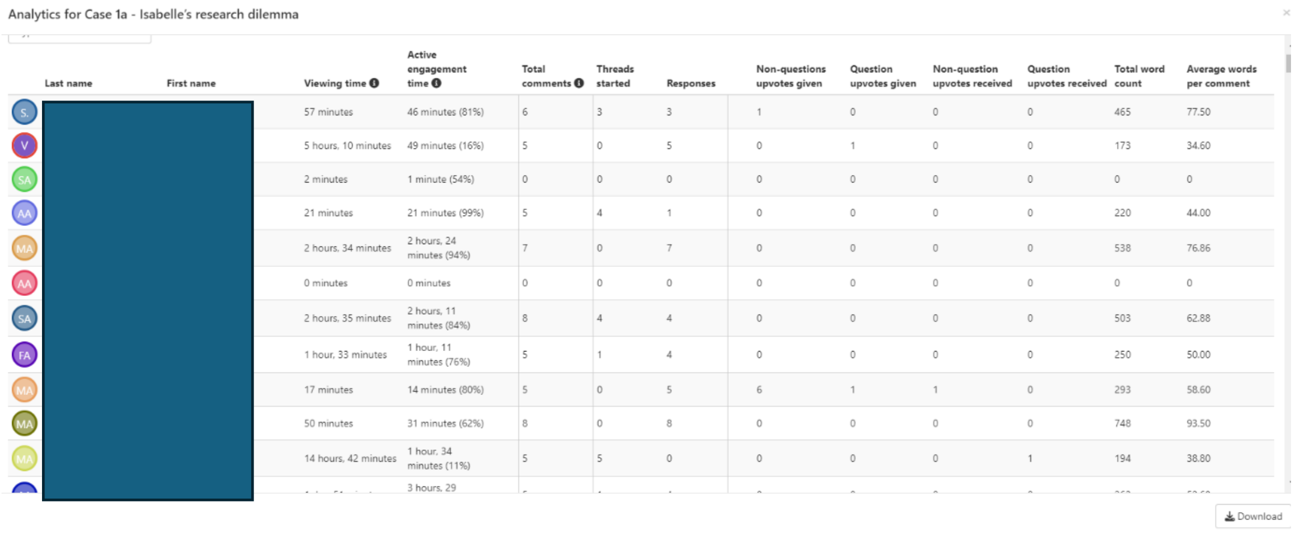
Figure 4: 2022-23 Canvas Engagement Data with Assessed Weekly Learning Conversations

Figure 4 above shows amongst 902 students, there are about 13.6k page reviews on average and over 900 participations of the weekly reflection assignment which directly links to the reflective learning conversations.

From the above Figures, it is obvious that students have had better engagement when there was a collaborative learning community in place. This was clearly demonstrated in 2021-22 when the community reading tool Perusall was implemented and in 2022-23 when group learning conversations were required. However, it could be argued that the engagement attributes to the direct link of the learning materials to the assessment.

4.2 Student engagement with Perusall

Between the AI-mediated community learning environment Perusall and the human-led learning conversations, the role of AI in support of the collaborative learning and student engagement seems apparent.



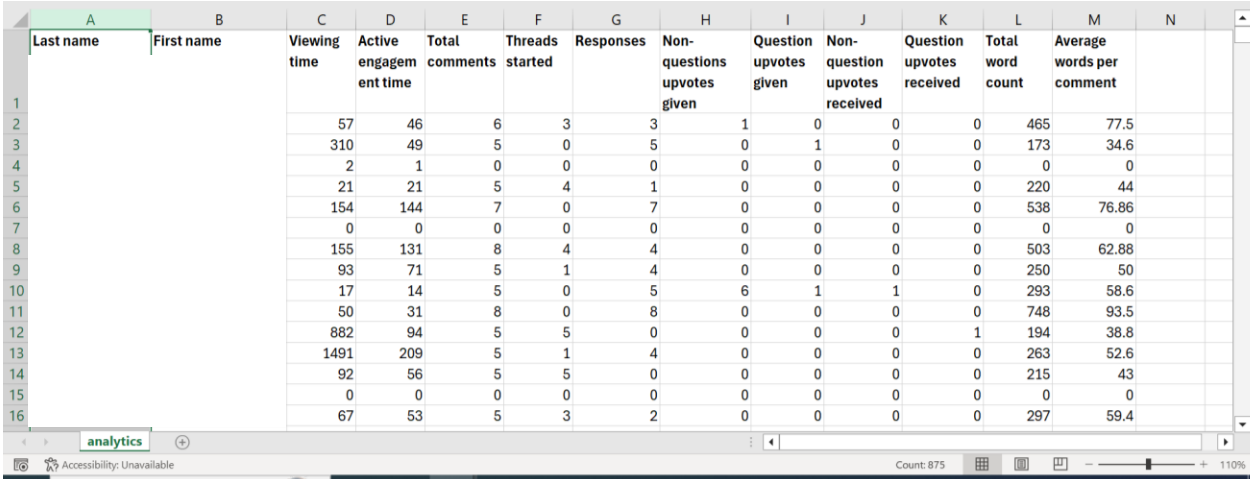
Analytics for Case 1a - Isabelle's research dilemma

Last name	First name	Viewing time	Active engagement time	Total comments	Threads started	Responses	Non-questions upvotes given	Question upvotes given	Non-question upvotes received	Question upvotes received	Total word count	Average words per comment
[redacted]	[redacted]	57 minutes	46 minutes (81%)	6	3	3	1	0	0	0	465	77.50
[redacted]	[redacted]	5 hours, 10 minutes	49 minutes (16%)	5	0	5	0	1	0	0	173	34.60
[redacted]	[redacted]	2 minutes	1 minute (54%)	0	0	0	0	0	0	0	0	0
[redacted]	[redacted]	21 minutes	21 minutes (99%)	5	4	1	0	0	0	0	220	44.00
[redacted]	[redacted]	2 hours, 34 minutes	2 hours, 24 minutes (94%)	7	0	7	0	0	0	0	538	76.86
[redacted]	[redacted]	0 minutes	0 minutes	0	0	0	0	0	0	0	0	0
[redacted]	[redacted]	2 hours, 35 minutes	2 hours, 11 minutes (84%)	8	4	4	0	0	0	0	503	62.88
[redacted]	[redacted]	1 hour, 33 minutes	1 hour, 11 minutes (76%)	5	1	4	0	0	0	0	250	50.00
[redacted]	[redacted]	17 minutes	14 minutes (80%)	5	0	5	6	1	1	0	293	58.60
[redacted]	[redacted]	50 minutes	31 minutes (62%)	8	0	8	0	0	0	0	748	93.50
[redacted]	[redacted]	14 hours, 42 minutes	1 hour, 34 minutes (11%)	5	5	0	0	0	0	1	194	38.80
[redacted]	[redacted]		3 hours, 29									

Figure 5: Perusall Engagement Data Overview with Analytics Breakdown (Perusall View)

Figure 5 above illustrates a Perusall view of the students' engagement analytics with each case study / learning material. It contains a breakdown of students' 'Viewing time' (the total amount of time the student had the content open), 'Active engagement time' (student move the curse or type something at least once every 2 minutes), 'Total comments' (the sum amount of the threads stated by the student and the number of responses they have posted), 'Threads started', 'Responses', 'Non-questions upvotes given', 'Question upvotes given', 'Non-question upvotes received', 'Question upvotes received', 'Total word count', 'Average words per comment'.

Similar to Figure 5, Figure 6 below shows a downloadable version of student engagement with one of the case studies (learning materials) as an example. The downloaded view mirrors the Perusall view in a different format.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Last name	First name	Viewing time	Active engagement time	Total comments	Threads started	Responses	Non-questions upvotes given	Question upvotes given	Non-question upvotes received	Question upvotes received	Total word count	Average words per comment	
2			57	46	6	3	3	1	0	0	0	465	77.5	
3			310	49	5	0	5	0	1	0	0	173	34.6	
4			2	1	0	0	0	0	0	0	0	0	0	
5			21	21	5	4	1	0	0	0	0	220	44	
6			154	144	7	0	7	0	0	0	0	538	76.86	
7			0	0	0	0	0	0	0	0	0	0	0	
8			155	131	8	4	4	0	0	0	0	503	62.88	
9			93	71	5	1	4	0	0	0	0	250	50	
10			17	14	5	0	5	6	1	1	0	293	58.6	
11			50	31	8	0	8	0	0	0	0	748	93.5	
12			882	94	5	5	0	0	0	0	1	194	38.8	
13			1491	209	5	1	4	0	0	0	0	263	52.6	
14			92	56	5	5	0	0	0	0	0	215	43	
15			0	0	0	0	0	0	0	0	0	0	0	
16			67	53	5	3	2	0	0	0	0	297	59.4	

Figure 6: Persuall Engagement Data with Each Case Study (Downloaded View)

It is worth noting the 'Threads started' column which refers to students building up their conversations and engage with each other. The community and communication are also reflected in the 'Responses' and upvoted questions. Another point to note is the total count of the rows indicates 875 participants/students, which as a matter of fact there were only 778 students on the module. It indicates some students have created duplicates accounts, mostly one with their personal account and the other with the University student account. Given students must use their University's account so to record their final grade, the row 7 and 15 with 0s seem to be duplicates where students did not engage.

In addition, please note the student's final grade is automatically graded by the AI algorithm of Perusall. The tutor could set up the parameter such as viewing time, active engagement time, comments, responses, threads, word count of the comments. It assists consistent marking, avoid manual marking and has largely reduced human variation in marking.

4.3 Student assessment performance comparison

Having viewed students' engagement with the learning materials and the impact of community learning tool in support of student engagement, this section outlines the student module performance related to their engagement with the weekly learning materials.

Table 2: Assessment performance breakdown and comparison

	2020-21 100% online delivery		**2021-22 Blended with Perusall**		**2022-23 On-campus with No Persuall**	
Total No.	**463**		**778**		**902**	
90+	**2**	**0%**	**477**	**61%**	**473**	**52%**
80-89%	9	2%	85	11%	77	9%
70-79%	32	7%	41	5%	102	11%
60-69%	106	23%	31	4%	66	7%

50-59%	144	31%	12	2%	28	3%
40-49%	73	16%	15	2%	21	2%
20-39%	48	10%	40	5%	19	2%
1-19%	16	3%	17	2%	2	0%
0	**33**	**7%**	**60**	**8%**	**114**	**13%**
Overall Failure Rate	36.7%		16.9%		17.2%	
Average Grade	**48.4%**		**79.4%**		**72.8%**	

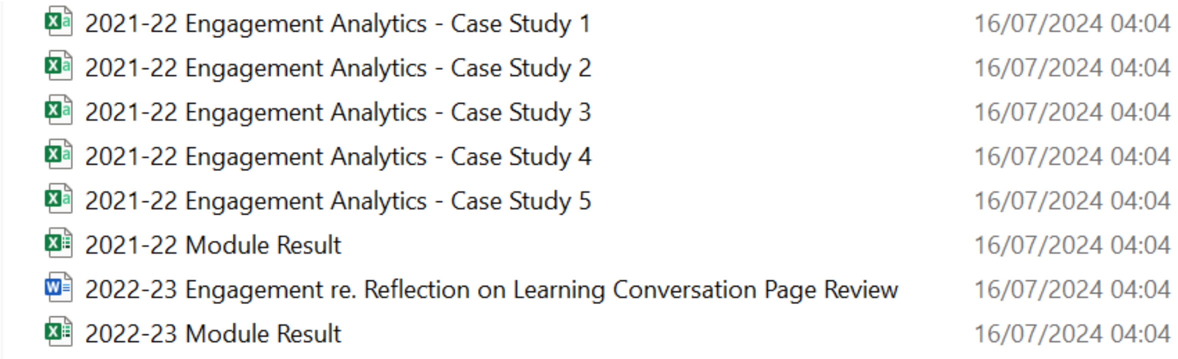
It is clear from Table 2 that where Persuall applied, the module assessment has achieved the highest average grade, lowest failure rate and most high achievers. It is vital to note similar outstanding module grade performance in both 2021-22 and 2022-23 when community learning took place with 2021-22 being AI mediated online learning community and 2022-23 being people led face-to-face community learning community in place. That suggests collaborative learning does improve students' performance as stated in hypothesis 1.

However, it is interesting to note that the growing number of students who achieved 0 normally refers to students who have not engaged with the assessment in the three years, 7% for 2020-21, 8% for 2021-22 and 13% for 2022-23. That entails that the more students registered on a module; the higher number of students are likely to fail to engage with the assessment. Though they might have engaged with the learning materials at the beginning or during their learning journey, this does not necessarily lead to the completion of the assessment. This addresses the challenges of managing large cohort of students.

what is more, Perusall the AI tool does seem to mediate positive association between collaborative learning and student engagement and performance. However, besides AI, human facilitated collaborative learning also have a positive associate with student engagement and performance.

4.4 Machine learning regression analysis of Perusall engagement

In order to assess the impact of AI, i.e. Perusall community reading tool on student engagement and performance, machine learning regression models were also applied to analyse the results. This helps to understand the relationship between student engagements and student performances (see Appendix 1). Student engagement data with Persuall for 2021-22 was stored in CSV files and student grades are stored in Excel files (see Figure 7) below.

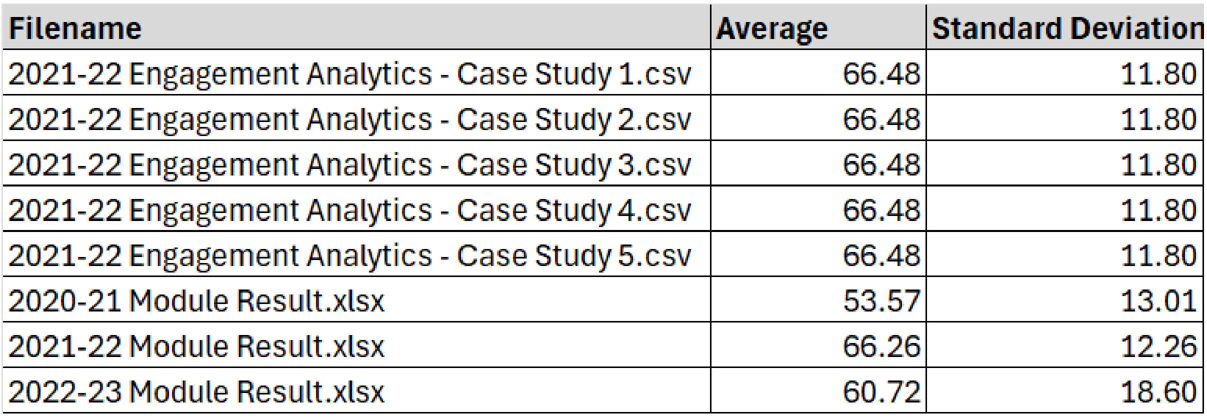


2021-22 Engagement Analytics - Case Study 1	16/07/2024 04:04
2021-22 Engagement Analytics - Case Study 2	16/07/2024 04:04
2021-22 Engagement Analytics - Case Study 3	16/07/2024 04:04
2021-22 Engagement Analytics - Case Study 4	16/07/2024 04:04
2021-22 Engagement Analytics - Case Study 5	16/07/2024 04:04
2021-22 Module Result	16/07/2024 04:04
2022-23 Engagement re. Reflection on Learning Conversation Page Review	16/07/2024 04:04
2022-23 Module Result	16/07/2024 04:04

Figure 7. Student Records datasets.

Table 3 below shows the statistical analysis of the student results at different academic years. It is worth noting, the below overall module result concurs with the element assessment result as outlined in Table 2 with 2021-22 being the best, followed by 2022-23 the second and 2020-21 the third.

Table 3. The Average and the Standard Deviation of Student Records dataset

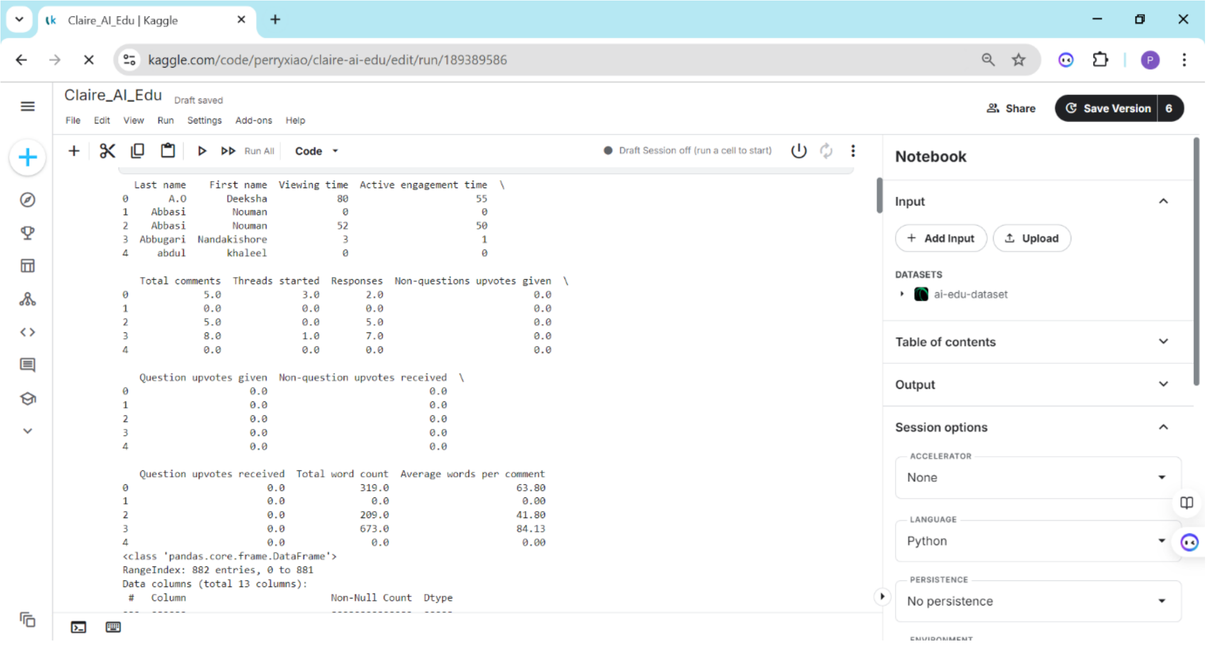


Filename	Average	Standard Deviation
2021-22 Engagement Analytics - Case Study 1.csv	66.48	11.80
2021-22 Engagement Analytics - Case Study 2.csv	66.48	11.80
2021-22 Engagement Analytics - Case Study 3.csv	66.48	11.80
2021-22 Engagement Analytics - Case Study 4.csv	66.48	11.80
2021-22 Engagement Analytics - Case Study 5.csv	66.48	11.80
2020-21 Module Result.xlsx	53.57	13.01
2021-22 Module Result.xlsx	66.26	12.26
2022-23 Module Result.xlsx	60.72	18.60

To analyse the data, a Machine Learning python code has been developed on Kaggle: https://www.kaggle.com/code/perryxiao/claire-ai-edu with the aim to use Machine Learning algorithms to analyse the datasets to see if there are correlations between the students' engagement with their attainment.

Firstly, all the five CSV files containing the student engagement information for five case studies were opened, then the total student engagement of all five case studies were analysed. Finally, the student's engagement data and the students' assessment grades were merged and stored in an Excel file.

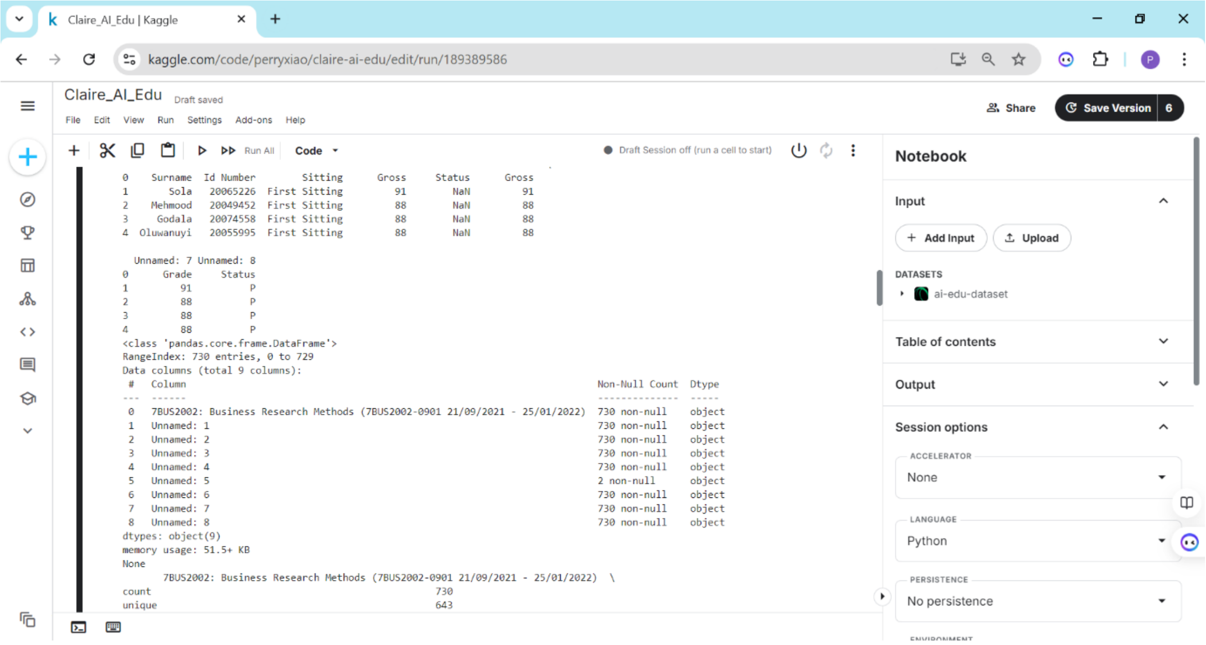
Figure 8 below shows a typical student engagement CSV file, 2021-22 Engagement Analytics - Case Study 1.csv file. This file contains the student's engagement information, such as Viewing time, Active engagement time, total comments, Threads started and so on.



```
   Last name   First name  Viewing time  Active engagement time  \
0        A.O      Deeksha            80                      55
1     Abbasi       Nouman             0                       0
2     Abbasi       Nouman            52                      50
3   Abbugari  Nandakishore            3                       1
4      abdul      khaleel             0                       0

   Total comments  Threads started  Responses  Non-questions upvotes given  \
0             5.0              3.0        2.0                          0.0
1             0.0              0.0        0.0                          0.0
2             5.0              0.0        5.0                          0.0
3             8.0              1.0        7.0                          0.0
4             0.0              0.0        0.0                          0.0

   Question upvotes given  Non-question upvotes received  \
0                     0.0                            0.0
1                     0.0                            0.0
2                     0.0                            0.0
3                     0.0                            0.0
4                     0.0                            0.0

   Question upvotes received  Total word count  Average words per comment
0                        0.0             319.0                      63.80
1                        0.0               0.0                       0.00
2                        0.0             209.0                      41.80
3                        0.0             673.0                      84.13
4                        0.0               0.0                       0.00
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 882 entries, 0 to 881
Data columns (total 13 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
```

Figure 8. The dataset from 2021-22 Engagement Analytics - Case Study 1.csv file.

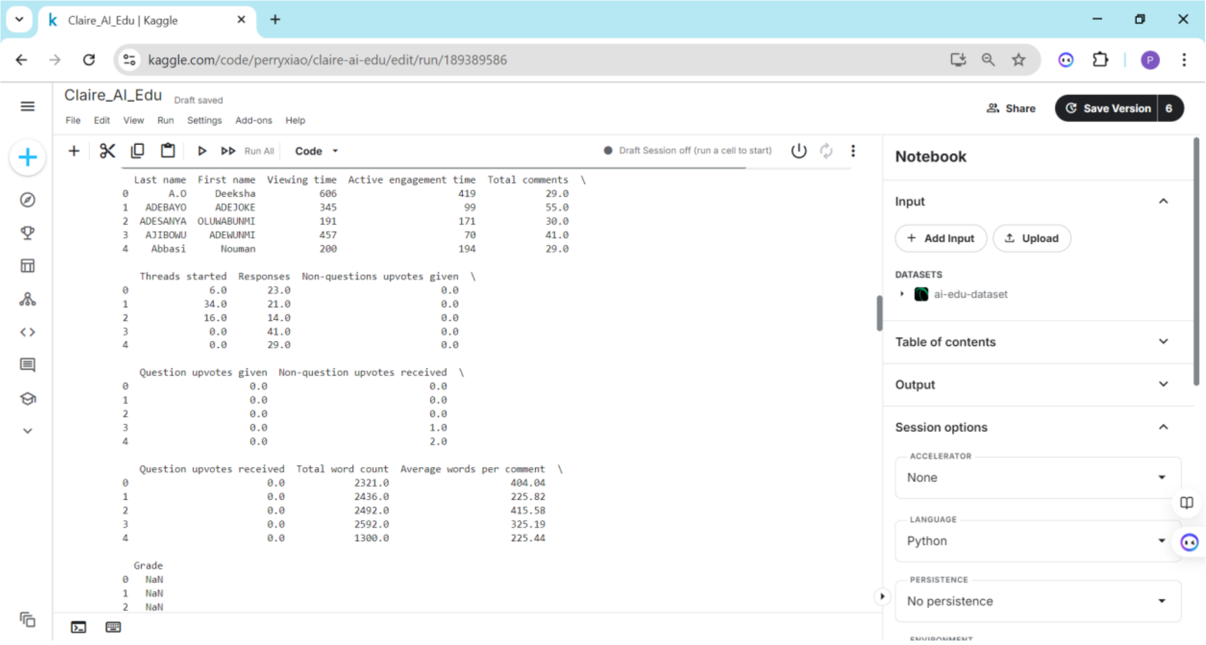
Figure 9 shows the dataset from the Excel file, 2021-22 Module Result.xlsx file. This file contains the students' grades.



```
0    Surname  Id Number       Sitting      Gross      Status      Gross
1       Sola   20065226  First Sitting        91         NaN         91
2    Mehmood   20049452  First Sitting        88         NaN         88
3     Godala   20074558  First Sitting        88         NaN         88
4  Oluwanuyi   20055995  First Sitting        88         NaN         88

  Unnamed: 7 Unnamed: 8
0      Grade     Status
1         91          P
2         88          P
3         88          P
4         88          P
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 730 entries, 0 to 729
Data columns (total 9 columns):
 #   Column                                                                Non-Null Count  Dtype
---  ------                                                                --------------  -----
 0   7BUS2002: Business Research Methods (7BUS2002-0901 21/09/2021 - 25/01/2022)  730 non-null    object
 1   Unnamed: 1                                                            730 non-null    object
 2   Unnamed: 2                                                            730 non-null    object
 3   Unnamed: 3                                                            730 non-null    object
 4   Unnamed: 4                                                            730 non-null    object
 5   Unnamed: 5                                                            2 non-null      object
 6   Unnamed: 6                                                            730 non-null    object
 7   Unnamed: 7                                                            730 non-null    object
 8   Unnamed: 8                                                            730 non-null    object
dtypes: object(9)
memory usage: 51.5+ KB
None
       7BUS2002: Business Research Methods (7BUS2002-0901 21/09/2021 - 25/01/2022)  \
count                                                730
unique                                               643
```

Figure 9. The dataset from 2021-22 Module Result.xlsx file.

Then, the two files of students' engagement and grade were merged according to the student's First name and Second name as shown in Figure 10 below.



```
  Last name  First name  Viewing time  Active engagement time  Total comments  \
0       A.O     Deeksha           606                     419            29.0
1   ADEBAYO     ADEJOKE           345                      99            55.0
2  ADESANYA  OLUWABUNMI           191                     171            30.0
3   AJIBOWU    ADEWUNMI           457                      70            41.0
4    Abbasi      Nouman           200                     194            29.0

   Threads started  Responses  Non-questions upvotes given  \
0              6.0       23.0                          0.0
1             34.0       21.0                          0.0
2             16.0       14.0                          0.0
3              0.0       41.0                          0.0
4              0.0       29.0                          0.0

   Question upvotes given  Non-question upvotes received  \
0                     0.0                            0.0
1                     0.0                            0.0
2                     0.0                            0.0
3                     0.0                            1.0
4                     0.0                            2.0

   Question upvotes received  Total word count  Average words per comment  \
0                        0.0            2321.0                     404.04
1                        0.0            2436.0                     225.82
2                        0.0            2492.0                     415.58
3                        0.0            2592.0                     325.19
4                        0.0            1300.0                     225.44

  Grade
0   NaN
1   NaN
2   NaN
```

ACCELERATOR: None

LANGUAGE: Python

PERSISTENCE: No persistence

Figure 10. The final merged dataset.

Multiple models were built to evaluate the student's performances. First, we have analysed each case study separately with different models, but the results are not very good, as shown in Table 4.

Table 4. The multiple model regression for 2021-22 Engagement Analytics - Case Study 1.csv file

File	Model	Train Mean Squared Error	Train R^2 Score	Test Mean Squared Error	Test R^2 Score
2021-22 Engagement Analytics - Case Study 1.csv	SVR	0.97	0.03	0.72	0.05
2021-22 Engagement Analytics - Case Study 1.csv	DecisionTreeRegressor	0.04	0.96	1.68	-1.20
2021-22 Engagement Analytics - Case Study 1.csv	RandomForestRegressor	0.19	0.81	0.77	-0.01
2021-22 Engagement Analytics - Case Study 1.csv	GradientBoostingRegressor	0.22	0.78	0.89	-0.17
2021-22 Engagement Analytics - Case Study 1.csv	KNeighborsRegressor	0.71	0.29	0.97	-0.27
2021-22 Engagement Analytics - Case Study 1.csv	MLPRegressor	655.48	-654.48	401.52	-524.87
2021-22 Engagement Analytics - Case Study 1.csv	XGBRegressor	0.04	0.96	1.06	-0.39

Then, we decided to use all five case studies together. Below Figure 11 – 21 show the overall results of different models. The results shows that Gradient Booting, Random Forest, and K Nearest Neighbour give the best results. The Decision Tree model and XGB model perform very well for training data, but not very well for the testing data.

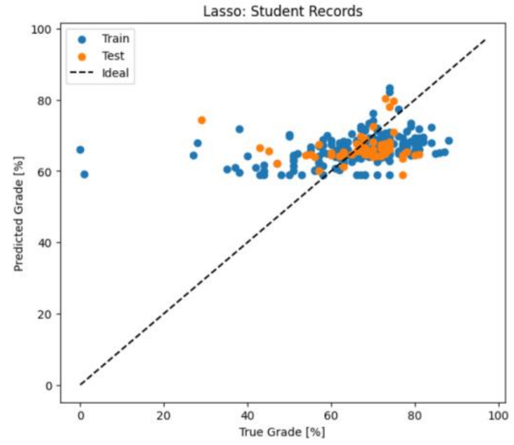


Figure 11. The Lasso model results.

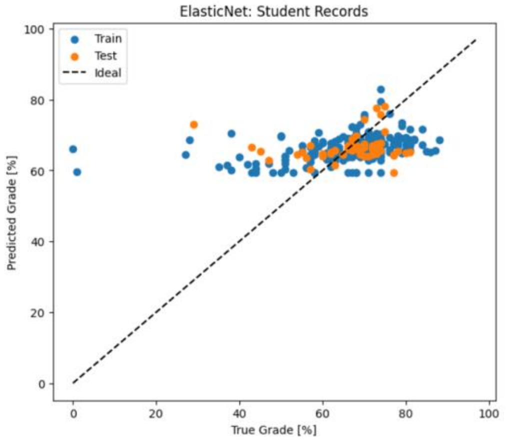


Figure 12. The ElasticNet model results.

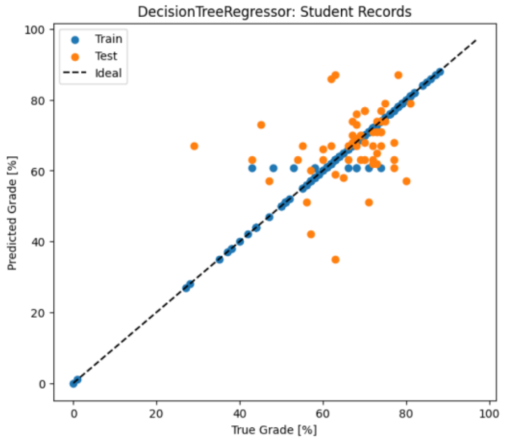


Figure 13. The Decision Tree model results.

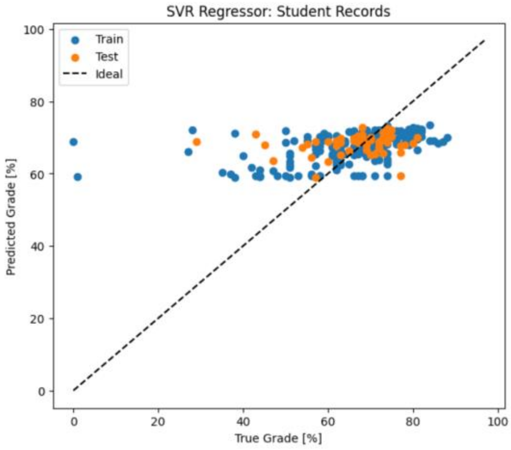


Figure 14. The SVM model results.

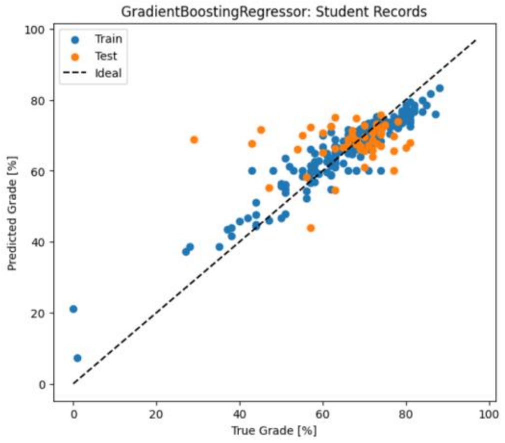


Figure 15. The Gradient Boosting model results.

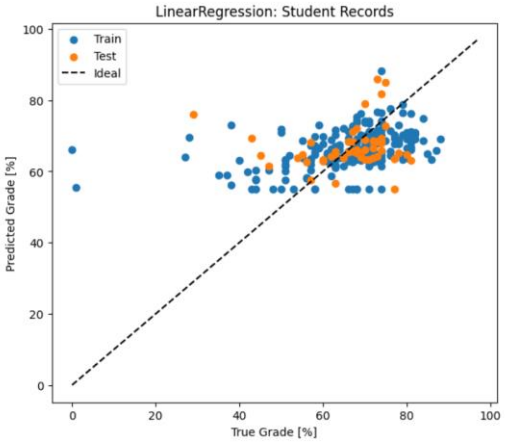


Figure 16. The Linear Regression model results.

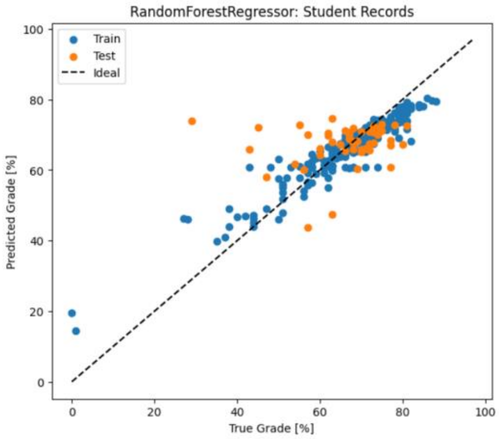


Figure 17. The Radom Forest model results.

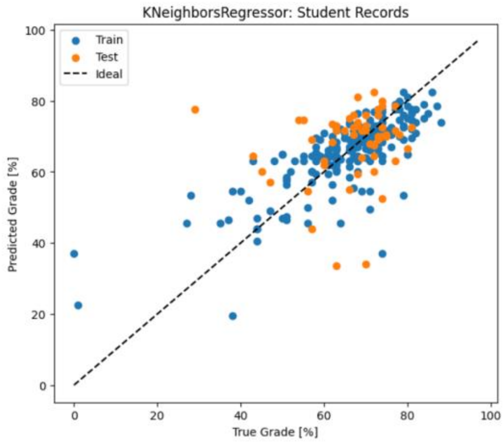


Figure 18. The K Nearest Neighbour model results.

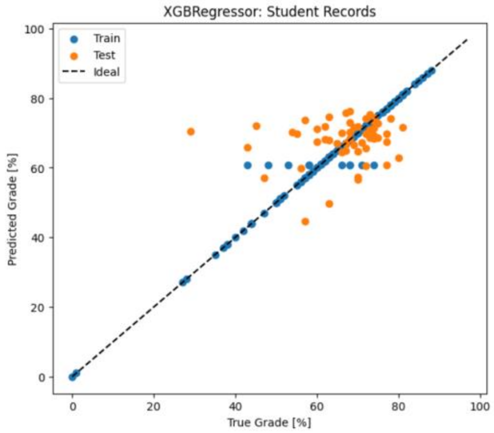


Figure 19. The XGB model results.

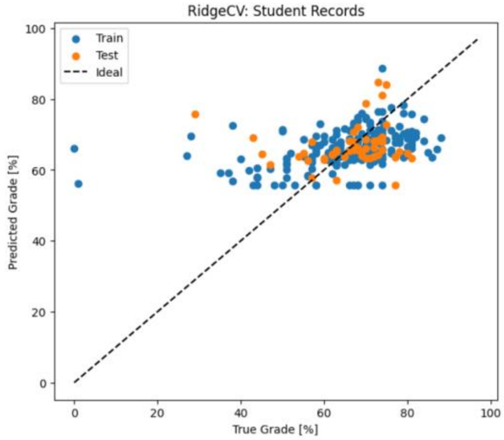


Figure 20. The RidgeCV model results.

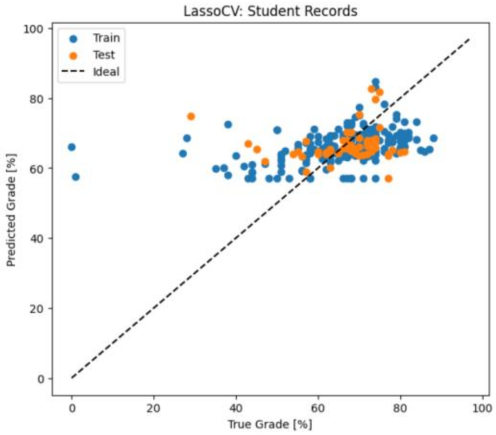


Figure 21. The LassoCV model results.

Amongst all the various models, it appears that Gradient Booting, Random Forest, and K Nearest Neighbour give the best results and seem to be the most effective models measuring the key features which affects students' final grade when they engage with Perusall. Figure 22 shows the features importance.

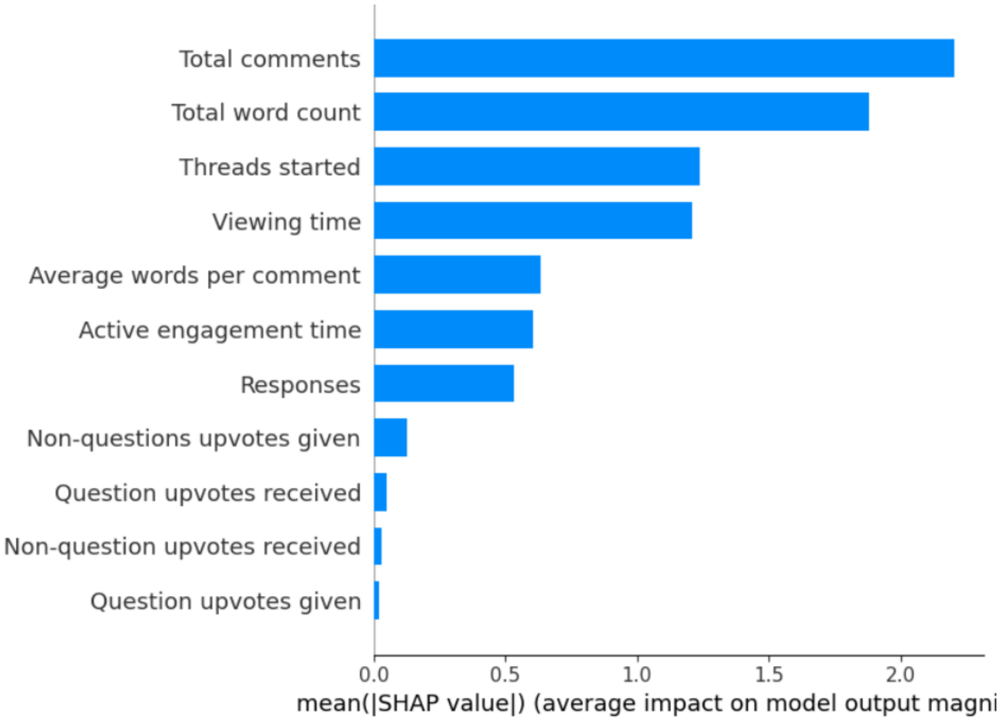


Figure 22. The feature importance for Random Forest model.

As it illustrates that the following features in Perusall seem to be the most important features determine students' final achievement:

1. Total Comments
2. Total Word Count
3. Thread Started
4. Viewing Time
5. Average words per comment
6. Active engagement time
7. Responses

On the contrary, the results that the following are the least important features.

1. Question upvotes given
2. Non-question upvotes received
3. Question upvotes received
4. Non-questions upvotes given

These results suggest that students who like to comments, like start new comments, and have more words in comments are likely to do well in the module, with total comments being the most important determining feature. The results also suggest that question upvotes given and received, as well as non-question upvotes given and received have little correlation with students' performance.

5.0 Discussion and conclusion

In conclusion, this research has assessed students' engagement and performance in a UK postgraduate course. Three academic years' (2020-21, 2021-22 and 2022-23) data on the student engagement with learning materials was presented and students' performance across the three years were compared. Machine learning regression analysis was conducted for students' engagement with Persuall, an online community social reading tool with AI facilitated grading functionality.

This research affirms that collaborative learning does improve student engagement and performance as demonstrated in both 2021-22 and 2022-23 academic year's student engagement and attainment data. In both academic year, collaborative learning took place in encouraging students to engage with the learning materials which also links to the assessment. In 2021-22, collaborative learning took place in the form of Persuall, an AI-enhanced social reading and community-based learning tool where students can view each other's comments, learn from one another, provoke and comment one another, despite the regression analysis reveals the final grade was mainly determined by the student's own comments and the word count of the comments. Whereas in 2022-23, students firstly engaged in group discussions known as learning conversations in their tutorials and then followed by individual reflections of the group learning conversations. Both online AI tool and face-to-face human oriented collaborative learning proven to be effective in engaging students in learning and striving for better academic performance.

However, it is worth noting that collaborative learning community does encourage students' engagement with learning but not necessarily assessment. This is apparent in human-led group learning environment where students are all encouraged to participate; while when it comes to the follow-up individual assessment, the completion rate dropped. Nonetheless, it seems linking learning materials to assessments does enhance student engagement.

In addition, this research also reveals the web-based AI community social learning tool, i.e. Perusall does seem to mediate the positive association between collaborative learning and student engagement and attainment as suggested in hypothesis 2. The students' academic performance over a period of three years shows that using AI aided software does improve students' academic performance. This is also supported by Machine Learning analysis of 2021-22 academic year's student engagement and performance data within Perusall which has discovered the correlations between student engagement and performance. That means, academics can use AI software to understand and track student engagement so to intervene and improve student learning and performance.

However, it appears that collaborative learning community does improve student performance regardless the community is mediated by AI or human. This occurs to the social constructivist learning theory with reference to the two key theoretical underpinning, i.e. Constructivist Learning Theory and Social Learning Theory. Meanwhile, the role of AI may also be argued by effective assessment design as demonstrated in the 2022-23 student engagement and performance data which is likely to suggest that people facilitated learning community seems to be equally effective.

However, it worth noting the analysis and conclusion of the research did not consider of the different modes of delivery for being online, blended and on-campus. In 2020-21, the year of pandemic when teaching was transformed to online, students were isolated from each other and the new online delivery might have impacted on their engagement and performance. Whereas in 2021-22, blended delivery was offered to students where face-to-face reminders might have also impacted students' learning and engagement. The adoption of Persuall enabled learners to an online social learning community where they can interact and learn from each other. In addition, students' level of engagement with the learning materials were simultaneously assessed by Perusall's generative AI grading algorithm, supplemented by automated just-in-time feedback for students which appear to be encouraging and effective. Coming to 2022-23, module delivery was converted back to face-to-face and learning communities were formed in tutorials where group conversations were facilitated by tutors, followed by individual assessment directly relate to the engaged discussions. Students' engagement with the learning conversation assessment and their academic attainment was evident though not as high as when Perusall AI-mediated tool was employed.

Theoretical contribution. This study has the following main contributions to the literature. This study applies constructivist learning and social learning theories in the AI education context, where the previous literature focuses on either single theory or applies to conventional education context. We have studied Perusall software which uses AI technology in collaborative learning to demonstrate that AI not only is an external social influence in student engagement and learning but also promote collaboratively learning constructively through promoting internal cognitive processes. Due to this internal and external influences on student learning, we argue that the two theories are complementary to each other and can be applied in improving student learning and performance. Our study provides empirical evidence to support this argument and prove that AI-adopted technologies such as Perusall help create more interactive learning environments which benefit student learning and performance as a result.

Practical contribution and implication. This study highlights the importance of setting up the AI tool, such as Perusall and its infrastructure for effective and consistent measurement and assessment to engage and motivate students. Appropriate use of AI mediation tools can effectively and efficiently engage students at scale; save time for manual grading which is more likely to have variations and discrepancies. In particular, universities are having an increase in student number in recent years. This will continue to last in the future. Using AI in student learning, it is easier for instructors to manage large classes by maintaining quality, consistency and student educational experience. This study also indicates that appropriate pedagogical curriculum design building in collaborative learning community can also facilitate social learning and constructive learning environment, which will in turn enhance student engagement and academic performance. This is applicable in not only conventional human-centred designed learning space but also digitalised technology-based learning environment. Whether the collaborative learning environment embeds AI technologies such as Perusall, if the environment embraces the collaborative learning design functions, student performance is almost the same as their performance under the environment with AI curriculum design. Nonetheless, AI makes the collaborative curriculum design more consistent and effective because AI replaces the conventional and manual design.

Future research may need to explore students' perspectives on the delivery mode and how it might have impacted on their engagement with social learning and community learning. Meanwhile, students' experience of the collaborative community learning via an online AI platform and group of real people might also worth exploring to determine the role of AI in community-based learning.

References

Adams, B. and Wilson, N.S. (2020). Building Community in Asynchronous Online Higher Education Courses Through Collaborative Annotation, *Journal of Educational Technology Systems*, 49(2), 250-261.

Ateh, C. M. (2024). Social platforms as Tools for Inclusive Pedagogy: Case of Perusall for Collaborative Reading. *Annals of Social Sciences & Management Studies*, 10(3).

Babu and Wooden. (2023). Managing the strategic transformation of higher education through artificial intelligence. *Administrative Sciences*, 13(9), 196.

Bada, S. O., & Olusegun, S. (2015). Constructivism learning theory: A paradigm for teaching and learning. *Journal of Research & Method in Education*, 5(6), 66-70.

Bandura, A., & Walters, R. H. (1977). Social learning theory (Vol. 1, pp. 141-154). Englewood Cliffs, NJ: Prentice hall.

Bharath, D.M.N. and Brownson, S. (2021). Perusall: Read, connect, discuss! *Journal of Public Affairs Education*, 27(3), 373-375.

Blayone, T. J., vanOostveen, R., Barber, W., DiGiuseppe, M., & Childs, E. (2017). Democratizing digital learning: theorizing the fully online learning community model. *International Journal of Educational Technology in Higher Education*, 14, 1-16.

Bolton, P. Lewis, J. and Gower, M. (2024) "International Students in UK Higher Education" *House of Commons Library* 20 September 2024

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. IEEE Access, 8, 75264-75278.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.

Draper, N. R., & Smith, H. (1998). Applied regression analysis.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 29(5), 1189-1232.

Gash, H. (2014). Constructing constructivism. *Constructivist foundations*, 9(3), 302-310.

Gligorea, I., Cioca, M., Oancea, R., Gorski, A. T., Gorski, H., & Tudorache, P. (2023). Adaptive learning using artificial intelligence in e-learning: a literature review. *Education Sciences*, 13(12), 1216.

Gordon, M. (2009). Toward a pragmatic discourse of constructivism: Reflections on lessons from practice. *Educational studies*, 45(1), 39-58.

HESA (2024) *Higher Education Student Statistics: UK 2022/23 released* 8 August 2024

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in human behaviour*, 19(3), 335-353.

Miller, K., Lukoff, B., King, G., and Mazur, E. (2018). Use of a Social Annotation Platform for Pre-Class Reading Assignments in a Flipped Introductory Physics Class, *Frontiers in Education*, 7 March.

Mukhalalati, B. A., & Taylor, A. (2019). Adult learning theories in context: a quick guide for healthcare professional educators. *Journal of medical education and curricular development*, 6, 2382120519840332.

Mvududu, N., & Thiel-Burgess, J. (2012). Constructivism in practice: The case for English language learners. *International Journal of Education*, 4(3), 108-118.

Perusall (2025) Perusall Available at: https://www.perusall.com/ Accessed on 20/02/2025

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106.

Saaida, M. B. (2023). AI-Driven transformations in higher education: Opportunities and challenges. *International Journal of Educational Research and Studies*, 5(1), 29-36.

Saunders, M.N.K., Lewis, P. and Thornhill, A. (2019) Research Methods for Business Students. 8th Edition, New York: Pearson

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Tran, V. D. (2013). Theoretical Perspectives Underlying the Application of Cooperative Learning in Classrooms. *International Journal of Higher Education*, 2(4), 101-115.

Trowler, V. (2010). Student engagement literature review. *The Higher Education Academy*, 11(1), 1-15

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.

Appendix 1. List of machine learning models used in this study

Decision Tree: Splits the data based on features to build a tree structure for prediction. It's easy to understand but prone to overfitting [Quinlan, J. R. (1986)].

ElasticNet: Combines the penalties of Lasso and Ridge regression, providing a balance between feature selection and regularization [Zou, H., & Hastie, T. (2005)].

Extreme Gradient Boosting (XGB): An efficient and scalable implementation of gradient boosting [Chen, T., & Guestrin, C. (2016)].

Gradient Boosting (GB): Builds an ensemble of weak learners in a sequential manner to minimize the loss function [Friedman, J. H. (2001)].

K-nearest neighbors (KNN): Predicts based on the k nearest data points in the feature space. Simple but sensitive to the choice of k [Cover, T., & Hart, P. (1967)].

Lasso (Least Absolute Shrinkage and Selection Operator): Helps in feature selection by shrinking the coefficients of less important features to zero. It can handle multicollinearity [Tibshirani, R. (1996)].

LassoCV: Lasso regression with cross-validation for hyperparameter tuning.

Linear Regression: Assumes a linear relationship between the input features and the target variable. Simple and interpretable [Draper, N. R., & Smith, H. (1998).].

Random Forest: An ensemble of decision trees, reduces overfitting and provides better generalization [Breiman, L. (2001).

RidgeCV: Ridge regression with cross-validation for hyperparameter tuning [Hoerl, A. E., & Kennard, R. W. (1970)].

Support Vector Machine (SVM): Finds the optimal hyperplane that separates the data with maximum margin. Can be used for regression by modifying the objective function [Cortes, C., & Vapnik, V. (1995)].

Responsible AI in Action: Assessing Variations in Perspectives Between C-Level Managers and AI Developers

Completed Research

Pouria Akbarighatar

Department of Information Systems, University of Agder

Ilias O. Pappas

Department of Computer Science, Norwegian University of Science and Technology

Department of Information Systems, University of Agder

Polyxeni Vassilakopoulou

Department of Information Systems, University of Agder

Abstract

Organizations worldwide embrace Artificial Intelligence (AI) initiatives across diverse fields, from healthcare to transportation. While AI may offer significant benefits, there are also concerns related to issues such as bias, privacy, transparency, reliability, and accountability. Responsible AI (RAI) frameworks and principles have emerged to address these issues. However, operationalizing RAI remains a challenge. In this paper, we report findings from a survey of C-level managers and AI developers from Europe and North America. We identify areas of consensus and divergence between these practitioner groups, revealing that organizations often struggle to implement RAI due to differing perspectives. By shedding light on these challenges, our research contributes to the literature on operationalizing RAI, offering three suggestions to bridge the gap between high-level guidelines and real-world practices, particularly in areas such as fairness, inclusiveness, benevolence, transparency, and intelligibility.

Keywords: Responsible AI, RAI principles, AI practices, C-level managers, AI developers.

1.0 Introduction

Organizations engage in Artificial Intelligence (AI) initiatives aiming to reap benefits. For instance, AI is used for optimizing crop yields, diagnosing and treating diseases, enhancing road safety, reducing financial fraud, and supporting children at risk (Amugongo et al., 2023; Ho et al., 2019; Stilgoe, 2018; Stohr et al., 2024; Van Esch et al., 2019; Wall, 2018). However, AI also poses significant challenges related to its alignment with human values and societal expectations (Fumagalli et al., 2022; Mateen, 2018; Rinta-Kahila et al., 2023). Responsible AI (RAI) aims to address these challenges ensuring that AI is introduced in organisations ethically, sustainably, and respectfully towards human values and society (Pappas et al., 2023). RAI has been the answer of policymakers and companies to the growing concerns of stakeholders including the public sector, big tech companies, and governments (Clarke, 2019, European Commission, 2019, Floridi et al., 2018, Microsoft, 2020).

Prior research has shown that there are significant differences in the perspectives on RAI between practitioners and the general public (Jakesch et al., 2022) and also, between practitioners and lawmakers (Khan et al., 2023). Specifically, these prior findings point to differences in how significant different RAI principles are for different groups highlighting the importance of paying attention to who gets to define and plan RAI initiatives. There is a growing body of research examining practitioners' awareness of RAI and their perspectives on the challenges associated with RAI operationalization (Pant et al., 2024, Stahl et al., 2022, Vakkuri et al., 2020). However, there has been limited focus on how different groups of practitioners within organizations, particularly those directly involved in the development and deployment of AI systems, assess the significance of RAI principles.

These practitioners, such as C-level managers and AI developers, are central to translating RAI principles into actionable strategies. Their involvement in planning and building AI initiatives makes them key stakeholders in ensuring AI systems are developed responsibly. As the first step in responsible AI use is ensuring it is developed following key principles, it becomes crucial to examine these groups. The divergence among practitioner groups regarding the prioritization of RAI principles could lead to internal tensions within organizations striving to implement RAI effectively. Conflicting priorities may hinder collaboration, create obstacles, and result in inconsistent strategies and practices, ultimately impacting RAI effectiveness. Our research aims to fill this gap by investigating the perspective of two different practitioner groups: C-level managers and AI developers. Specifically, we aim to answer the question: *What are the C-level managers and AI developer's perceptions of Responsible AI principles*?

This study contributes insights with both theoretical and practical implications. First, a nuanced understanding of the diverse viewpoints influencing real-world RAI application contributes to better conceptualizing the sociotechnical dynamics of RAI operationalization. Second, from a practice perspective, understanding these dynamics can inform the development of tailored training and resources that address specific concerns of different practitioner groups and can contribute to better intraorganizational communication and collaboration.

The paper is organized as follows. Section 2 includes the related literature, Section 3 the research method, and Section 4 the results and analysis. The discussion and implications are presented in Section 5, followed by limitations and research directions in Section 6. Finally, Section 7 presents the conclusions.

2.0 Related Literature

2.1. Responsible AI

Information Systems (IS) scholars have engaged in understanding how AI should be managed by problematizing the unintended consequences or the potentially unethical use of AI (Berente et al., 2021). In practice-oriented literature, numerous inquiries also have taken place. A notable report, published by the Organisation for Economic Co-operation and Development (OECD) in early 2019, stands out. This report synthesizes insights from over 70 documents that discuss ethical AI principles across various sectors. The documents originate from a range of sources, spanning industry players like Google, IBM, and Microsoft, governmental entities such as the Montreal Declaration and the Lords Select Committee, and academic institutions including the Future of Life Institute, IEEE, and AI4People. It comprises five complementary value-based principles: inclusive growth, fairness, transparency, security and safety, and accountability.

In a study that reviewed 84 documents on ethical AI, the prevalent themes were transparency, justice and fairness, non-maleficence, responsibility, and privacy, each appearing in over 50% of cases (Jobin et al., 2019). Moreover, a systematic analysis of the ethical technology literature by (Royakkers et al., 2018) underscored recurring themes encompassing privacy, security, autonomy, justice, human dignity, technology control, and power equilibrium. These themes collectively 'define' responsible AI as technology that is (a) beneficial and respectful towards individuals and the environment (beneficence); (b) resilient and secure (non-maleficence); (c) reflective of human values (autonomy); (d) fair (justice); and (e) transparent, accountable, and comprehensible (explicability).

When examining the European Commission’s High-Level Expert Group report's ethical principles, a consistent pattern emerges (European Commission, 2019). The report outlines four ethical principles, deeply rooted in fundamental rights, that must be upheld to ensure the trustworthy development, deployment, and use of AI systems. The first principle prioritizes respecting human autonomy and freedom (respect for human autonomy). The second emphasizes that systems should neither cause harm nor worsen existing issues for humans (prevention of harm). The third underscores the necessity for fairness throughout AI's lifecycle (fairness). Lastly, explicability proves essential for establishing and maintaining user trust in AI systems. This mandates transparent processes, clear communication of AI system capabilities and commitments, and comprehensible decisions for those directly and indirectly impacted. The absence of such information impedes the ability to challenge decisions effectively (explicability).

ISO 22989:2022 and ISO 24038 provide definitions and detailed explanations of the concept of trustworthiness, encompassing elements such as robustness, reliability, transparency, explainability, interpretability, accountability, safety, privacy, and fairness. All these concepts align with the categories established by OCED and the European Commission. For instance, transparency, interpretability, expandability, and accountability, share a common goal from varying perspectives, reinforcing each other. Collectively, these principles advance AI systems' understandability. Additionally, principles connected to avoiding harm and positive impacts, such as safety, privacy, benevolence, and non-maleficence, uphold AI's beneficence nature. Similarly, fairness and inclusiveness aim to eradicate disparities, ensure equal opportunities, and prevent marginalization.

Despite the advancements in conceptualising RAI, for organizations delivering AI services, achieving RAI begins with the understanding of key actors in organizations including managers and developers. The importance of practitioner's awareness and buy-in is frequently highlighted by researchers exploring the challenges organizations face in operationalizing RAI (Davison 2023; Amugongo et al., 2023; Barredo Arrieta et al., 2020; Kazim & Koshiyama, 2021). Our research focuses on exploring the perceptions of these actors.

Principle (P_i)	Description	Sources
P1: Benevolence and Nonmaleficence	Indicate that AI technology is designed to promote good and maximize benefits, all the while avoiding harm and minimizing risks.	(European Commission., 2019; Microsoft AI, 2022; Clarke., 2019; Floridi et al., 2018).
P2: Reliability and Safety	AI systems should aim to prevent failures and accidents ensuring intended performance.	(ISO:24028, 2020; Microsoft AI, 2022; Clarke., 2019)

P3: Privacy	Freedom from intrusion into an individual's private life or affairs when it happens due to improper or illegal collection and use of their data.	(ISO:24028, 2020; Microsoft AI., 2022).
P4: Security	Security refers to protecting data and controlling access based on authorization levels.	(ISO:24028, 2020; Microsoft AI., 2020).
P5: Accountability	Accountability refers to taking responsibility, providing justifications for actions, responding to inquiries, and being liable.	(ISO:24028., 2020; Microsoft AI., 2022; Clarke., 2019)
P6: Explainability	Explainability refers to providing comprehensive information about AI's inner workings.	(ISO:22989., 2020; Microsoft AI., 2022; Clarke., 2019)
P7: Intelligibility	Intelligibility refers to enabling humans who use or manage AI to understand the reasoning of an AI system.	(ISO:24028., 2020)
P8: Transparency	Transparency entails disclosing AI system details, like performance, limitations, components, measures, design goals, data sources, for a decision, prediction, or recommendation.	(IS:22989., 2020; Microsoft AI., 2022; Clarke., 2019; Floridi et al., 2018)
P9: Inclusiveness	Inclusiveness refers to involving diverse individuals and perspectives, regardless of their unique circumstances.	(OECD., 2018; Microsoft AI., 2022)
P10: Fairness	AI systems must be designed to ensure impartial treatment, and prevention of discriminatory outcomes.	(OECD., 2018; 2020; Microsoft AI., 2022; Clarke., 2019; Floridi et al., 2018)

Table 1. Responsible AI principles and their descriptions

2.2. Practitioners´ Perspectives on Responsible AI

Researchers have been examining the implementation of RAI in practice and practitioners’ perspectives. In a study published in 2024, Pant and colleagues reported findings from a survey of 100 AI practitioners aimed at understanding AI practitioners’ awareness of AI ethics and their *challenges* in incorporating ethics. They found that the majority had a reasonable familiarity with AI ethics especially those related to privacy protection and security (Pant et al., 2024). Vakkuri and colleagues conducted a multiple case study across AI development companies in Finland engaged in healthcare AI and found that developers are aware of the importance of ethics and exhibit concerns towards ethical issues (Vakkuri et al., 2019). Holstein and colleagues explored the perspectives of Machine Learning practitioners on their challenges and needs related to fair systems (Holstein et al., 2019). They found most were aware of the "Fairness" principle. Veale and colleagues conducted interviews with public-sector Machine Learning practitioners in high-stake contexts, to understand the challenges they face in understanding and following public values in their work with a focus on fairness and accountability (Veale et al., 2018). In a study performed by Akbarighatar et al. (2023), AI practitioners were interviewed regarding their RAI practices. They found that practitioners are generally aware of Responsible AI (RAI) and recognize the necessity of using RAI principles to guide AI development Moreover, organizations are introducing new roles, practices, and governance structures aligned with RAI principles to mitigate the risks associated with AI use and strengthen their competitive advantages. While prior research has explored AI practitioners' views related to RAI, limited attention has been paid to perspectives across subgroups. This is an important gap, as no studies have assessed how different practitioner communities evaluate key RAI principles and whether viewpoints diverge or coalesce. As prior research has shown significant differences in the perspectives on RAI between practitioners and the general public (Jakesch et al., 2022) and also, between practitioners and lawmakers (Khan et al., 2023) it is possible that there is variation among AI practitioner subgroups. Understanding variation or alignment among practitioner subgroups is critical. Divergent assessments of principles' significance may create tensions within organizations. Conflicting priorities may hinder collaboration, possibly resulting in misaligned strategies and practices. This study aims to address this gap by investigating potential disparities in practitioner views across two key subgroups: developers and C-level managers.

3.0 Research Method

We developed a survey to collect data from AI developers and C-level managers of organizations engaged in AI development. C-level managers play crucial roles in implementing and overseeing roles compared to others. For instance, C-level managers often provide the overall vision and strategic direction for AI initiatives, ensuring they align with the company's values and ethical standards. So, it is important to know the C-level manager’s understanding of AI ethics principles and challenges. AI developers are critical in resolving operational practices and challenges in implementing AI initiatives in organizations. These technical experts are directly involved in the development and deployment of AI systems, making their insights invaluable for understanding the practical aspects of RAI. Their expertise helps bridge the gap between theoretical principles and real-world application, ensuring that ethical guidelines are practically integrated into AI systems. Including C-level managers and AI developers in the research allows us to capture a comprehensive view of how RAI principles are operationalized and the technical hurdles that may arise during this process.

We identified potential respondents through professional LinkedIn groups like “Artificial Intelligence and Business Analytics” and the “Ethical AI Database” website, targeting responsible or ethical AI companies, as well as general AI companies. This strategy ensured a robust and representative sample for our study. Before responding to the questions, participants were given a description of these principles, which can be found in Table 1. The sample (N=130) comprised professionals with significant experience in AI roles. The majority held graduate degrees, with 33.3% possessing a PhD and 53.1% a master's-level qualification.

We developed a survey to collect data on practitioners’ perspectives on the different RAI principles drawing from literature as outlined in Table 1. To evaluate the content validity of these principles, we engaged a panel of six experts with substantial academic and practical experience in responsible AI. Four of these experts had over 15 years of industry experience in data science and AI, while the remaining two were senior academics specializing in Information Systems in organizations. We provided the experts with definitions of each principle and asked them to answer the survey questions. Additionally, we sought their recommendations for improving or refining questions. Their feedback led to minor modifications and clarifications in the definitions, reinforcing the content validity of our instrument.

We then distributed the revised survey instrument to four C-level technology managers. These managers were selected from companies that specialize in responsible AI solutions for AI firms and possess extensive experience in implementing RAI principles. Taking their input into account, we carefully revised and refined the definitions of the principles to ensure they were more concise and understandable. The questions are included in Appendix A. They are matched to the 10 principles we identified in the literature (Table 1) and relate to practitioners´ perspectives on importance, alignment, and operationalization. To gain insights into participant demographics, we collected also information such as their age, gender, professional background, years of experience, and organizational affiliations, allowing us to better understand the diverse perspectives within our participant pool. Table 2 provides a summary of descriptive statistics for our sample.

Working experience			**Working experience in AI**		
Category	C-level Managers N=60	AI Developers N=70	**Category**	C-level Managers N=60	AI Developers N=70
Fewer than one year	0%	0%	Fewer than one year	0%	0%
1-3 years	24%	14%	1-3 years	36%	32%
4-6 years	20%	21%	4-6 years	33%	31%
7-10 years	14%	23%	7-10 years	19%	26%
More than 10 years	52%	42%	More than 10 years	12%	11%
Familiarity with the concept of responsible AI			**Education level**		
Category	C-level Managers N=60	AI Developers N=70	**Category**	C-level Managers N=60	AI Developers N=70
Expert	29%	15%	PhD	28 %	31 %
Very familiar and involved actively	31%	35%	Master	54 %	52%
Very familiar and some involvement	31%	19%	Bachelor	15%	20%
Familiar but never involved	7%	23%	**Where**		
Hear about it	2%	8%	Europe	55%	52%
Never heard	0,0%	0%	North America	45%	48%

Table 2. Descriptive statistics of the sample

4.0 Analysis and Findings

To understand the differences in perceptions between the two groups of participants, we began by testing the normality of the ten principles. This would validate whether parametric analysis is acceptable or not. From the Kolmogorov-Smirnov and Shapiro-Wilk test results, the p values were significant which indicates the data set is not normally distributed and consequently the non-parametric analysis is in order. The results of the normality test are presented in Annex B. We performed nonparametric statistical analysis (De Winter et al., 2016; Myers & Sirois, 2014) to examine the differences and similarities between the perspectives of C-level managers and AI developers. Nonparametric methods are particularly appropriate for data that do not meet the assumptions required for parametric tests, such as normal distribution or homogeneity of variance. The Mann-Whitney U test was employed to statistically assess the differences in perceptions between C-level managers and AI developers regarding the importance, alignment, and operationalization of RAI principles.

Test Statistics	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Mann-Whitney U	1417.0	1726.0	1715.5	1981.5	1814.0	1771.0	1577.5	1452.0	1622.0	1574.0
Wilcoxon W	4118.0	4430.0	4416.5	4682.5	4515.0	4472.0	4278.5	4153.0	4323.0	4275.0
Z	-3.233	-1.432	-1.889	-.504	-1.286	-1.483	-2.403	-3.017	-2.182	-2.420
Asymp. Sig. (2-tailed)	**.001**	.082	.072	.615	.198	.138	**.016**	**.003**	**.029**	**.016**

Table 3, Mann-Whitney U Test on the perceived importance of RAI principles

The findings reveal significant differences between the two groups in certain principles in terms of how they assess their importance, alignment, and operationalization. Specifically, these differences are significant in principles of benevolence and non-maleficence, intelligibility, transparency, inclusiveness, and fairness (see Tables 3, 4, and 5). Conversely, no significant differences were observed between the groups in terms of reliability and safety, privacy, security, accountability, and explainability.

Test Statistics	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Mann-Whitney U	1417.0	1726.000	1715.500	1981.500	1814.000	1771.000	1577.500	1452.000	1622.000	1574.000
Wilcoxon W	4118.000	4430.000	4416.500	4682.500	4515.000	4472.000	4278.500	4153.000	4323.000	4275.000
Z	-3.233	-1.332	-1.889	-.504	-1.286	-1.483	-2.403	-3.017	-2.182	-2.420
Asymp. Sig. (2-tailed)	**.001**	.082	.072	.615	.198	.138	**.016**	**.003**	**.029**	**.016**

Table 4, Mann-Whitney U Test on perceived alignment of RAI principles

Test Statistics	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Mann-Whitney U	1417.000	1726.000	1715.500	1981.500	1814.000	1771.000	1577.500	1452.000	1622.000	1574.000
Wilcoxon W	4118.000	4430.000	4416.500	4682.500	4515.000	4472.000	4278.500	4153.000	4323.000	4275.000
Z	-3.233	-1.332	-1.889	-.504	-1.286	-1.483	-2.403	-3.017	-2.182	-2.420
Asymp. Sig. (2-tailed)	**.001**	.082	.072	.615	.198	.138	**.016**	**.003**	**.029**	**.016**

Table 5, Mann-Whitney U Test on perceived operatiolization of RAI principles

In the cases of benevolence and non-maleficence, intelligibility, transparency, inclusiveness, and fairness, the mean ranks indicate that C-level managers perceive the importance, alignment, and operationalization of these RAI principles more highly than AI developers do (Table 6). This suggests that C-level managers assign greater significance to these principles and their integration into AI practices compared to the AI developers.

Ranks					Ranks				
	Role	N	Mean Rank	Sum of Ranks		Role	N	Mean Rank	Sum of Ranks
Benevolence-Importance	C-level managers	57	77.14	4397.00	Transparency-Operationalization	C-level managers	57	76.68	4370.50
	AI developers	73	56.41	4118.00		AI developers	73	56.77	4144.50
	Total	130				Total	130		
Benevolence-Alignment	C-level managers	57	74.04	4220.00	Inclusiveness-Importance	C-level managers	57	73.54	4192.00
	AI developers	73	58.84	4295.00		AI developers	73	59.22	4323.00
	Total	130				Total	130		
Benevolence-Operationalization	C-level managers	57	76.11	4338.00	Inclusiveness-Alignment-	C-level managers	57	74.11	4224.50
	AI developers	73	57.22	4177.00		AI developers	73	58.77	4290.50
	Total	130				Total	130		
Intelligibility-Importance	C-level managers	57	74.32	4236.50	Inclusiveness-Operationalization	C-level managers	57	75.53	4305.00
	AI developers	73	58.61	4278.50		AI developers	73	57.67	4210.00
	Total	130				Total	130		
Intelligibility-Alignment	C-level managers	57	71.46	4073.50	Fairness-Importance	C-level managers	57	74.39	4240.00
	AI developers	73	60.84	4441.50		AI developers	73	58.56	4275.00
	Total	130				Total	130		
Intelligibility-Operationalization	C-level managers	57	70.04	3992.50	Fairness-Alignment	C-level managers	57	75.46	4301.50
	AI developers	73	61.95	4522.50		AI developers	73	57.72	4213.50
	Total	130				Total	130		
Transparency-Importance	C-level managers	57	76.53	4362.00	Fairness-Operationalization	C-level managers	57	74.95	4272.00
	AI developers	73	56.89	4153.00		AI developers	73	58.12	4243.00

	Total	130				Total	130		
Transparency-Alignment	C-level managers	57	77.20	4400.50					
	AI developers	73	56.36	4114.50					
	Total	130							

Table 6, Mann-Whitney U Test and Mean ranks on principles

5.0 Discussion and Implications

The findings of this study also reveal that C-level managers assign greater overall significance to RAI principles and their integration into AI practices compared to AI developers. This discrepancy suggests that strategic leaders may prioritize RAI principles more highly than the technical teams responsible for their implementation. Notably, this gap persists, particularly in comparison to well-established principles like privacy and security. This observation aligns with existing literature (Floridi et al., 2018; Haresamudram et al., 2023; Robert et al., 2020), which underscores the importance of improving our understanding of principles such as fairness, and inclusiveness among practitioners. Bridging the gap between high-level frameworks and technical practices is essential for ensuring that these principles are not merely discussed but effectively applied within organizations' practices.

Our findings indicate that there are principles in which there are no differences between C-level managers and AI developers including, privacy, security, reliability, accountability, and explainability. This alignment indicates that the strategic priorities of C-level managers, which are often discussed in guidelines are consistent with the practical implementation concerns of AI developers, which are typically represented in algorithmic codes or practices. Then we have principles where the perceptions among C-level managers and AI developers diverge. These are: fairness, inclusiveness, benevolence, transparency, and intelligibility in particular, C-level managers often consider certain principles to be more important than their AI developer counterparts.

This variance can create a disconnect in how principles are understood and applied within an organization. C-level managers, who are primarily focused on strategic oversight, seem to have distinct expectations for these principles compared to AI developers, who are more involved in the technical implementation of AI systems. This divergence can lead to inconsistencies in the application of guidelines and practices across different AI initiatives. These findings emphasize the necessity of bridging the gap between strategic and technical perspectives. Based on insights from nonparametric statistical analysis, we provide a set of suggestions to help bridge the gap between high-level RAI principles and their practical implementation.

First, it may be useful for organizations to prioritize securing the necessary resources (e.g., funding, infrastructure, tools) and developing essential competencies (e.g., awareness, skills, expertise, cross-functional collaboration) to ensure effective and cohesive implementation of RAI principles. This strategic investment in both tangible and intangible assets is crucial for successfully operationalizing RAI principles within organizational frameworks (Akbarighatar, 2024; Morley et al., 2020).

Second, as we move from principles that are viewed as high priority by practitioners to those of lower rank, the difficulty of their operationalization increases. This underscores the importance of contextualizing RAI principles within organizations (Dolata et al., 2022). For instance, a dedicated team can actively bridge the gap between C-level managers and AI developers by taking contextual factors into account during implementation. This team can ensure that high-level guidelines are effectively translated into practical applications, addressing specific organizational needs and constraints. Additionally, regular communication and feedback loops between these groups can help identify and mitigate challenges over time, fostering a more cohesive and effective implementation of RAI principles (Vassilakopoulou et al., 2022).

Finally, from a practical perspective, the findings of this study suggest directions for managers and AI developers to consider in developing and implementing RAI practices within their organizations. While all principles address important aspects of responsibility and are interconnected, it is crucial to understand how to initiate their implementation and plan effectively to achieve the desired outcomes. This, in turn, can inform their efforts to define the organization's RAI maturity and readiness for AI processes (Akbarighatar et al., 2023a). Organizations must develop a clear and actionable roadmap and implementation strategies for integrating RAI into their operations. Moreover, the comparative insights between C-level managers and AI experts can help organizations bridge the divide between high-level decision-making and on-the-ground practice. By addressing the differences in how these key stakeholder groups perceive and prioritize RAI principles, organizations can foster a more collaborative and coherent approach to RAI operationalization, leading to more consistent and effective RAI implementation.

6.0 Conclusion

The study uncovered misalignments between RAI perspectives of C-level managers and AI developers. These findings provide valuable theoretical and practical contributions. Firstly, by developing insights on the diverse viewpoints shaping real-world RAI implementation, it extends current knowledge on the sociotechnical dynamics of operationalizing RAI particularly in terms of principles. Secondly, in a practical sense, apprehending these viewpoints informs the development of training and resources addressing practitioner subgroups. It can also be used as a basis for strengthening intra-organizational collaboration and communication for RAI implementation.

While the study offers valuable insights, it is important to acknowledge and address inherent limitations. The study predominantly focuses on assessing perceptions around the importance, alignment, and operationalization of RAI principles, potentially neglecting other critical dimensions. Furthermore, the study currently only includes an analysis of quantitative data. Further work can involve incorporating data from interviews that can help in developing a better understanding of practitioners' perspectives.

References

Akbarighatar, P., Pappas, I. & Vassilakopoulou, P. 2023. Practices for Responsible AI: Findings from Interviews with Experts. *Proceedings of the Americas Conference on Information Systems (AMCIS).*

Akbarighatar, P., Pappas, I., & Vassilakopoulou, P. (2023a). A sociotechnical perspective for responsible AI maturity models: Findings from a mixed-method literature review. *International Journal of Information Management Data Insights*, 3(2), 100193.

Akbarighatar, P. (2024). Operationalizing responsible AI principles through responsible AI capabilities. *AI and Ethics,* pp. 1-15.

Amugongo, L. M., Kriebitz, A., Boch, A., & Lütge, C. (2023). Operationalising AI ethics through the agile software development lifecycle: A case study of AI-enabled mobile health applications. *AI and Ethics*, pp. 1-18.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. *Information Fusion*, 58, 82–115.

Berente, N., Gu, B., Recker, J. and Santhanam, R., 2021. Managing artificial intelligence. *MIS Quarterly*, *45*(3).

Clarke, R. 2019. Principles and business processes for responsible AI. *Computer Law & Security Review,* 35**,** 410-422.

Davison, R. (2023). The practitioner perspective. Information Systems Journal, 33(6), 1455-1458.

De Winter, J. C. F., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, 21(3), 273–290.

Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. Information Systems Journal, 32(4), 754–818.

European Commission. 2019. *Ethics guidelines for trustworthy AI* [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [Accessed 30 November 2022].

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U. & Rossi, F. 2018. AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines,* 28**,** 689-707.

Fumagalli, E., Rezaei, S., & Salomons, A. (2022). OK computer: Worker perceptions of algorithmic recruitment. *Research Policy,* 51(2), 104420.

Fornell, C. and Larcker, D.F., 1981. Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, *18*(1), pp.39-50.

Gefen, D., Rigdon, E.E. and Straub, D., 2011. Editor's comments: an update and extension to SEM guidelines for administrative and social science research. *MIS quarterly*, pp.iii-xiv.

Haresamudram, K., Larsson, S., & Heintz, F. (2023). Three Levels of AI Transparency. *Computer*, 56(2), 93–100.

Ho, C. W. L., Soon, D., Caals, K., & Kapur, J. (2019). Governance of automated image analysis and artificial intelligence analytics in healthcare. *Clinical Radiology,* 74(5), 329–337.

Holstein, K., Wortman Vaughan, J., Daumé Iii, H., Dudik, M. & Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? Proceedings of the 2019 CHI conference on human factors in computing systems, 2019. 1-16.

ISO:22989. (2022). ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. Retrieved from https://www.iso.org/standard/74296.html.

ISO:24028. (2020). ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence . Retrieved from https://www.iso.org/standard/77608.html).

Jakesch, M., Buçinca, Z., Amershi, S., & Olteanu, A. (2022). How Different Groups Prioritize Ethical Values for Responsible AI. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 310–323.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.

Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. Patterns, 2(9), 100314.

Khan, A. A., Akbar, M. A., Fahmideh, M., Liang, P., Waseem, M., Ahmad, A., Niazi, M. & Abrahamsson, P. 2023. AI ethics: an empirical study on the views of practitioners and lawmakers. *IEEE Transactions on Computational Social Systems,* 10**,** 2971-2984.

Mateen, H. (2018). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy: Cathy O'Neil. Broadway Books, 2016. 268 Pages. *Berkeley Journal of Employment and Labor Law*, 39(1), 285–292.

Merhi, M. I. (2023). An Assessment of the Barriers Impacting Responsible Artificial Intelligence. *Information Systems Frontiers*, 25(3), 1147–1160.

Microsoft, "Empowering responsible AI practices | Microsoft AI." Accessed: Mar. 26, 2024. [Online]. Available: https://www.microsoft.com/en-us/ai/responsible-ai.

Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices, *Sci. Eng. Ethics*, vol. 26, no. 4, pp. 2141–2168, Aug. (2020).

Myers, L., & Sirois, M. J. (2014). Spearman Correlation Coefficients, Differences between. *Wiley StatsRef: Statistics Reference Online*.

OECD. 2019. *Artificial Intelligence Principles* [Online]. Available: https://oecd.ai/en/ai-principles [Accessed 30 November 2022].

Pant, A., Hoda, R., Spiegler, S. V., Tantithamthavorn, C. & Turhan, B. 2024. Ethics in the age of AI: an analysis of ai practitioners' awareness and challenges. *ACM Transactions on Software Engineering and Methodology,* 33**,** 1-35.

Pappas, I. O., Mikalef, P., Dwivedi, Y. K., Jaccheri, L., & Krogstie, J. (2023). Responsible Digital Transformation for a Sustainable Society. *Information Systems Frontiers*, 25(3), 945–953.

Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., & Gregor, S. (2023). Managing unintended consequences of algorithmic decision-making: The case of Robodebt. *Journal of Information Technology Teaching Cases*, 204388692311655.

Robert, L. P., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). Designing fair AI for managing employees in organizations: A review, critique, and design agenda. *Human–Computer Interaction*, 35(5–6), 545–575.

Royakkers, L., Timmer, J., Kool, L. and Van Est, R., 2018. Societal and ethical issues of digitization. *Ethics and Information Technology*, *20*, pp.127-142.

Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K. & Jiya, T. 2022. Organisational responses to the ethical issues of artificial intelligence. *AI & Society,* 37**,** 23-37.

Stohr, A., Ollig, P., Keller, R., & Rieger, A. (2024). Generative mechanisms of AI implementation: A critical realist perspective on predictive maintenance. *Information and Organization*, 34(2), 100503.

Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. S*ocial Studies of Science*, 48(1), 25–56.

Vakkuri, V., Kemell, K.-K. & Abrahamsson, P. Implementing ethics in AI: initial results of an industrial multiple case study. *International Conference on Product-Focused Software Process Improvement, 2019*. Springer, 331-338.

Vakkuri, V., Kemell, K.-K., Kultanen, J. & Abrahamsson, P. 2020. The current state of industrial practice in artificial intelligence ethics. *Ieee Software,* 37**,** 50-57.

Van Esch, P., Black, J. S., & Ferolie, J. (2019). Marketing AI recruitment: The next phase in job application and selection. C*omputers in Human Behavior*, 90, 215–222.

Veale, M., Van Kleek, M. & Binns, R. 2018 Published. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making.

Vassilakopoulou, P., Parmiggiani, E., Shollo, A., & Grisot, M. (2022). Responsible AI: Concepts, critical perspectives and an Information Systems research agenda. *Scandinavian Journal of Information Systems*, 34(2).

Wall, L. D. (2018). Some financial regulatory implications of artificial intelligence. *Journal of Economics and Business*, 100, 55–63.

Appendix A: Survey

1. Age
2. Gender
3. What is your current job title?
4. What is your highest level of education completed?

   High School Diploma or Equivalent   Bachelor   Master   PhD

5. How many years of experience do you have?

   0-1   1-3   3-5   >5

6. In which country do you currently work?
7. Number of employees in your Organisation:

   1-10   10-100   100-250   250-500   >500

8. Industry

Finance   Healthcare   Energy   Manufacturing   Commerce   Entertainment   Professional Services   Public Services   Other: specify

9. How familiar are you with the concept of responsible AI (RAI)?

Never heard about it	Heard about it - slightly familiar	Familiar but never involved in RAI practices	Very Familiar and some involvement in RAI practices	Very familiar and involved in RAI practices actively.	Expert

10. When did your organization start implementing RAI practices?

Have not started and have no plans to start soon	Have not started yet but plan to start within the next 12 months	Less than a year ago	1-3	3-5	>5

The following questions relate to different tenets of Responsible AI. After reading the definition for each tenet, please respond to the associated questions.

1. For the AI-infused initiatives in our organization, we view benevolence and non-maleficence as an important consideration.
2. The AI-infused initiatives in our organization align with the benevolence and non-maleficence tenet.

3. Benevolence and non-maleficience checks for AI-infused initiatives are implemented in our organization.
4. For the AI-infused initiatives in our organization we view reliability and safety as an important consideration.
5. The AI-infused initiatives in our organization align with the reliability and safety tenet.
6. Implementing reliability and safety checks for AI-infused initiatives is important in our organization.
7. For the AI-infused initiatives in our organization we view privacy as an important consideration.
8. The AI-infused initiatives in our organization align with the privacy tenet.
9. Implementing privacy checks for AI-infused initiatives is important in our organization.
10. For the AI-infused initiatives in our organization we view security as an important consideration.
11. The AI-infused initiatives in our organization align with the security tenet.
12. Implementing security checks for AI-infused initiatives is important in our organization.
13. For the AI-infused initiatives in our organization we view accountability as an important consideration.
14. The AI-infused initiatives in our organization align with the accountability tenet.
15. Implementing accountability checks for AI-infused initiatives is important in our organization.
16. For the AI-infused initiatives in our organization we view explainability as an important consideration.
17. The AI-infused initiatives in our organization align with the explainability tenet.
18. Implementing explainability checks for AI-infused initiatives is important in our organization.
19. For the AI-infused initiatives in our organization we view intelligibility as an important consideration.
20. The AI-infused initiatives in our organization align with the intelligibility tenet.
21. Implementing intelligibility checks for AI-infused initiatives is important in our organization.
22. For the AI-infused initiatives in our organization we view transparency as an important consideration.
23. The AI-infused initiatives in our organization align with the transparency tenet.
24. Implementing transparency checks for AI-infused initiatives is important in our organization.

25. For the AI-infused initiatives in our organization we view inclusiveness as an important consideration.
26. The AI-infused initiatives in our organization align with the inclusiveness tenet.
27. Implementing inclusiveness checks for AI-infused initiatives is important in our organization.
28. For the AI-infused initiatives in our organization we view fairness as an important consideration.
29. The AI-infused initiatives in our organization align with the fairness tenet
30. Implementing fairness checks for AI-infused initiatives is important in our organization.

Appendix B: Normality test

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
P11	.257	130	<.001	.840	130	<.001
P12	.196	130	<.001	.870	130	<.001
P13	.203	130	<.001	.886	130	<.001
P21	.237	130	<.001	.802	130	<.001
P22	.258	130	<.001	.850	130	<.001
P23	.252	130	<.001	.819	130	<.001
P31	.286	130	<.001	.708	130	<.001
P32	.235	130	<.001	.797	130	<.001
P33	.249	130	<.001	.772	130	<.001
P41	.283	130	<.001	.735	130	<.001
P42	.248	130	<.001	.815	130	<.001
P43	.256	130	<.001	.813	130	<.001
P51	.172	130	<.001	.874	130	<.001
P52	.183	130	<.001	.907	130	<.001
P53	.197	130	<.001	.903	130	<.001
P61	.193	130	<.001	.896	130	<.001
P62	.165	130	<.001	.923	130	<.001
P63	.190	130	<.001	.918	130	<.001
P71	.159	130	<.001	.931	130	<.001
P72	.170	130	<.001	.930	130	<.001
P73	.160	130	<.001	.929	130	<.001
P81	.180	130	<.001	.884	130	<.001
P82	.151	130	<.001	.916	130	<.001
P83	.176	130	<.001	.909	130	<.001
P91	.148	130	<.001	.906	130	<.001
P92	.144	130	<.001	.919	130	<.001
P93	.154	130	<.001	.916	130	<.001
P101	.188	130	<.001	.886	130	<.001
P102	.142	130	<.001	.915	130	<.001
P103	.164	130	<.001	.906	130	<.001

Pij refers to a value indexed by i and j, where: i = 1, 2, … 10 represents the Responsible AI (RAI) principles; and j = 1 indicates the importance of the perceived principles; j = 2 indicates perceived alignment; j = 3 indicates perceived operationalization.

Inclusivity and Diversity in Conversational AI Development: Challenges and Opportunities for Misinformation

Boineelo R Nthubu
Lancaster University, b.nthubu1@lancaster.ac.uk

Niki Panteli
Lancaster University, n.panteli1@lancaster.ac.uk

Research In progress

Abstract

Conversational AI development lacks inclusivity and diversity, which contributes biases that have the potential to create and spread misinformation. This study explores how inclusivity and diversity in Conversational AI development can be fostered to restrict misinformation. The study follows a qualitative methodology including semi-structured interviews and focused groups with 30 conversational AI users and developers. This research in progress paper is expected to make a theoretical contribution to research on misinformation and the role of inclusive and diverse conversational AI. It contributes to the literature on responsible AI, and the literature on chatbots and digital assistants use within workplaces. The study is also expected to have practical implications that will benefit users, organisations and society through actionable recommendations to foster conversational AI development with characteristics of inclusivity and diversity.

Keywords: Conversational AI, Misinformation, AI developers, Inclusivity, Diversity.

1.0 Introduction

The rise of conversational AI, also known as chatbots and digital assistants, presents increased and new misinformation challenges as well as opportunities for mitigation strategies (Xu et al., 2023; Hajli, et al., 2022; Gunson et al., 2021). Users interact with conversational AI due to their informational needs (Verne et al., 2022; Jackson and Panteli, 2024). However, information produced by conversational AI is not always accurate leading to potential misinformation (ibid.). The misinformation has an impact on organisations, government service, and the society at large. Cases have been presented where conversational AI used to offer critical citizen services give incorrect responses (Verne et al., 2022). These incorrect responses can limit access of essential government services by citizens that really need them such as welfare benefits. There are also concerning cases in mental health, where a man committed suicide at the advice of a chatbot (Atillah, 2023), and recently a young boy (14), encouraged by conversational AI, similarly committed suicide (Duffy, 2024). Such cases highlight that information provided by conversational AI could be harmful. At a time where people are seeking advice from an AI chatbot instead of a healthcare professional, and with the instant accessibility to the information conversational AI can provide, there is a growing pressure for misinformation mitigation strategies.

The position is taken that AI development lacks diversity and inclusivity, posing problems for misinformation. AI development teams' homogeneity (western and male), with less women, the global south, and other underrepresented groups perspectives, can unintentionally inject biases into the development of AI systems (Shams et al., 2023). Additionally, communities such as women, the LGBTQ+ community, senior citizens, Africans, and disabled persons are often overlooked during AI development (Fosch-Villaronga and Poulsen, 2022) further perpetuating biases in AI systems. The information generated by biased AI algorithms may reflect the biases and perspectives of a narrow demographic, topic, or community, leading to biased and inaccurate information that can cause societal harm and perpetuate stereotypes and marginalisation of underrepresented groups.

To mitigate biases in conversational AI algorithms and the risks of misinformation, increasing diversity and inclusivity in AI development is essential (Shams et al., 2023; Bragazzi et al., 2023). This paper aims to examine the role of Conversational AI developers not just in managing but also, more importantly, limiting and even restricting misinformation. Following these, the research questions guiding the study are:

How does conversational AI development contribute to misinformation?
What is the role of developers in tackling misinformation?
How can diversity and inclusivity be fostered in developing conversational AI systems to avoid misinformation?

2.0 Literature Review

2.1 Biases Presenting Misinformation Challenges in Conversational AI

The prevalence of misinformation in AI algorithms is influenced by various biases inherent but not limited to the algorithmic development. These biases, including, data bias, algorithmic bias, and gender bias contribute to AI's susceptibility to producing or amplifying misinformation. Data bias, one of the most common types of bias means the data used to train the algorithm is not representative of the general population (Nazer et al., 2023). If a dataset overrepresents certain groups or viewpoints, the model will reflect these biases, leading to skewed or misrepresentative outcomes. The responses from these outcomes can be easily shared, spreading misinformation. For example, a conversational AI may provide biased responses if their data sources are not representative of specific communities or topic (Bragazzi et al., 2023).

Another bias encountered in development is gender bias, and this is primarily caused by algorithmic design and training datasets (Hall and Ellis, 2023). Gender bias can contribute towards misinformation by delivering responses that reflect gendered assumptions rather than factual accuracy.

Algorithm bias may lead to AI systems giving responses that may indirectly favour particular types of information, viewpoints or groups. Biases related to problem formulation may lead to wrong outcomes if users formulate question in a different way from how the algorithm questions were formulated (Srinivasan and Chander, 2021).

2.2 Inclusivity and Diversity Can Mitigate Misinformation

More frequently studies are finding that the lack of diversity in development teams is a contributing factor to many of the biases in AI algorithms. For example, Nadeem et al., (2020) found that the lack of diversity in developers was a major contributing factor to gender bias in AI algorithms. This is also reported in a recent review (Hall and Ellis, 2023). Ultimately, AI systems reflect the biases of their developers, especially if development teams lack diversity in e.g. gender, ethnicity, and cultural background (Adams and Khomh, 2020). The homogeneity of development teams can limit a team's ability to anticipate and address the needs of diverse populations, inadvertently embedding biases that may contribute to misinformation.

Although ruling out such biases is not straightforward due to the different bias types, to mitigate as many biases as possible, teams and organisations involved in AI development should be made more diverse (Adams and Khomh, 2020; Hall and Ellis, 2023; Nadeem et al., 2020). Inclusive AI development leads to more robust models capable of giving less biased and inaccurate responses. When diverse perspectives are included in the design process, AI systems would be better at restricting misinformation, as developers from various backgrounds can identify unique risks and biases. Following this understanding, it is important to explore further how inclusivity and diversity can be fostered in conversational AI development to restrict misinformation. This investigation aligns with Abdelhalim et al., (2024)'s call to further explore inclusivity and diversity in AI technology in order to foster a responsible integration of technology in organisations. Similarly, the OECD's responsible AI principles, a set of guidelines for AI actors, call for AI actors to address misinformation amplified by AI throughout the AI system lifecycle, including development (OECD, 2024). Ultimately, this study not only offers a novel perspective to restricting misinformation in conversational AI development, but also has potential for fostering meaningful change towards responsible AI.

3.0 Methodology

We plan to employ qualitative methods, including semi-structured interviews and focus groups with both AI developers and users of conversational AI anticipating between 30 participants. Questions to be asked include:

Please describe a specific incident you encountered or heard about where conversational AI provided inaccurate or misleading information?

Describe an incident you encountered or heard about where conversational AI responses seemed biased or less inclusive of specific communities, social issues or topic?

What is your view about biases that may be inherent features of conversational AI?

What improvements could conversational AI developers make to better provide accurate information on diverse communities and topics to prevent the spread of misinformation?

Following the interviews and focused groups, we plan to analyse the data following (Gioia et al., 2013).

4.0 Tentative Implications

The study is still at an early stage but we anticipate to be able to showcase some of the empirical findings at the conference. The study is expected to make a theoretical contribution to research on misinformation and the role of inclusive and diverse conversational AI. It is anticipated that this will contribute to the literature on responsible AI as well as to the literature on the use of chatbots and digital assistants within workplaces. The study is also expected to have practical implications that will benefit users, organisations and society through actionable recommendations on how to develop conversational AI with characteristics of inclusivity and diversity.

Références

Abdelhalim, E., Anazodo, K. S., Gali, N., & Robson, K. (2024). A framework of diversity, equity, and inclusion safeguards for chatbots. *Business Horizons*.

Adams, B., & Khomh, F. (2020). The diversity crisis of software engineering for artificial intelligence. *IEEE Software, 37*(5), 104-108.

Atillah, E.I (2023). Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change. Available at [https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-]. Accessed on [31 October 2024].

Bragazzi, N. L., Crapanzano, A., Converti, M., Zerbetto, R., & Khamisy-Farah, R. (2023). Queering Artificial Intelligence: The Impact of Generative Conversational AI on the 2SLGBTQIAP Community. A Scoping Review. *A Scoping Review (August 22, 2023)*.

Daffy, C. (2024) This mom believes Character.Ai is responsible for her son’s suicide | CNN Business. Available at [https://edition.cnn.com/2024/10/30/tech/teen-suicide-character-ai-lawsuit/index.html?Date=20241030&Profile=CNN,CNN+International,cnn&u

tm_content=1730321222&utm_medium=social&utm_source=facebook,linkedin,threads&fbclid=IwY2xjawGShk5leHRuA2FlbQIxMQABHSdr169RpanyReW1uik3gOwDx3T9IbcAe6hndwRvLs5VHodz5OejpUn7cQ_aem_rltI8mujbKSqQZg3nBjLVQ&sfnsn=scwspwa]. Accessed on [04 November 2024].

Fosch-Villaronga, E., & Poulsen, A. (2022). Diversity and inclusion in artificial intelligence. *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice*, 109-134.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational research methods*, *16*(1), 15-31.

Gunson, N., Sieińska, W., Yu, Y., Garcia, D. H., Part, J. L., Dondrup, C., & Lemon, O. (2021, September). Coronabot: A conversational ai system for tackling misinformation. In *Proceedings of the Conference on Information Technology for Social Good* (pp. 265-270).

Hajli, N., Saeed, U., Tajvidi, M., & Shirazi, F. (2022). Social bots and the spread of disinformation in social media: the challenges of artificial intelligence. *British Journal of Management*, *33*(3), 1238-1253.

Hall, P., & Ellis, D. (2023). A systematic review of socio-technical gender bias in AI algorithms. *Online Information Review*, *47*(7), 1264-1279.

Jackson, S., & Panteli, N. (2024). AI-Based Digital Assistants in the Workplace: An Idiomatic Analysis. *Communications of the Association for Information Systems*, *55*(1), 22.

Nadeem, A., Abedin., & Marjanovic, O. (2020) Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies. *ACIS 2020 Proceedings*. 27.https://aisel.aisnet.org/acis2020/27

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, *2*(6), e0000278.

OECD. (2024) OECD AI Principles. Available at [https://oecd.ai/en/ai-principles]. Accessed on [05 November 2024].

Shams, R. A., Zowghi, D., & Bano, M. (2023). AI and the quest for diversity and inclusion: A systematic literature review. *AI and Ethics*, 1-28

Srinivasan & Chander, 2021. Biases in AI systems. *Communications of the ACM*, *64*(8), 44-49.

Verne, G, B, Steinstø, T, Simonsen, L, & Bratteteig, T (2022) How Can I Help You? A chatbot's answers to citizens' information needs. *Scandinavian Journal of* Information Systems: Vol. 34: Iss. 2, Article 7. Available at: https://aisel.aisnet.org/sjis/vol34/iss2/7

Xu, Fan, S., & Kankanhalli, M. (2023). Combating Misinformation in the Era of Generative AI Models. In ACM ICM, 9291–9298.

Navigating Digital Interaction and Privacy: A Comparative Analysis of Social Media Engagement Among Consumers in Thailand and the UK

Anabel Gutierrez[1], Khanyapuss Punjaisri[2], Simon O'Leary[3]
[1]Royal Holloway University of London, [2]NIDA Business School, National Institute of Development Administration, Thailand, [3]Canterbury Christ Church University

Abstract

Consumers engage in social media activities, such as liking, commenting, sharing, and creating content, often revealing their identities, even if it is an online persona. The rise of social commerce allows users to purchase brands directly through social media, but it raises privacy concerns. Brands can personalise messages to drive engagement, yet data monetisation has created tensions among technology, privacy, consumers, regulators, and firms. This paper explores the links between UK and Thai consumers' social media activities, their privacy concerns, and perceptions of brand intrusiveness in relation to purchase intentions via social commerce. Results indicate limited differences in engagement levels between Thai and UK consumers, with active participation reducing feelings of intrusiveness and privacy concerns. However, some differences exist among those users who do have privacy concerns affecting their purchase intention, an interesting finding given that regulations surrounding online privacy are emerging at varying rates globally.

Keywords: Privacy Concerns, Intrusiveness, Social Media Engagement

1.0 Introduction

Over recent decades, the evolution of Social Network Sites (SNSs) has created a convergence of social commerce, e-commerce, and social media. This convergence holds significant potential for fostering enduring brand-consumer relationships through multiple social media interactions, promoting positive attitudes towards social media advertising and increasing purchase behaviours (Gutierrez et al., 2023). Social media content, acting as a narrative tool, influences consumer perceptions and attitudes toward reality, conveying stories and values to broad audiences (Tsay-Vogel, Shanahan and Signorielli, 2018). By leveraging tracking systems, brands are able to personalise messages across platforms, enhancing the relevance of and interactivity with social media advertising. This in turn motivates consumers to engage with and purchase from brands (Alalwan, 2018). However, the use of digital technologies by firms for data monetisation and sharing also creates tensions between the technology and related concerns about consumer privacy and regulators (Quach, Thaichon, Martin, Weaven, and Palmatier, 2022). Moreover, individuals' concerns about personal data collection and its use are diverse, being shaped by multiple factors such as culture, profession, and legislation, as well as personal attributes such as personality and past experience (Mutambik, Lee, Almuqrin, Zhang, and Homadi, 2023). This research explores the impact of privacy concerns and intrusiveness on purchase intention in social commerce among UK and Thai consumers, based on their social media engagement with brands.

2. Literature Review

2.1 Social Media Activities Engagement

According to Hollebeek, Glynn, and Brodie (2014), consumer-brand engagement is a consumer's positive brand-related cognitive, emotional, and behavioural activity during, or related to, consumer/brand interactions. This definition resonates with the social media activities (SMAs) of consumers engaging with brands on social media, and is captured in consumers' creation and contribution of SMAs beyond merely consuming social media content (Schivinski et al., 2016). Along with the rise of social media marketing and interactive channels, consumers are exposed to brand-related content, such as social media advertising, when they actively participate in social media activities. This increased engagement in social media activities has led to a deeper connection between consumers and brands (Gutierrez et al., 2023) transforming the nature of engagement, increasing also the amount of data available. The increased consumer-brand engagement in social media enables brands to analyse user-generated content that includes images, text, and audio to optimise conversion rates, measuring content efficiency or creating personalised content (Nazir et al., 2023), enhancing the opportunities to build long term relationships between brand and consumers.

Despite the varied levels of social media penetration and usage across different countries, it is undeniable that multiple consumers worldwide have engaged with brands via different social media activities. Consumers in both the UK and Thailand are engaged with social commerce, although Thailand has one of the highest shares of social commerce buyers worldwide at 94%, compared to 56% in the UK (Statista, 2024a). Privacy concerns on social media platforms have been a topic worldwide, and with the expectation of a rise in privacy incidents and data breaches, regulatory bodies worldwide are implementing stricter measures, as seen in the European Union's

General Data Protection Regulation. Worldwide data protection laws, including the GDPR, Australian Privacy Act, and CPRA, aim to restrict personal data acquisition, utilisation, retention, and transmission. While regulations on data privacy vary across nations, the GDPR has emerged as a universal benchmark (Quach et al., 2022). However, users' awareness of their country's privacy laws also varies, with the UK reporting 57% of users' awareness while Thailand is not listed (Statista. 2024b).

2.2 Purchase Intention and Social Commerce

The emergence of social commerce has enabled consumers to buy goods or services from a brand within a social media platform, and it eliminates the need to navigate away from the platform. While not overtly advertising, clickable advertisements in a social media feed, influencer posts with direct purchase links, and multimedia content can lead directly to a dedicated e-commerce platform (Mutambik et al., 2023) impacting purchase intention, however, little is known if purchase intention differs between countries hence, we propose:

H1: Social media activities influence purchase intention in both countries, the UK and Thailand.

When consumers actively engage with social media, they encounter brand messages in different forms, including advertisements. These brand messages aim to shape consumers' attitudes towards the brand, ultimately intending for consumer conversion, and subsequently impacting their intention to make a purchase (McClure & Seock, 2020). Despite initial concerns about privacy, the more consumers engage in social media activities, the more the social contacts and friendships formed reduce such privacy concerns (Tsay-Vogel et al., 2018). However, each country has different ways and levels of interacting with social media that may have an impact on privacy concerns, hence we propose:

H2: Social media activities influence privacy concerns in both countries, the UK and Thailand.

Similarly, personalised and interactive social media advertising, whilst suggesting a level of intrusiveness, also enhances consumer-brand image fit, leading to more likes and a positive relationship between social media advertising and brand response (Carlson, Hanson, Pancras, Ross, and Rousseau-Anderson, 2022). However, cultural factors on how social media activities are done in both countries may impact the intrusiveness perception:

H3: Social media activities influence the perceived intrusiveness of social media advertising in both countries, the UK and Thailand.

2.3 Privacy Concerns and Intrusiveness

Privacy concerns (PCs) among consumers often arise due to a perceived lack of control of their personal data and uncertainties about how retailers will manage such information during commercial transactions or communications (Okazaki, Eisend, Plangger, de Ruyter and Grewal, 2020). Additionally, research indicates that privacy concerns mediate relationships between factors like ad intrusiveness, ad intent, privacy self-efficacy, and privacy violation, ultimately predicting social media fatigue (Bright, Logan and Lim, 2022) reducing purchase intention:

H4: Privacy concerns influence purchase intention in both countries, the UK and Thailand.

Social media data can be monetised in various ways by organisations via user-generated content, geospatial or biometric data, and web tracking technologies to develop targeted marketing strategies and to personalise content, products, and experiences, thus strengthening customer relationships (Quach et al., 2022). Trackability enables brands to personalise their message to consumers, enhancing perceived relevance and congruence to consumers (Carlson et al., 2022). Thus, such information exchange with brands on social media can be perceived as fair. Although personalised brand messaging on social media can enhance perceived relevance, it may also be viewed as intrusive, resulting in unfavourable brand attitudes on how brands use consumers data which also could decreases purchase intention (Rana & Arora, 2021).

H5: Intrusiveness of social media advertising influences privacy concerns in both countries, the UK and Thailand.

H6: Intrusiveness of social media advertising influences purchase intention in both countries, the UK and Thailand.

Figure 1 represents the hypothesized relationships of this study.

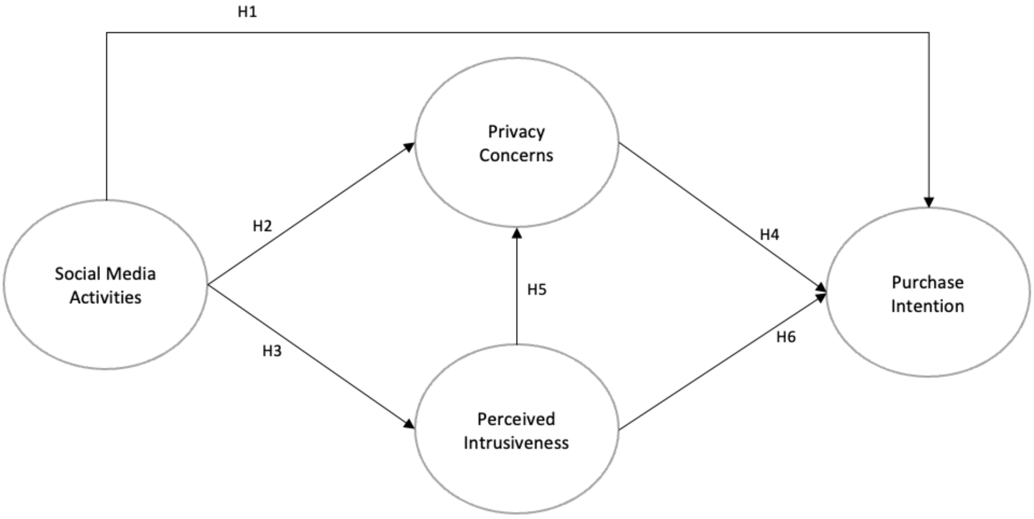


Figure 1 Conceptual Model

3. Method

This study uses a descriptive research design, using a questionnaire/survey-based research strategy. A Qualtrics panel administered the online survey to recruit and collect data from social media users in the UK and Thailand. The questionnaire began with an overview of the research aim and objectives to provide respondents with the necessary information regarding the survey, including how their anonymity and confidentiality would be reassured. Prior to survey completion, consent from each respondent was requested. Various screening questions were used to ensure that respondents reside within the UK or Thailand and were active on social media, with some level of exposure to social media activities, including social media advertising. The survey was completed by 1090 respondents (UK: 549, Thailand: 501).

The questionnaire contained four measures addressing the hypotheses. Seven-point Likert scales were used, ranging from 1 ‘strongly disagree’ to 7 ‘strongly agree’. All constructs were adapted from prior studies. Social Media activities (SMAs), representing consumers’ contribution and creation of social media activities, were measured based on the scales developed by Schivinski et al. (2016). Privacy concerns (PCs) were adapted from Gutierrez, O'Leary, Nripendra, Dwivedi, and Calle (2019). Intrusiveness (INT), was from Riedel, Weeks, and Beatson (2018). Finally, purchase intention (PIN) was adapted from Alalwan (2018). All scales were used in a reflective measurement model.

4. Results

4.1 Confirmatory Factor Analysis

Given the measures were adapted from previous studies, an exploratory factor analysis on all 23 items was conducted. The results of the EFA demonstrate that there are some high cross-loadings (higher than 0.6) between three items of PCs and INT. Therefore, in total, six items were deleted. Then, the reliability of scale tests was completed, using PAWS Statistics 18. The Cronbach's alphas for all constructs are in the acceptable range (0.809 - 0.953).

Then, to assess the validity of scales, a confirmatory factor analysis (CFA) was performed on AMOS v23, leading to further item deletions. Table 1 provides a list of all items used in the final model along with the composite reliabilities and its psychometric properties, including the composite reliability (CR) and average variance extracted (AVE). The discriminant validities were also satisfied for both UK and Thailand as shown in Table 2.

Constructs	**Item**	**Composite Reliability/** loadings **(Thailand)**	**AVE Thailand**	**Composite Reliability/** loadings **(UK)**	**AVE UK**
Social Media Activities (SMAs)		**0.909**	**0.714**	**0.926**	**0.759**
	I post videos/photos that show my brands	0.834		0.874	
	I comment on posts/videos/photos related to my brands	0.848		0.878	
	I share my brand related posts	0.861		0.920	
	I like posts/videos/photos related to my brands	0.836		0.809	
Intrusiveness (INT)		**0.825**	**0.611**	**0.796**	**0.567**
	I feel that advertisements on my social media are a waste of my time	0.755		0.669	
	I feel that advertisements that cover a significant portion of the screen when I am viewing social media are too obtrusive	0.836		0.811	
	I feel that advertisements in my social media are intrusive when they are not relevant to me	0.751		0.772	
Privacy concerns (PCs)		**0.816**	**0.689**	**0.887**	**0.797**
	I am concerned that companies are collecting too much information about me through my social media data	0.773		0.893	
	It bothers me when I do not have control over how companies use my social media data	0.884		0.892	

Purchase intention (PIN)		**0.939**	**0.756**	**0.954**	**0.807**
	I will buy products from brands that are promoted on social media	0.837		0.861	
	I desire to buy products from the brands that are promoted on advertisements on social media	0.874		0.913	
	I am likely to buy products from the brands that are promoted on social media	0.873		0.906	
	I plan to purchase products from the brands that are promoted on social media	0.866		0.907	
	My willingness to buy products from the brands I engage with on social media is generally high	0.897		0.903	

Table 1: List of All Items Used and Psychometric Values

4.2 Multi-group invariance test

Prior to testing the hypotheses, measurement invariance was performed on AMOS v23. The validity of the structure of the model across the two groups was assessed. First, the configural invariance was tested, and the assumption was held since the chi-square difference ($\Delta\chi^2$ = 12.858, $\Delta\chi^2$df = 10; p = 0.23) between the unconstrained (χ^2(142) = 328.431) and fully constrained (χ^2(152) = 341.289) is non-significant. The metric and scalar invariance assumptions were also held since there is no difference between CFI (ΔCFI = 0.00 for metric invariance and 0.002 for scalar invariance) and the differences between Gamma (ΔGamma = 0.0004 and 0.002) and MC (ΔMC = 0.001 and 0.007) values are less than the cut-off values (<0.02 for ΔCFI, 0.015 for ΔGamma and 0.02 for ΔMC), as recommended by Fan and Sivo (2009). Therefore, measurement invariance was considered satisfactory.

	INT		**SMAs**		**PCs**		**PIN**	
	TH	**UK**	**TH**	**UK**	**TH**	**UK**	**TH**	**UK**
INT	**0.782**	**0.753**						
SMAs	-0.384	-0.357	**0.845**	**0.871**				
PCs	-0.685	-0.612	0.232	0.216	**0.830**	**0.893**		
PIN	-0.365	-0.372	0.726	0.842	0.307	0.245	**0.870**	**0.898**

Table 2: Discriminant Validity Values

4.3 Hypotheses test

Table 3 presents the measures according to the path coefficients for and across the countries. AMOS v23 was used to run structural equation modelling (SEM) to test the hypothesized relationships. Regarding H1, social media activities of consumers (SMAs) are found to positively influence purchase intention (PIN) for both Thailand (β = 0.692; *p* = .000) and the UK (β = 0.813; *p* =.000), with no statistically significant path difference (χ^2 = 1.117, *p* = > .05). Thus, H1 is rejected. However, the result shows that the relationship between consumers' social media activities (SMAs) and privacy concerns (PCs) is positive but not statistically significant for both Thailand (β = 0.037; *p* > .05) and UK (β = 0.037; *p* > .05). Thus, H2 is rejected. The relationship between consumers' social media activities (SMAs) and perceived intrusiveness of social media advertising (INT) is negative and statistically significant for Thailand (β = -0.384; *p* = .000) and UK (-0.357; *p* = .000) with non-significant path difference as χ^2 = 0.264, *p* = > .05. Thus, H3 is supported. Furthermore, consumers' privacy concerns (PCs) are found to negatively influence purchase intention (PIN) of Thai respondents as β = -0.148; *p* < .01, whilst this relationship is not statistically significant for the UK respondents since β = -0.031; *p* > .05). H4 is partially supported. Yet, the path difference remains non-significant between the two countries (χ^2 = 3.444, *p* = > .05). Then, H5 depicting the relationship between perceived intrusiveness of social media advertising (INT) and privacy concerns (PCs) is statistically significant in Thailand (β = 0.699; *p p* = .000) and UK (0.613; *p* = .000), with non-significant path difference (χ^2 = 0.688, *p* = > .05). Thus, H5 is accepted. The relationship between perceived intrusiveness of social media advertising (INT) and purchase intention (PIN) are non-significant for both Thailand and the UK since β = 0.002; *p* > .05 and β = -0.063; *p* > .05) respectively. No path difference was evident either for this relationship. Thus, H6 is rejected.

	Hypothesised path	**Thailand**	**UK**	**Results**	**Path differences: χ^2(*p*-value)**
H1	SMAs → PIN	0.692***	0.813***	Supported	1.117 (0.29) (supported)
H2	SMAs→ PCs	$0.037^{n.s.}$	$0.003^{n.s.}$	Rejected	-
H3	SMAs→ INT	-0.384***	-0.357***	Supported	0.264 (0.608) (supported)
H4	PCs → PIN	-0.148**	$-0.031^{n.s.}$	Partially supported	3.444 (0.063) (supported)
H5	INT → PCs	0.699***	0.613***	Supported	0.688 (0.407) (supported)

H6	INT → PIN	0.002$^{n.s.}$	-0.063$^{n.s.}$	Rejected	-

Table 3: Path Analysis and Path Differences

5. Discussion and conclusion

The contribution of this study is two-fold. First, it highlights the crucial role of consumers' social media activities with brands as a lubricant for social commerce. In line with previous studies (Gutierrez et al., 2023), this study corroborates that the social media activities of consumers, such as contribution to and creation of brand-related posts, positively influence their intention to buy on social commerce in both countries. Some past studies argue that the trackability of social media marketing may render advertising too personalised, thereby increasing perceived intrusiveness (Aguirre, et al., 2015). However, this study has identified that consumers perceive social media advertising as less intrusive when actively engaging in social media activities. Since social media activities include content creation and contribution, consumers' attention to social media advertising enhances their ability to process brand-related content (Jung, 2017). Also, personalised social media advertising improves the perceived fit between consumers and brand advertising. Therefore, consumers engaged in social media activities may find the benefits of relevant brand information outweigh their concerns about intrusiveness. Their informational and social motivations (Alalwan, 2018) seem to explain why this study finds no support for the link between social media activities and their concerns over privacy. For consumers who engage in social media activities, privacy concerns are not prominent (Tsay-Vogel et al., 2018). Yet, when considering the relationship between intrusiveness and privacy concerns, this study demonstrates that their privacy concerns are triggered when they find social media advertising intrusive, and some research has shown that consumers feel unsettled, even creeped out by over-personalisation of ads (Krause et al., 2022). However, in our results, intrusiveness does not itself affect their purchase intention.

Second, this study reveals that consumers in Thailand and UK barely differ from each other regarding such relationships even though the regulatory regimes and patterns of social media usage differ considerably in the two countries. However, there does appear to be a small negative relationship between privacy concerns and purchase intention amongst Thai consumers, while UK consumers' intention to buy is not affected by privacy concerns. Despite the regulations on privacy becoming stricter worldwide, the pace of change in personal data processing regulations varies between countries and regions. Whilst areas such as the UK, the European Union (EU), and Australia have implemented robust legislation governing data that may help reassure consumers to an extent, a notable number of countries lack comprehensive, or indeed any, data protection laws. According to the United Nations Conference on Trade and Development (2021), around 71% of countries have some form of data protection legislation, and 9% have proposed laws in development. However, the absence of legislation in 15% of countries and the lack of available data in 5% highlights a substantial gap in regulatory coverage affecting consumer attitudes and behaviours. Indeed, Thai consumers have faced a series of data leaks from both private organisation and state institutions in recent times. For example, in 2020, a significant data breach occurred when a Thai health app leaked the personal information of millions of users, including sensitive medical data (Sriyai, 2024). There are also data breaches in Thailand from the private sector, perhaps explaining why privacy concerns appear to play a significant role in reducing purchase intention in Thailand.

As AI continues to shape the landscape of social media interaction and consumer behaviour, its implications for privacy and engagement will be paramount, particularly in diverse markets like Thailand and the UK. Future considerations must prioritise transparency in AI-driven data usage to mitigate privacy concerns, especially given recent global data breaches that have heightened consumer scepticism. While AI can enhance personalisation and improve engagement, it must be implemented to respect consumer autonomy and not manipulate purchasing decisions unduly (Labrecque, et al., 2024). Educating consumers about how their data is used and empowering them with control over their information will be critical in building trust. As regulations evolve globally, brands must stay compliant and proactive in their approach to data protection, aligning their strategies with emerging laws while fostering data governance and ethical frameworks around AI technologies (Manning, et al., 2023; Craigon, et al., 2023). Ultimately, balancing AI's benefits with consumer rights protection will be essential for maintaining trust and fostering positive interactions in the ever-evolving digital marketplace.

References

Aguirre, Elizabeth, et al. "Unraveling the Personalization Paradox: The Effect of Information Collection and Trust-Building Strategies on Online Advertisement Effectiveness." Journal of Retailing, vol. 91, no. 1, 2015, pp. 34–49, https://doi.org/10.1016/j.jretai.2014.09.005.

Alalwan, A. A. (2018). Investigating the impact of social media advertising features on customer purchase intention. *International Journal of Information Management,* 42, 65-77.

Bright, L.F., Logan, K. & Lim H.S. (2022). Social media fatigue and privacy: An exploration of antecedents to consumers' concerns regarding the security of their personal information on social media platforms. *Journal of Interactive Advertising*, 22(2), 125-140.

Carlson, J., Hanson, S., Pancras, J., Ross, W. & Rousseau-Anderson, J. (2022). Social media advertising: How online motivations and congruency influence perceptions of trust. *Journal of Consumer Behaviour*, 21(2), 197-213.

Craigon, P., Sacks, J., Brewer, S., Frey, J., Gutierrez, A., Jacobs, N., Kanza, s., Manning, L., Munday, S., Wintour, A. & Pearson, S. (2023) Ethics by Design: Responsible Research & Innovation for AI in the Food Sector. *Journal of Responsible Technology*, Volume 13. Available at: https://doi.org/10.1016/j.jrt.2022.100051

Fan, X. & Sivo, S.A. (2009). Using [Delta] goodness-of-fit indexes in assessing mean structure invariance. *Structure Equation Modeling: A Multidisciplinary Journal*, 16(1), 54-69.

Gutierrez, A., O'Leary, S., Nripendra, P.R., Dwivedi, Y.K. & Calle, T. (2019). Using privacy calculus theory to explore for entrepreneurial directions in mobile location-based advertising: Identifying intrusiveness as the critical risk factor. *Computers in Human Behavior*, 95, 295-306.

Gutierrez, A., Khanyapuss, P., Desai, B., Alwi, S., O'Leary, S., Chaiyasoonthorn, W. & Chaveesuk, S. (2023). Retailers, don't ignore me on social media! The importance of consumer-brand interactions in raising purchase intention - Privacy the Achilles heel. *Journal of Retailing and Consumer Services*.

Hollebeek, L.D., Glynn, M.S. & Brodie, R.J. (2014). Consumer brand engagement in social media: Conceptualization, scale development and validation. *Journal of Interactive Marketing*, 28(2), 149-165.

Jung, A.R. (2017). The influence of perceived ad relevance on social media advertising: An empirical examination of a mediating role of privacy concern. *Computers in Human Behavior*, 70, 303-309.

Krause, K., Groeppel-Klein, A., Friderich, S. N., & Schmitz, M. (2022). Effects of Personalization and Ad Algorithm Disclosure on Perceived Creepiness. Advances in Consumer Research, 50, 208–209.

Labrecque, L., Peña, Y.P., Leonard, H. & Leger, R. (2024) Not All Sunshine and Rainbows: Exploring the Dark Side of AI in Interactive Marketing. *Journal of research in interactive marketing* 18(5), 970–999.

Li, H., Edwards, S.M., & Lee, J-H. (2002). Measuring the Intrusiveness of Advertisements: Scale Development and Validation. *Journal of Advertising*, 31(2), 37-47.

Manning, L., Brewer, S., Craigon, P., Frey, J., Gutierrez, A., Jacobs, N., Kanza, s., Munday, S., Pearson, S. & Sacks, J. (2023) Reflexive governance architectures: Considering the ethical implications of autonomous technology adoption in food supply chains. *Trends in Food Science & Technology*. Available at: https://doi.org/10.1016/j.tifs.2023.01.015

McClure, C. & Seock, Y.K., (2020). The role of involvement: Investigating the effect of brand's social media pages on consumer purchase intention. *Journal of Retailing and Consumer Services*, 53, 101975.

Mutambik, I., Lee, J., Almuqrin, A., Zhang, J. Z., & Homadi, A. (2023). The Growth of Social Commerce: How It Is Affected by Users' Privacy Concerns. *Journal of Theoretical and Applied Electronic Commerce Research*, 18(1), 725–743.

Nazir, S., Khadim, S., Asadullah, M. A. & Syed, N. Exploring the Influence of Artificial Intelligence Technology on Consumer Repurchase Intention: The Mediation and Moderation Approach. *Technology in society*, 72, 102190.

Okazaki, S., Eisend, M., Plangger, K., de Ruyter, K., & Grewal, D. (2020). Understanding the Strategic Consequences of Customer Privacy Concerns: A Meta-Analytic Review. *Journal of Retailing*, *96*(4), 458–473. https://doi.org/10.1016/j.jretai.2020.05.007

Quach, S., Thaichon, P., Martin, K. D., Weaven, S., & Palmatier, R. W. (2022). Digital technologies: tensions in privacy and data. *Journal of the Academy of Marketing Science*, 50(6), 1299–1323. https://doi.org/10.1007/s11747-022-00845-y

Rana, M. & Arora, N. (2021). How Does Social Media Advertising Persuade? An Investigation of the Moderation Effects of Corporate Reputation, Privacy Concerns and Intrusiveness. *Journal of Global Marketing*, 35(3), 248-267.

Riedel, A.S., Weeks, C.S., & Beatson, A.T. (2018). Am I intruding? Developing a conceptualisation of advertising intrusiveness. *Journal of Marketing Management*, 34(9-10), 750-774. DOI: 10.1080/0267257X.2018.1496130

Schivinski, B., Christodoulides, G. & Dabrowski, D. (2016). Measuring consumers' engagement with brand-related social-media content: Development and validation of a scale that identifies levels of social-media engagement with brands. *Journal of Advertising Research*, 56(1), 64-80.

Sriyai, S. (2024). Thailand's Public Sector Data Breaches Erode Public Trust – And Might Undermine E-Government. Retrieve from: https://fulcrum.sg/thailands-public-sector-data-breaches-erode-public-trust-and-might-undermine-e-government/ (Last accessed, September 9, 2024)

Statista (2024a). Percentage of online consumers buying from social networks in selected countries worldwide in 2024. Retrieved from: https://www.statista.com/statistics/1252481/social-buyers-worldwide-countries/ . (Last accessed: March 1, 2024)

Statista (2024b). Awareness of internet users worldwide of their country's privacy laws as of June 2024, by country. Retrieved from: https://www.statista.com/statistics/1441448/privacy-laws-awareness-global-by-country/. (Last accessed: March 1, 2024)

Trade and Development Report (2021) From recovery to resilience: the development dimension. Retrieved from: https://unctad.org/publication/trade-and-development-report-2021. (Last accessed: December 1, 2023)

Tsay-Vogel, M., Shanahan, J., & Signorielli, N. (2018). Social media cultivating perceptions of privacy: A 5-year analysis of privacy attitudes and self-disclosure behaviors among Facebook users. *New Media & Society, 20*(1), 141-161.

An Empirical Investigation into Initial Opinion Formation on Social Media Platforms

Venu Bhaskar Puthineedi
Trinity Business School, Trinity College Dublin, puthinev@tcd.ie
Ashish Kumar Jha
Trinity Business School, Trinity College Dublin, akjha@tcd.ie

Completed Research

Abstract

Digital social media platforms are flooded with vast amounts of unknown and often unverified information, where users are constantly exposed to new content that can quickly influence their perspectives. The initial formation of opinions in this environment is critical, as first impressions shape immediate reactions and can significantly influence user behavior, political leanings, and consumption patterns. Understanding how these opinions form is essential for fostering responsible information sharing and combating misinformation. This research explores the interaction between cognitive processes and the digital identity of information sources on opinion formation. Study one, involving 320 participants, reveals that initial opinions begin to solidify after consuming five pieces of information. Study two, with 180 participants, investigates factors such as cognitive structures, personality traits, and socio-demographic characteristics, finding that users are more likely to trust information from profiles with unverified titles (e.g., Dr. or Doctor) than from verified expert influencers.

Keywords: Digital platforms, initial opinions, digital identity.

Introduction

"Suddenly, Everybody's an Expert"[1]

- New York Times (February 3, 2000)

This news headline was published by the New York Times in the early 2000s, during the nascent days of social media. Notably after the pandemic, there has been a growing trend of individuals presenting themselves as experts online, often in areas far beyond their domain of expertise. One issue with individuals suddenly becoming self-proclaimed experts is that many topics have multiple perspectives, yet people frequently promote "facts" (whether presumed or accurate) from just one side, posing as authorities. When consuming information on social media, users often fail to verify the authenticity of what they read, leading to the spread of half-truths and, ultimately, misinformation (Moravec et al., 2022, Laato et al., 2020).

Extensive research highlights the role of pre-existing beliefs and confirmation bias as significant contributors to the spread of misinformation (Modgil et al., 2021). However, a more

[1] https://www.nytimes.com/2000/02/03/technology/suddenly-everybody-s-an-expert.html

intriguing question arises: in situations where users lack pre-existing beliefs or opinions, how do they consume information and form their initial opinions?

Understanding the factors that drive the formation of initial opinions on social media is crucial for developing effective strategies that mitigate adverse outcomes and encourage more balanced information consumption. Individual differences in cognitive schemas, personality traits, and sociodemographic factors influence opinion formation (Chan et al., 2023), highlighting the need for tailored information systems to promote responsible social media use. Therefore, it is essential to study the factors influencing the formation of initial opinions. This research seeks to answer the question: *When do individuals form their initial opinions while consuming information on social media platforms, and what factors contribute to this process?*

The study is operationalized through two experiments. The first experiment addresses the question of when initial opinions form as individuals consume new information on social media. Further, tests if social media users are actually not concerned with the veracity of the information. The second experiment aims to identify the factors that influence the formation of initial opinions. This research contributes to theory and practice by revealing two key findings: first, the minimum number of information points required to form opinions in individuals is *five*; second, social media users are more likely to trust information from profiles with unverified titles than from those of expert influencers.

Background and literature

Extensive research highlights that pre-existing beliefs and confirmation bias are significant factors driving the spread of misinformation (Moravec et al., 2020, Kim et al., 2019). However, an intriguing question arises: in scenarios where users lack pre-existing beliefs and opinions, how do they consume information and form initial opinions? Surprisingly, there is limited literature on how users form initial opinions. Therefore, our research aims to address this gap by investigating how users consume information and form initial opinions in the absence of pre-existing beliefs and opinions.

When consuming information in a hedonic mode, individuals tend to engage with content that aligns with their existing beliefs (Kapoor et al., 2018). The ability to perceive, process, and retain information is determined by the amount of cognitive resources available, as explained by cognitive load theory and information processing theory (Hu et al., 2017). These resources are influenced by the complexity of the information, its relevance to understanding the topic, and the user's interest (Plass et al., 2010).

In contrast individuals consume new information of interest, they utilize cognitive resources to perceive and process the information. As they continue to consume information on similar topics, the remaining cognitive resources diminish. As this amount of information consumed increases and cognitive resources decrease, users may experience cognitive effects such as reduced attention focus, information overload, decision fatigue, and decreased cognitive performance (Laato et al., 2020). Consequently, individuals tend to make quick decisions based on the initial information consumed that is etched in their memory. This implies that the initial information consumed is responsible for the formation of initial opinions, which are then used as a basis for making decisions on new information. The crucial question to be addressed here is: *At what point are these initial opinions formed*?

Critical Information Load

The point at which individuals transition from being novices to exhibiting expert behavior in a topic solely based on the volume of information consumed is termed the point of critical information load. At this juncture, individuals become inundated with information that surpasses their cognitive capacity, prompting them to rely on existing knowledge to assess new information. As a result, they assume a sense of expertise based solely on the information they have consumed, illustrating the phenomenon of cognitive overload. This phenomenon is explained by the Dunning-Kruger effect, which states that individuals with lower levels of competence in a particular domain are more likely to overestimate their abilities compared to those with higher levels of competence (Kruger and Dunning, 1999).

After reaching the critical information load threshold, individuals lacking prior knowledge on a topic tend to exhibit behaviors akin to experts, assessing and validating new information through the lens of their initial opinions and beliefs. According to cognitive response theory, exposure to information congruent with one's existing opinions triggers confirmation bias, wherein individuals selectively seek, interpret, and favor information that aligns with their preexisting beliefs (Klayman and Ha, 1987). Conversely, when confronted with information that contradicts their opinions, individuals may experience cognitive dissonance, resulting in reduced engagement with the information (Festinger, 1962).

Based on Cognitive response theory and arguments form Kruger and Dunning (1999), individuals acting like experts after the point of critical information load should experience confirmation bias if the new information point aligns with the initial opinions formed and experience cognitive dissonance if the initial opinions are challenged. If the new information consumed reinforces the initial opinions, then individuals tend to like, share, and believe more, resulting in increased engagement with it (Kim and Dennis, 2019). However, if the new information is contrary to the initial opinions, then engagement with the information reduces.

Process of Digital Engagement

Extensive literature emphasizes the impact of source credibility on user behavior and information consumption (Cheung et al., 2012). However, a less-explored aspect is source's digital identity. Social identity theory posits that individuals derive a significant portion of their identity from the social groups to which they belong (Hogg, 2016). As per this theory, individuals construct their identity through a blend of self-presentation, interaction, and participation within social groups.

Similarly, social media users craft digital identities that reflect their personal interests, professional affiliations, and cultural indicators (Kreps, 2010). These digital identities can be perceived differently by other users based on their cognitive schemas for consuming information. Schemas are pre-existing cognitive structures that aid in organizing and interpreting information (McVee et al., 2005). The digital identity serves as one of these schemas, influencing how users perceive and interpret information shared by the source. In line with this, users on social media tend to trust and engage with information that aligns with their pre-existing cognitive structures. Figure 1 describes the formation of initial opinions in individuals consuming new information of interest on social media platforms.

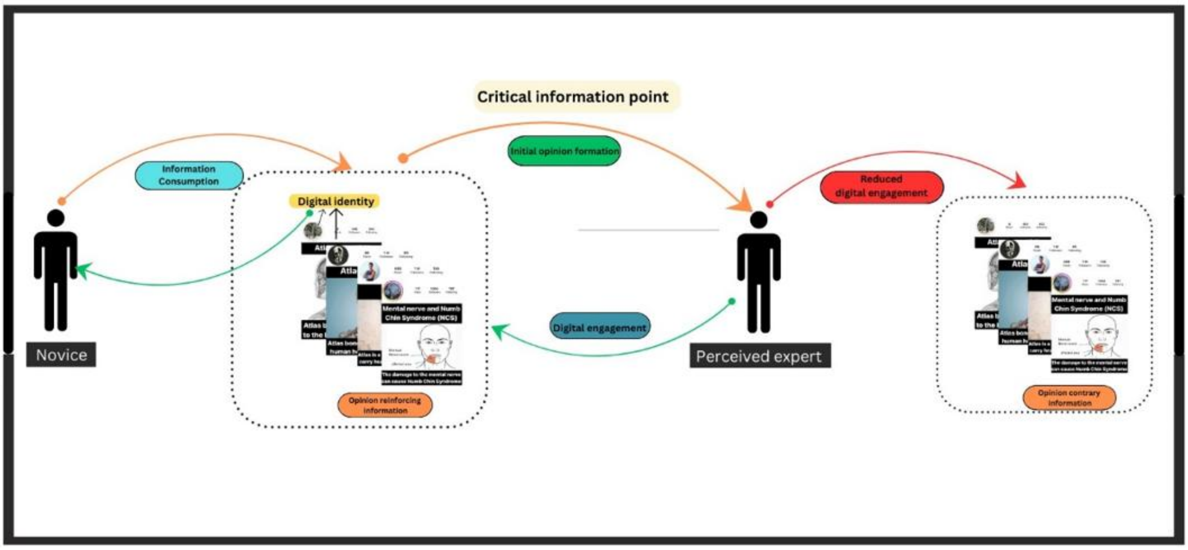

Critical information point
Information Consumption
Initial opinion formation
Reduced digital engagement
Digital identity
Novice
Digital engagement
Perceived expert
Mental nerve and Numb Chin Syndrome (NCS)
Opinion reinforcing information
Opinion contrary information

Figure 1: Initial opinion formation on social media platforms in hedonic mode.

Hypothesis Development

Information and cognitive load

In a landscape inundated with news stories, opinions, and perspectives, users may feel compelled to form rapid judgments based on initial information encountered in their feed that aligns with their interests . According to cognitive load theory, in such environments, users' ability to perceive, process, and retain this information is dependent on the availability of cognitive resources (Hu et al., 2017). Within these resources, the complexity of the information, its relevance to understanding the topic, and the user's interest in the topic all play significant roles (Plass et al., 2010).

According to Kim et al. (2019), users on social media platforms consume information for pleasure and fun (i.e., hedonic mode of information consumption), the cognitive system-1 is active, where users spend less cognitive effort to analyze information (Moravec et al., 2020). When users encounter new information on social media platforms related to their interests, they expend the available cognitive resources to process the information (Plass et al., 2010). The nature of the landscape, they are compelled to form opinions. Further, when users consume information that aligns with their initial opinions, they tend to expend the remaining cognitive resources. Thus, leading to the formation and reinforcement of initial opinions regarding the new topic of interest.

Moreover, individuals may experience cognitive consonance when presented with information that aligns with new opinions, further reinforcing their confidence in those beliefs and motivating continued engagement (Klayman and Ha, 1987). Additionally, users may engage with and share information as a means of social validation, seeking a sense of belongingness within their social network (James et al., 2017). Therefore, we hypothesize that with increased information load, individuals are more likely to engage with, share, and believe information on social media platforms related to their topics of interest, even if they have not previously encountered it.

H1: Increased information load leads individuals to like (H1a), share (H1b), and believe (H1c) more posts on social media, even with minimal prior knowledge.

Point of critical information load and veracity

According to Dunning-Kruger effect, individuals with lower levels of competence in a particular domain are more likely to overestimate their abilities when compared to those with higher levels of competency (Kruger and Dunning, 1999). When consuming new information on social media with limited cognitive resources, individuals with little to no prior knowledge of the topic may develop a sense of confidence or perceived expertise in the topic. In such scenarios, the initial information consumed plays a crucial role in the formation of opinions and beliefs on a new topic. If the initial new information consumed is responsible for the formation of initial opinions and might influence the user's behaviour regarding new information points consumed, then the most important question that might arise is: *what is the point of critical information load?*

Based on the research done by Miller (1967) the short-term memory of individuals can hold up to seven plus or minus two information points. In the hedonic mode of information consumption, the users perceive, evaluate, and consume information by expending minimum cognitive resources and recalling information that is easily available (i.e., short-term memory). Based on this theory, an initial experiment is conducted to identify the point of critical information load, and the results suggest that the minimum number of points at which opinions are formed and users become perceived experts in a topic is five.

According to cognitive response theory, if the information aligns with the initial beliefs of the users, then confirmation bias comes into play, and the interaction with the information doesn't change (Klayman and Ha, 1987). However, when consuming information that contradicts the initial beliefs, the users experience cognitive dissonance, which reduces the interaction with the information. Therefore, at the point of critical information load, if the social media users tend to behave like experts in a topic they haven't encountered before, then based on these beliefs, their interaction with further information points should vary based on the nature of the further information points presented to them. Hence, we hypothesize that:

H2: Beyond the critical information point, subsequent information conflicting with initial exposure is less likely to be liked (H2a), shared (H2b) and believed (H2c).

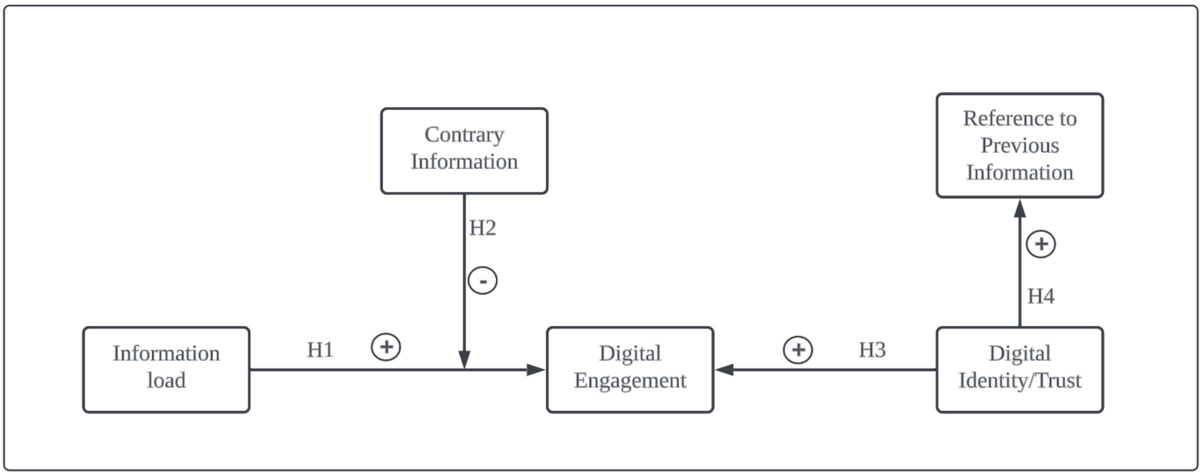


Figure 2: Conceptual framework of the research.

Digital Identity and information load

The vast amount of research revolves around source credibility theory, which provides a clear picture on various source characteristics that can influence information consumption interaction and dissipation (Cheung et al., 2012). However, little research dwells on the concepts of perceived digital identity which transcends the projected source reputation which concerns the number of followers, likes, amount of interaction, etc (Jha and Shah, 2021).

Social identity theory posits that individuals derive a significant portion of their identity from the social groups they belong to (Mishra et al., 2012). According to social identity theory, individuals construct their identity through a combination of aspects like self-presentation, interaction, and participation in social groups (Ellemers and Haslam, 2012). Analogously, social media users create digital identities that reflect their traits such as personal interests, professional affiliations, and cultural indications. Other users can perceive the created digital identity differently based on their schema for consuming information.

The schemas are the pre-existing cognitive structures that help organize and interpret information (McVee et al., 2005). The digital identity acts as one of the schemas that influence how users perceive and interpret the information shared by the source. Credibility and trust in the digital identity play a crucial role in how users accept and engage with the information (Kelton et al., 2008). According to this the users on social media trust information and engage with it, if the information presented aligns with their pre-existing cognitive structures.

Based on the cognitive learning theory the new information consumed will be stacked, altered, or discarded based on the nature of the information (Fox, 1997). If the information aligns with the pre-existing cognitive structures and interests, then the information is stacked. The digital identity of the source is an aspect responsible for creating a certain schema or cognitive structure for the consumption and retention of information. The information consumed will influence further information consumed only if the initial information is retained. Hence, to test this line of argument, we hypothesize that:

H3: The likeability (H3a), shareability (H3b), and believability (H3c) of subsequent posts on digital social media platforms are impacted by the digital identity of the source.

Digital identity and confirmation bias

According to cognitive response theory, confirmation bias is the tendency to interpret, search for, recall, and favor information in a way that confirms one's pre-existing beliefs (Westerwick et al., 2017). Pre-existing beliefs come into play if the users have prior knowledge about the topic of information being consumed. In scenarios where there is little to no prior knowledge, social media users resort to information consumption schema or cognitive structures to consume and interact with the information. This implies that the formation of initial opinions in social media users depends on the cognitive structures (McVee et al., 2005).

When consuming new information on the topic of interest, if the information fits in the cognitive structures, then users consume it, and based on this initial information, the opinions and beliefs are formed or updated. Once these initial opinions are formed they act as the pre-existing beliefs along and will be responsible for how users interact with further information on the same topic. This implies that the users are recalling or referring back to the initial information when consuming further information.

Based on the cognitive load theory when consuming new information on any topic, the users will expend less cognitive resources and interact with the information based on existing beliefs (Sweller, 2011). So we argue that the inference or reference to the previous information is dependent on the cognitive structures of the users which is dependent on the trust digital identity of the source. Hence we hypothesise that:

H4: The trust in the digital identity of the initial information will impact the propensity of the information reception to refer back to the original information points.

Methodology

This research is a two-part study. The first study is focused on understanding the formation of initial beliefs and opinions in users while consuming information on social media platforms, and the second study revolves around the influence of sources on the formation of initial opinions and how various factors influence social media users' behavior.

Study-1

Factorial design

This design enables an in-depth examination of how information load, the nature of the information (reinforcing or challenging), and its veracity (true or false) affect participants' information processing and behavior. The experimental setup is designed to generate data that will confirm or refute the stated hypotheses. Table 1 outlines the experimental design, where, for example, the subjects in group 3T/1F indicate three true initial information points followed by one false subsequent information point. The subjects are then prompted to express their likeliness of liking, sharing, and believing the subsequent information on a seven-point Likert scale. Apart from the quantified opinions, demographical features such as age, gender, education qualification, social media usage, and income are captured.

	3/1		5/1	
Opinion-reinforcing	3F/1F	3T/1T	5F/1F	5T/1T
Opinion-challenging	3F/1T	3T/1F	5F/1T	5T/1F

Table 1: Experimental design of study 1

The experimental design comprises three factors, with two levels in each level, culminating in a three-factor, eight-level matrix. The first factor pertains to the number of information points presented to participants. This factor comprises two conditions: the first group is exposed to four information points, designated as the **3/1** condition, where the first three information points are classified as initial information, while the fourth serves as subsequent information. Similarly, the second group is exposed to six information points, referred to as the **5/1** condition, where the first five information points are treated as initial information, and the sixth point operates as subsequent information, as delineated in Table 1.

Participants, based on their assigned group, initially receive either three or five pieces of initial information, followed by an evaluation of their opinion formation upon exposure to the subsequent information point. This evaluation measures participants' propensity to like, share, and believe the subsequent information on a seven-point Likert scale. This level evaluates the influence of information load on users' information consumption and provides a basis to evaluate the first set of hypotheses. Each information point utilized in the experiment is represented by a distinct Instagram post.

The second factor focuses on the nature of the subsequent information point, with two levels: **opinion-reinforcing** and **opinion-challenging** information. While the initial information remains consistent across both groups at this level, the critical variable that shifts is the subsequent information. For participants in the reinforcing information condition, the subsequent information aligns with the initial information. Conversely, for those in the challenging information group, the subsequent information contradicts the initial information. The incorporation of this level expands the design into a two-factor, four-level matrix and aids in studying the formation of initial opinions.

Finally, to evaluate the impact of information authenticity on participants' opinions, a third factor is introduced to the experimental grid, where the veracity of information becomes a key factor, resulting in a three-factor, eight-level experimental design. Participants are exposed to either true (denoted as T) or false (denoted as F) information. In the reinforcing group, participants who receive true initial information are provided with true subsequent information, while those presented with false initial information continue with false subsequent information. In contrast, participants in the challenging group who receive true initial information are later presented with false subsequent information, and those with false initial information encounter true subsequent information.

Stimuli

The second section of the study dived into the subject's perceptions while consuming the information presented to them. In this phase, users encountered information presented in a format resembling posts on the Instagram platform. Our study was focused on understanding the influence of initial information on the consumption and interaction of information in the future and the formation of beliefs and opinions. This posed the challenge of creating information points that were new to the subjects, and at the same time, the new information points needed to be in line with the subjects' interests. This challenge was addressed by choosing information points from the medical domain.

In the study, less familiar anatomical concepts, such as the atlas bone and the mental nerve, were employed strategically to mitigate the potential influence of pre-existing beliefs and opinions among the subjects regarding a specific concept or domain. Despite these anatomical concepts being relatively unknown to the general public, selecting participants for the study who lacked awareness of these concepts yet expressed an interest in engaging with such information on social media posed a notable challenge.

Participants

The study was conducted in the medical context. To identify the participants who lacked awareness about the stimuli used in the study yet had an interest in engaging with such content on social media platforms, a pre-screening survey was conducted to identify appropriate subjects for the study. The pre-screening survey was designed to identify subjects interested in medical information and follow medical pages on social media who don't work in the medical industry. The short-listed subjects were then requested to take part in the main study.

The study was floated on Prolific in November 2023, and 303 short-listed adults from the US volunteered to complete the survey. The subjects were randomly split into 8 groups mentioned in Table 1. Among the 303 responses, 22 were incomplete, 18 failed the attention checks and 17 responses were deemed to be invalid as they were patterned responses, resulting in a total of 246 valid responses.

Study-2

Design and stimuli

The second phase of the research focuses on examining the impact of information sources on user interaction and referencing of prior information. Utilizing information from the first study, this phase assesses the influence of different sources on information interaction and reference to previous information. This phase delves into the effects of various types of social media profiles in the medical domain, specifically on Instagram.

The study aims to comprehend the influence of four distinct profiles disseminating medical information on Instagram. These profiles include a doctor profile with a high number of followers, an expert influencer with a substantial following, a doctor profile with a low number of followers, and a common man profile with a limited following.

The experiment was structured into two sections. The first section collected participants' demographic information. In the second section, participants were randomly assigned to one of the four predefined profile groups. Each participant was then exposed to two segments, with each segment containing six discrete information points (five initial information points and one subsequent information point). Following the presentation of these points, participants were asked to rate their likelihood of liking, sharing, and believing the subsequent information.

Responses were measured using a seven-point Likert scale, ranging from 1 (extremely unlikely) to 7 (extremely likely).

The effect of the profile variable on information consumption provided a foundation for evaluating the third set of hypotheses. Additionally, participants were asked to assess their likelihood of referring to the previous information point while interacting with the current one. This approach enabled a nuanced examination of the relationship between the assigned profile and patterns of information interaction and consumption. The stimuli for the second study mirror those of the first study, with the only variation being the profile sharing the information.

Participants

The second study also required participants who were unaware of the stimuli used in the study yet were interested in engaging with medical content on social media platforms. So, another short pre-screening study similar to the first pre-screening study was conducted on prolific to identify appropriate subjects. The study was floated on Prolific in December 2023, and 221 short-listed adults from the US volunteered to complete the survey. The subjects were randomly split into 4 groups. Of the 221 responses, 17 were incomplete, and 12 failed the attention checks. Upon further analysis, 13 responses were deemed to be invalid as they were patterned responses, resulting in a total of 179 valid responses.

Results and Analysis

The findings from descriptive statistics of the experimental data underscore a notable absence of significant correlations among the independent variables. This absence indicates a lack of multicollinearity concerns, fortifying the statistical independence of the variables. The results from randomization tests underscore the efficacy of the randomization methodology employed, as no statistically significant differences were observed in the characteristics of subjects across different experimental groups. These findings, established at a 95% confidence level, validate the successful random assignment of subjects for both studies.

We implemented Ordered Logit models to study the influence of user-level factors and experimental variables on the dependent variables likability, shareability, believability, and reference to previous information. Tables 2 and 3 present the ordered logit results for studies 1 and 2. Study 1 is focused on examining the influence of information level factors (**information load, nature of information,** and **veracity of information**) on the dependent variables, whereas study 2 is focused on understanding the influence of profile (**profile** and **trustworthiness of profile)** on the dependent variables.

Variable	**Likeability**	**Shareability**	**Believability**	**Reference to Previous Information**
(Intercept)				
Gender	-0.504 (0.222) *	-0.557 (0.231) *	-0.251 (0.228)	-0.241 (0.220)
Age	0.018 (0.015)	0.005 (0.015)	-0.013 (0.015)	-0.010 (0.015)
Education Qualification	-0.028 (0.159)	0.162 (0.162)	0.164 (0.168)	0.153 (0.161)
Work Experience	-0.022 (0.016)	-0.009 (0.016)	0.022 (0.016)	0.022 (0.016)
Annual Salary	-0.025 (0.063)	-0.043 (0.063)	-0.121 (0.066)	-0.025 (0.062)
Hours on SM	0.057 (0.030)	0.065 (0.036)	-0.008 (0.030)	0.024 (0.033)
Number of SM	0.059 (0.090)	-0.007 (0.090)	-0.008 (0.092)	-0.044 (0.089)
Amount of News	0.137 (0.096)	0.208 (0.099) *	0.134 (0.099)	0.259 (0.098) *
Information load	0.245 (0.230)	0.715 (0.235) **	0.921 (0.238) **	0.751 (0.235) **
Nature of information	-0.579 (0.231) *	-0.580 (0.232) *	-0.546 (0.237) *	-0.162 (0.229)
Veracity of information	0.316 (0.230)	0.302 (0.231)	0.124 (0.232)	0.046 (0.231)
AIC	905.289	888.576	823.237	893.498
observations	246	246	246	246

Values in parentheses represent standard error; *** p<0.01, ** p<0.05, * p<0.1

Table 2: Study 1 ordered logit analysis.

	Likeability	**Shareability**	**Believability**	**Reference to Previous Information**
Gender	-0.319 (0.280)	-0.283 (0.282)	-0.208 (0.289)	0.426 (0.285)
Age	0.010 (0.013)	0.012 (0.013)	0.017 (0.013)	-0.004 (0.012)
Education Qualification	-0.057 (0.067)	0.017 (0.065)	0.032 (0.072)	0.117 (0.063)
Work Experience	-0.001 (0.001)	-0.000 (0.001)	-0.001 (0.001)	-0.001 (0.001)
Annual Salary	-0.050 (0.063)	-0.045 (0.066)	-0.009 (0.063)	-0.005 (0.062)
Hours on SM	0.121 (0.085)	0.112 (0.082)	0.009 (0.085)	-0.036 (0.081)
Amount of News	0.006 (0.103)	0.084 (0.101)	0.063 (0.102)	0.079 (0.101)
Profile	-0.356** (0.135)	-0.167** (0.136)	-0.272(0.138).	-0.125 (0.136)
Trustworthy	0.694*** (0.114)	0.382*** (0.112)	0.736*** (0.117)	0.583*** (0.110)

AIC	621.407	635.236	586.657	642.597
Observations	179	179	179	179

Values in parentheses represent standard error; *** p<0.01, ** p<0.05, * p<0.1

Table 3: Study 2 ordered logit analysis.

Robustness Tests

Multiple checks and tests were systematically conducted from both theoretical and analytical perspectives to ensure the stability and generalizability of the results. Linear regression was employed to compare the results with the ordered logit, and the results are similar, which validates the findings. Furthermore, residual analysis of regression revealed linear residuals with no evidence of heteroscedasticity.

Likability, shareability, believability and reference to previous information served as the four dependent variables in our research. These variables were measured on a scale ranging from 1 to 7. Notably, our findings revealed that these dependent variables exhibited both left- and right-censoring, with values restricted to a minimum of 1 and a maximum of 7. Given these constraints, it was imperative to assess the results beyond these limitations. Therefore, we employed a Tobit regression model, allowing us to compare and validate the regression results. Encouragingly, these results mirrored the findings of the regression analysis, providing reassurance that the censored values did not compromise the validity of our results.

Discussion and Implications

Discussion

The results from Table 2 underscore the significant impact of *information load* on user responses. The regression analysis reveals positive influence of information load on shareability (coefficient 0.715, $p < 0.05$), believability (coefficient 0.921, $p < 0.01$), and reference to previous information (coefficient 0.751, $p < 0.05$). This suggests that as users consume more information points, ranging from three to five, there is a marked increase in shareability, believability, and reference to previous information, supporting hypotheses H1b and H1c. However, although likability increases with a coefficient of 0.245, this change is not statistically significant, failing to support hypothesis H1a.

These results align with the theory, indicating that as information load increases, confidence in beliefs and engagement also significantly increase. This lends support to the argument that the minimum threshold for forming initial opinions and beliefs among social media users is five information points. Furthermore, this aligns with the proposition made by Miller (1967) regarding the capacity of working memory to hold seven plus or minus two information points necessary for processing and storing information to form beliefs.

Hypotheses H2a, H2b, and H2c posited that beyond a critical information point, conflicting new information diminishes the likeability, shareability, and believability of initial information. From Table 2, it's evident that as the *nature of information* changes from reinforcing (1) to challenging (2), the likeability (coefficient -0.579, p<0.1), shareability (coefficient -0.580, p < 0.05) and believability (coefficient -0.546, p < 0.1) significantly decrease beyond the critical information load, i.e., after five information points, when encountering contradictory information. This implies that transitioning from reinforcing to conflicting information leads to a notable reduction in likeability, shareability, and believability.

To assess hypotheses H2a, H2b and H2c, a comparative analysis was conducted to determine the critical information point at which opinions are formed and serve as preexisting beliefs for subsequent information. Figure 3 depicts the average shareability and believability of information when presented with reinforcing and conflicting information among groups exposed to three and five information points. The results demonstrate that after consuming three information points, there is no significant variation in the average shareability and believability of subsequent information, regardless of whether it reinforces or conflicts with initial opinions. However, after consuming five information points, the average shareability significantly decreases (p < 0.05) from 4.017 to 2.951, and believability decreases significantly (p < 0.05) from 5.121 to 4.541 when encountering conflicting information providing support for the hypotheses H2a, H2b and H2c.

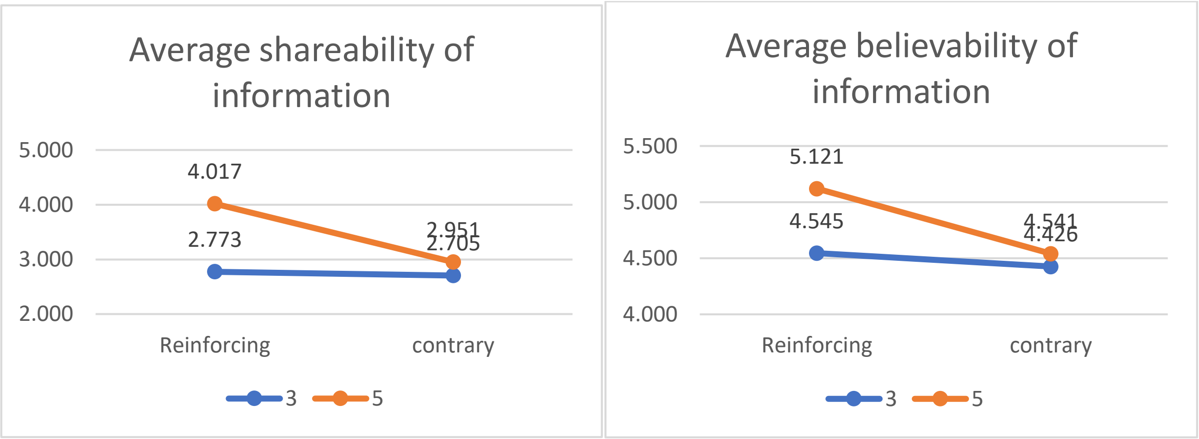


Figure 3: Average shareability and believability of information between groups presented with 3 and 5 information points

Table 3 offers compelling evidence that validates the third set of hypotheses, indicating that the digital identity of individuals (*profile*) influences the likability (coefficient -0.356, $p < 0.1$), shareability (coefficient -0.167, $p < 0.1$), and believability (coefficient -0.272, $p < 0.1$) of subsequent posts on digital social media platforms. These findings suggest that as the digital identity of the source value transitions across categories (1-Doctor with high following, 2-Influencer, 3-Doctor with low following, and 4-Common man), there is a significant decrease in both interaction and believability of the consumed information presented in Figure 4. This lends support to hypotheses H3a, H3b, and H3c.

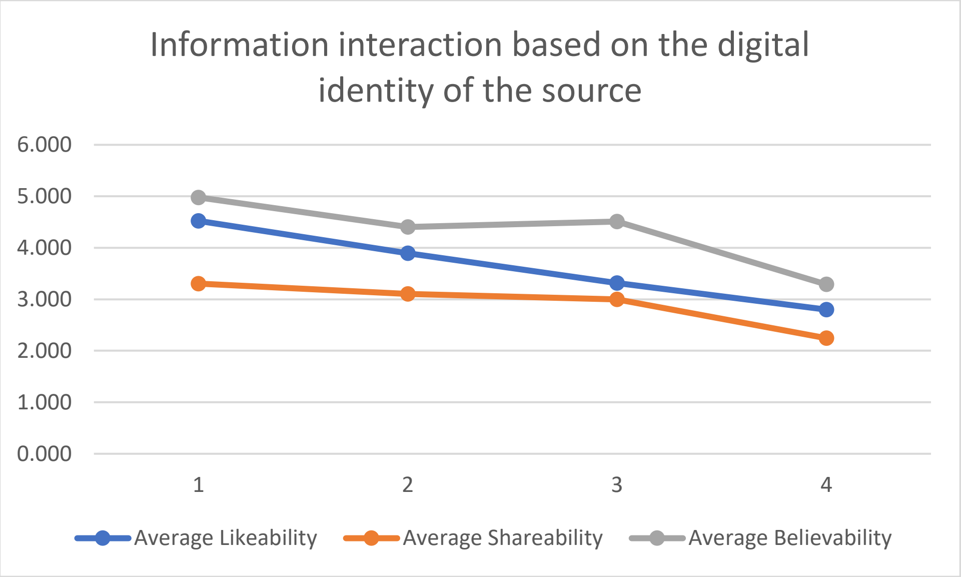


Figure 6: Information interaction of the subjects based on the digital identity of the information source.

Upon closer examination, we observe that subjects demonstrate a higher propensity to believe information disseminated by an unverified doctor with a low follower count compared to content shared by a verified social media influencer. This intriguing finding not only underscores the complex interplay between digital identity and user behavior but also carries significant practical implications for social media platforms.

The results from Table 3 reveal a significant influence of trust in the digital identity (*trustworthiness*) of the source on users' reference to previous information on social media platforms (coefficient 0.583, $p < 0.05$). This suggests that as trust in the digital identity of the source increases, users are more likely to reference previous information provided by the source when consuming new information on similar topics.

In summary, our study provides robust statistical support for all hypotheses except H1a. These findings affirm the central proposition that users form opinions after consuming five information points on social media platforms, highlighting the significance of the fifth information point as the point of critical information load. This underscores the pivotal role of initial information veracity in shaping user opinions on social media topics.

Additionally, we analyzed the impact of demographic features on information consumption. While our findings offer initial evidence suggesting that females tend to share and believe information more than males, further investigation is warranted. Furthermore, we observed that increased time spent on social media and higher information consumption correlate with heightened information shareability.

A significant finding of our research pertains to the veracity of initial information presented to the subjects. The categorical variable "veracity of information" denotes the nature of the initial information presented in the experiment, with category one indicating true information and category two indicating false information. Analysis from Table 2 indicates that the *veracity of information* does not significantly influence user interaction and believability. This suggests that beyond the point of critical information load, the initial information serves as the foundation for belief formation, acting as pre-existing beliefs. When subjects are exposed to false information, their beliefs align with this misinformation, highlighting the potential for false and misleading information to shape user perceptions.

Theoretical Implications

From a theoretical standpoint, results suggests that the sheer volume of information available exacerbates initial opinion formation and confirmation bias tendencies, prompting users to selectively engage with content that reinforces their existing beliefs. This finding validates arguments based on cognitive response theory, which posits that individuals' information processing is influenced by their pre-existing beliefs and cognitive biases.

This observation aligns with cognitive load theory (Plass et al., 2010), which suggests that when users encounter relevant information on social media, they allocate cognitive resources to process it. In a fast-paced digital environment, this leads to rapid opinion formation. As users find information that supports their initial views, they allocate resources selectively, reinforcing these beliefs. Cognitive load theory thus sheds light on how cognitive processing shapes information consumption and opinion formation on social media.

After consuming five information points, subjects presented with information contradicting the initial five points experienced a notable reduction in both interaction and believability with the sixth information point compared to subjects consuming information aligned with the initial set. This underscores the notion that the initial five information points serve as pre-existing beliefs regarding new information, emphasizing their pivotal role in shaping user opinions. This finding aligns closely with the arguments proposed by Miller (1967) which suggests that individuals encounter cognitive dissonance when confronted with information conflicting with their existing beliefs, and confirmation bias manifests when information aligns with pre-existing beliefs. Therefore, this result contributes significantly to the literature by validating this argument derived from cognitive response theory.

Practical Implication

Our study provides strong evidence supporting the existence of a "critical information point," identified here at five information points. This finding aligns with Kruger and Dunning's (1999) argument that individuals consuming information with limited cognitive resources may develop a false sense of expertise on a topic. In our experiments, users displayed greater trust in information from verified profiles (e.g., doctors with substantial followings) than from unverified profiles, such as lesser-known doctors and influencers. Notably, users showed higher belief in content from unverified doctors compared to influencers, reflecting a cognitive schema where "doctor" titles are perceived as highly credible.

Additionally, the veracity of information did not significantly alter user engagement or belief, emphasizing the powerful role of initial exposure in shaping user beliefs on social media. This suggests that social media platforms bear responsibility for fostering critical information processing to counter potential misinformation. Implementing verification for titles like "Dr." could further strengthen trust in content accuracy.

For users, these insights highlight the importance of recognizing potential biases in information processing, fostering more critical engagement. Digital identity management also emerges as crucial for content creators, as a trustworthy online presence can enhance the credibility and influence of shared information. Together, these findings underscore the role of digital identity and source credibility in shaping user engagement and opinions within the information ecosystem.

Opportunities for future work

As an initial step towards a broader understanding of human behavior within the IS context, our research offers valuable insights that can serve as a foundation for future investigations. Opening avenues for interdisciplinary research at the intersection of psychology, sociology, and information systems, enriching our understanding of the complexities inherent in online information consumption. From an experimental perspective, our research focused on utilizing information points from the medical domain and digital identities of sources within the medical field. However, this framework may undergo significant transformation when applied to other domains such as politics, entertainment, and sports.

Furthermore, while consuming information on social media platforms, users are exposed to a diverse array of content formats, including pictures, short and long videos, as well as meta-information such as comments. In our study, we specifically utilized pictures as information points; however, there exists potential to expand this investigation to different formats and combinations of information points commonly encountered on social media platforms. Future research endeavors could delve into investigating how users' schemas evolve in response to information consumption on social media platforms and the consequent impact on belief formation and decision-making processes.

Conclusion

This research highlights how information load influences opinion formation and belief development during social media use. Findings show that users begin to form opinions after consuming five content points on a topic, underscoring the importance of initial information in shaping user perceptions. This suggests the need for users to assess new information critically and for platforms to encourage responsible engagement. Additionally, the study reveals that trust in information is often influenced by the digital identity of the source, with users more likely to believe content from profiles featuring titles like "Dr." or "Doctor"—even if the profile is unverified with few followers. This calls for verification of professional titles to promote a safer, regulated online space. These findings offer valuable insights for Information Systems (IS) research, advancing our understanding of opinion formation within social media information consumption.

References

CHAN, T. K., LEE, Z. W., SKOUMPOPOULOU, D. & SITUMEANG, F. 2023. Judging the Wrongness of Firms in Social Media Firestorms: The Heuristic and Systematic Information Processing Perspective. *The Journal of the Association for Information Systems*.

CHEUNG, C. M.-Y., SIA, C.-L. & KUAN, K. K. 2012. Is this review believable? A study of factors affecting the credibility of online consumer reviews from an ELM perspective. *Journal of the Association for Information Systems,* 13**,** 2.

ELLEMERS, N. & HASLAM, S. A. 2012. Social identity theory. *Handbook of theories of social psychology,* 2**,** 379-398.

FESTINGER, L. 1962. A theory of cognitive dissonance, vol. 2. redwood. USA: Stanford University Press: Stanford, CA.

FOX, S. 1997. Situated learning theory versus traditional cognitive learning theory: Why management education should not ignore management learning. *Systems practice,* 10**,** 727-747.

HOGG, M. A. 2016. *Social identity theory*, Springer.

HU, P. J.-H., HU, H.-F. & FANG, X. 2017. Examining the mediating roles of cognitive load and performance outcomes in user satisfaction with a website. *MIS quarterly,* 41**,** 975-A11.

JAMES, T. L., LOWRY, P. B., WALLACE, L. & WARKENTIN, M. 2017. The effect of belongingness on obsessive-compulsive disorder in the use of online social networks. *Journal of Management Information Systems,* 34**,** 560-596.

JHA, A. K. & SHAH, S. 2021. Disconfirmation effect on online review credibility: An experimental analysis. *Decision Support Systems,* 145**,** 113519.

KAPOOR, K. K., TAMILMANI, K., RANA, N. P., PATIL, P., DWIVEDI, Y. K. & NERUR, S. 2018. Advances in social media research: Past, present and future. *Information Systems Frontiers,* 20**,** 531-558.

KELTON, K., FLEISCHMANN, K. R. & WALLACE, W. A. 2008. Trust in digital information. *Journal of the American Society for Information Science and Technology,* 59**,** 363-374.

KIM, A. & DENNIS, A. R. 2019. Says who? The effects of presentation format and source rating on fake news in social media. *Mis quarterly,* 43**,** 1025-1039.

KIM, A., MORAVEC, P. L. & DENNIS, A. R. 2019. Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems,* 36**,** 931-968.

KLAYMAN, J. & HA, Y.-W. 1987. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review,* 94**,** 211.

KREPS, D. 2010. My social networking profile: copy, resemblance, or simulacrum? A poststructuralist interpretation of social information systems. *European Journal of Information Systems,* 19**,** 104-115.

KRUGER, J. & DUNNING, D. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology,* 77**,** 1121.

LAATO, S., ISLAM, A. N., ISLAM, M. N. & WHELAN, E. 2020. What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? *European journal of information systems,* 29**,** 288-305.

MCVEE, M. B., DUNSMORE, K. & GAVELEK, J. R. 2005. Schema theory revisited. *Review of educational research,* 75**,** 531-566.

MILLER, G. A. 1967. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychology of Communication.* Penguin Books.

MISHRA, A. N., ANDERSON, C., ANGST, C. M. & AGARWAL, R. 2012. Electronic health records assimilation and physician identity evolution: An identity theory perspective. *Information Systems Research,* 23**,** 738-760.

MODGIL, S., SINGH, R. K., GUPTA, S. & DENNEHY, D. 2021. A confirmation bias view on social media induced polarisation during Covid-19. *Information Systems Frontiers***,** 1-25.

MORAVEC, P. L., KIM, A. & DENNIS, A. R. 2020. Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research,* 31**,** 987-1006.

MORAVEC, P. L., KIM, A., DENNIS, A. R. & MINAS, R. K. 2022. Do you really know if it's true? How asking users to rate stories affects belief in fake news on social media. *Information Systems Research,* 33**,** 887-907.

PLASS, J. L., MORENO, R. & BRÜNKEN, R. 2010. Cognitive load theory.

SWELLER, J. 2011. Cognitive load theory. *Psychology of learning and motivation.* Elsevier.

WESTERWICK, A., JOHNSON, B. K. & KNOBLOCH-WESTERWICK, S. 2017. Confirmation biases in selective exposure to political online information: Source bias vs. content bias. *Communication Monographs,* 84**,** 343-364.

Advancing Accountability on ICT Platforms
to Navigate AI Integration in S-T Systems
A New Paradigm for Interplay of Accountability and Interpretation

Gianni Jacucci
Department of Information Engineering and Computer Science, University of Trento, Italy

and

Mike Martin
Newcastle Business School, Northumbria University, UK

Completed Research

Abstract

In relational service or enterprise environments, involving multiple agencies and diverse cultural contexts, effective change management becomes complex, demanding a higher-level framework based on second-order cybernetic concepts such as principles, objectives, roles, responsibilities, intentions, and awareness. These concepts are crucial for ensuring accountability in human actions which go far beyond the basic functional data typically handled by institutional ICT platforms. Without an updated information paradigm, current ICT platforms remain inadequate for addressing accountability questions in complex socio-technical settings—whether analysed by humans or Trained Generative AI modules. In 2020, Martin and Wilson introduced a new, Information Communications (IC) paradigm designed for ICT platforms which addresses this problem. This paper utilizes IC and a conversation model to explore connections between AI and accountability and presents a resulting framework for second-order governance.. In collaboration with colleagues, we are developing a prototype conversational platform that enables second-order communications within information systems in ICT.

Keywords: AI, Accountability, Relational Service and Enterprise, ICT Information Paradigm, Second Order Cybernetics, Information on Communications.

> *"En route, le mieux c'est de se perdre.*
> *Lorsqu'on s'égare, les projets font place aux surprises*
> *et c'est alors, mais alors seulement que le voyage commence."*
> Nicolas Bouvier (1929 - 1998)

> *"Flectere si nequeo Superos, Acheronta movebo."*
> Virgilio, Eneide

1 Introduction

1.1 The Challenge

This paper explores the relationship between AI and Accountability in complex, multi-agency environments, such as multi-agency health and wellbeing systems and inter-organisational enterprise settings. Here, Accountability refers to the ability to allocate credit or blame for the outcomes of complex human activities particularly in periods of change. Managing change effectively in these environments, which often involve multiple peer stakeholders and diverse cultures, requires, in addition to first order functional data about human actions, a higher-level framework of second-order concepts, such as principles, objectives, intentions, and the awarenesses of actors performing change or communication acts. Governance, here, involves not only what and how questions but also why and with what motivation.

Second order information remains either implicit in current ICT environments or is partially captured in the external legal and commercial information of conditions of use, licenses and service level agreements. Without explicit and adequately represented second-order data, neither humans nor AI systems can adequately answer accountability questions in these complex settings.

Present powerful Trained Generative (TG) AI systems are based on Deep Learning (DL) models (Shrestha & Mahmood 2019) and/or Large Language Models (LLMs) (Ichien *et al.* 2024), that can take raw data and "learn" to generate statistically probable outputs when prompted. Obviously these systems cannot invent or take into account cybernetic second order data, that has not been provided to them.

Moreover, lacking both the personal subjective culture of human experience, as well as hermeneutic interpretation capabilities, they are ill-equipped to provide meaningful insights into human actions. AI DL systems, and LLMs, characterised by machine learning mechanisms based on the probabilistic evaluation of proposed massive data, extract meaning from facts of the world in a way that is completely different from human individual sedimentation, or socio-cultural amalgamation and homogenisation and does not causal causal links specific to the human sense making of intentionality. The path to using AI for Accountability is, therefore, fraught with challenges.

Recently, Augmented Intelligence might offer a promising option. It focusses on the enhancement of human capabilities through the collaborative efforts of humans and AI systems (Harfouche A. *et al.* 2023a); it creates *human-AI pairs* in which both elements may contribute towards interpretation.

1.2 Analysis of the problem specification in the IS Literature

Whatever their nature, AI routines can be used either stand alone, or in support to human activity (Harfouche A. *et al.* 2023a). In trying to understand data, they can be oriented by human knowledge and interpretation (Johnson 2022). They can also be equipped with some explanatory ability about their functioning and results (Harfouche L. *et al.* 2023b) Accountability of AI routines is normally intended as their eventual ability to account for the results they generate, that is to say, how or on what basis, they have produced their performance.

We address here a different kind of Accountability, with respect to AI systems: their eventual ability, when confronted with a given human socio cultural context, to account for the consequences of the human actions described therein: why, and with what consequences they have produced. Here, Accountability requires the availability of two necessary ingredients (Martin 2024a): a) the availability of a certain kind of data - *of cybernetic second order data like roles, responsibilities, intention, awareness, needed for accountability and second order governance* (Martin 2024a) -; and, b) the availability of *the ability of the hermeneutic interpretation of that data within the appropriate socio-cultural context implying a basic commitment to mutual understanding in human communicative actions* (Habermas1984) -. Both availabilities are questionable for current AI systems operating on current ICT platforms.

Current institutional information infrastructures operate under the now four-decade-old information paradigm of Data Processing and Distribution (DPD). DPD is limited to cybernetic first order functional data. (Martin and Wilson 2020) Here, the type of Accountability that we wish to pursue in socio technical contexts is not within reach: second order data are missing. Therefore, applying AI routines here will not work. And it will not work also because AI routines do not possess, as yet, personal culture and hermeneutic interpretation abilities. (Montagnani *et al.* 2021, Ichien *et al.* 2024, Skylar 2024)

In this paper we intend to explore the application of a recently proposed, innovative, structured communications, called Information Communications (IC) paradigm (Martin and Wilson 2020), to be introduced in institutional information infrastructure, capable of handling human conversations accessing and elaborating second order data, governance, and accountability. At the same time, we intend to reason on which type of current AI routines are best suited to pursue the performance of certain hermeneutic interpretation abilities.

1.3 More detailed analysis of motivation and objectives

Motivation: The phrases "Don't rush to automate incorrect processes" and "Garbage in, garbage out" are commonly cited warnings when discussing technological advancements. These cautions are particularly relevant in the push to integrate AI with natural language generation (NLG). Blind reliance on AI can lead to an expectation of insightful responses about human behaviour in the absence of the necessary input information or consideration of the system's interpretive limitations.

To analyse the connection between AI and the kind of Accountability we are interested in, consider the types of questions that need to be answered:

- *What happened, and what are the interpreted causal linkages?*
- *Who knew what, when, and what are the interpreted motivational linkages?*
- *Who informed whom, for what purpose, and with what consequences, and what are the conversational linkages and contexts?*

The first part of each question is empirical, while the second requires judgment and interpretation. The system in which such processes occur must be socio-culturally aware as well as socio-technically informed.

When assessing AI models, two critical factors must be understood: how the model functions and the data it has been provided. For those looking to use AI to enhance accountability in enterprise transformations, it is crucial that AI models operate at a cybernetic level beyond basic first order data processing and distribution and possess the interpretive abilities needed to accurately understand complex situations. So, we need characterize the current state of these aspects, and propose, in particular, a solution for accessing second-order governance through a model for human interaction, to be applied within Martin and Wilson's IC paradigm. At the same time, cultural hermeneutic interpretation affordance of different AI approaches needs to be addressed by examining human participation in structured *AI-human collaborative pairs*; as the adequacy for the task of current standalone AI routines appears, as yet, questionable.

Objectives:

This is a reflections paper. Addressing in particular two relevant concepts: 1) interpretation, and 2) cybernetics, in considering the adequacy of structural and infrastructural approaches to the complexity of emergent and evolving social service and enterprise contexts; in particular related to the recently started transition, from simply *transactional* to fully *relational* approaches to service and enterprise and to the infrastructures capable of supporting them. (Bartels & Turnbull 2022, Martin *et al.* 2024, Martin 2024a, Martin 2024b, Wilson *et al.* 2024)

Our approach starts with 1), reflecting on the fact that human interpretation of experience is both subjective (individual) and cultural (social). It continues by 2), considering the higher order cybernetics aspects of purposeful human actions: what is referred to, in everyday language, as 'meta-level thinking'. It then revisits, with cybernetic data analysis, the information paradigm applied in information platforms, pointing out limitations of the current data processing and distribution paradigm of present institutional information platforms.

These elements are crucial in examining the affordance of present AI routines in tackling accountability issues, given their difficulty with (subjective and cultural) interpretation, and given the cybernetics of the information paradigm they access, presently limited to first order functional data.

1.4 Challenges in Integrating AI and Accountability in ST contexts

Because of their features, AI as well as other emerging digital technologies, suffer the well-known issues of lack of accountability and transparency (Montagnani *et al.* 2021). From the perspective constructed by the reflections in this paper, we question the ability of AI systems, when confronted with a given human socio cultural context, to account for the consequences of the human actions described therein.

Despite the exceptional performance of large language models (LLMs) on a wide range of tasks involving natural language processing and reasoning, there has been sharp disagreement as to whether their abilities extend to more creative human abilities. A core example is the interpretation of novel metaphors: "the metaphors an LLM might generate are limited to those that human writers have already formed and planted into texts, thereby making humanity's store of metaphors available to be mined by LLMs". (Ichien *et al.* 2024). More: in a recent study of behavioural differences between expert humans and language models in wargame simulations, the LLM simulations cannot account for human player characteristics, showing no significant difference even for extreme traits, such as "pacifist" or "aggressive sociopath." (Skylar 2024)

As we have outlined, there are two main challenges in linking AI with Accountability in service and enterprise:

- **Challenge 1: Insufficient Data in Current ICT Platforms**: The data within current enterprise ICT platforms is inadequate for meeting accountability requirements, as it is based on data processing paradigms focused only on basic functional data.

- **Challenge 2: The Interpretive Nature of Accountability**: Accountability is inherently interpretive, making it difficult for AI systems, which lack these interpretive capabilities, to be effective.

These challenges are of core interest across many disciplines, from law to economics, medicine, science, and education. We refer to recent reviews on explainable AI systems (Mohseni et al. 2021), validation methods for AI systems (Myllyaho et al. 2021), evaluating explainable AI (Nauta et al. 2023), and NLG system evaluation (Sai et al. 2022). For example, in the legal field (Faggella 2021), AI can assist lawyers with due diligence and research, but it cannot replace the judgment required of a judge or jury. This highlights the limitations of AI in contexts requiring deep interpretation and judgment.

In this paper, we will explore how to overcome Challenge 1 by identifying the contextual information that should be integrated into AI systems to address accountability issues. This is achieved by examining how humans analyse purposeful human action, and retaining the kind of data they employ in doing it. To this end, it is useful to delve into four significant approaches to organizational interventions in complex socio-technical settings by putting one close to the other (Ciborra 1994, Jacucci 2007, Martin et al. 2009, Bednar and Welch 2014), emphasizing the importance of second-order cybernetic concepts in managing change (Jacucci 2024b).

1.5 Structure of the Paper

In the following sections, we address accountability, in the sense we have described above. For this we recall theory considerations on interpretation and analysis of human purposeful behaviour, in areas impacting subjectivity in human communication: cybernetic orders as the path to explicitly address roles and responsibilities. On the base of cybernetic concepts, we will discuss the evolution of the IC paradigm proposed by Martin and Wilson (2020), illustrating a conversation model that transcends the limitations of data processing and distribution (DPD). Then we consider interpretation in the context of different kinds of AI approaches.

Explicitly:

- **Section 2**: Theory of Interpretation and Cybernetics, as central concepts in Accountability
  - 2.1: Transition from transactional to relational of structural and infrastructural service and enterprise

    - 2.2: The meta-level: Interpretation requires parsing Cybernetic orders in purposeful human action
    - 2.3: In socio-technical contexts and systems Interpretation requires parsing “epistemic registers”
- **Section 3**: The new, conversational, Information Communications (IC) paradigm
- **Section 4**: Accountability in the new conversational IC paradigm
- **Section 5**: A model for purposeful human interaction capable of handling Accountability
- **Section 6**: Questioning the ability of different AI approaches to address Accountability issues

2 Theory of Interpretation and Cybernetics, as concepts of central importance in Accountability

Social theory helps us reflect on what human sense making is about. First of all, interpretation is subjective, and cultural (Langefors 1966, 1995; Schutz, in Wagner 1998). Secondly, phenomenological observation on human experience unveils that, besides facts, intentions matter, as do expectations, responsibilities, awarenesses and feelings: all second (or higher) order concepts involved in interpretation of social events.

In the following we start by recalling theoretical considerations appropriate to the complex social settings of *relational* service and enterprise, contrasted to the simpler case of *transactional* ones. (Martin *et al.* 2024) We consider this distinction explicitly, to capture the inherent complexity of the social interaction experience that we intend to take into account.

The key characteristic of *relationality* is that of direct and authentic human experience, requiring both human knowledgeable participation, and evaluation of the very extant experience. Evaluation is unavoidable, because authentic, communicative human actions are committed to mutual understanding. This fact links interpretation of second order information to communicative actions: the inherent complexity of relational experience in social interaction.

The traditional, *transactional* approach to social reality of service and enterprise and supporting infrastructure, represents an intrinsic simplification of human interaction. It reduces all structural and infrastructural system behaviours to measurable pre-conditions, (inter)activities and post-conditions. The respective information architectures and methods, which are mandated in procurement, and embedded in commercial information systems supply, demand that relationship processes must be reduced to clearly defined use-cases and business logics, recorded in pre-designed forms and data structures that are fixed and immutable in use and in practice. There is no concept of relationality that can be articulated and inscribed in management or information systems that are constructed within this paradigm. (Martin *et al.* 2024)

2.1 Towards social complexity: the Transition from *transactional* to *relational* in structural and infrastructural service and enterprise

Considerations of relationality - illuminating for the present endeavour - characterise *relational* service and enterprise and their technical infrastructures (Martin 2024a & 2024b, Martin *et al.* 2024): relations evolve, involve learning; relations are experiential, involving interpretation and judgement. Relational service and enterprise and their infrastructure therefore evolve, and involve ongoing interpretation. Here relational governance and human learning are (re)-istitutionalised in public service and enterprise infrastructure: this is a first important result of this analysis (Martin 2024a) followed here below by other four important results.

The next two foundational steps, linking interpretation to communicative actions, and cybernetic second order information to accountability, are performed with eminent theoretical inputs:

1) From Krippendorff's analysis – a cybernetics theorist - of the determinability of generic systems, relationality, as a socio-cultural phenomenon, is characterised as being *constitutively determinable* (Krippendorff 2009); this implies that:
    a. users must take part in its creation and evolution;
    b. being experiential – not just observable data -, it must include second-order cybernetic concepts and logics.

In step 1, based on Krippendorff's distinction, a disciplinary cross-over is required: the plan changes from social relationality to the informational and behavioural aspects of purposeful human activity. This shift in epistemic register from the conversational to informatics represents a second important result, exposing the reason why relationality demands a shift in information systems paradigm.

2) From Habermas definitions, authentic, communicative human actions are committed to mutual understanding. (Habermas 1984)

From step 2, we deduce that conversations must comprehend conversations about conversations, i.e., second order governance and accountability which represents the third important result.

We now link this to the last, socio-cultural epistemic register (Martin 2024a & 2024b, Martin *et al.* 2024): Relational socio-technical systems must live in, and co-evolution with, their socio-cultural contexts. They cannot be pre-designed, but must emerge and grow from the appropriation of a simple primordial system of "conversations about missions and purposes". This represents the initiation and emergence of the operation and governance of human enterprise as communities of interest and practice. This is the fourth important result.

Now, a fifth and final important result: in the contexts we are considering, the information technology supply relationship changes: from "solution" design and the development of commercial infrastructure to enabling and supporting user communities in learning and structuring their own infrastructure. (Martin 2024a & 2024b, Martin *et al.* 2024)

We see from the analysis of relational social contexts, that interpretation and cybernetics are essential ingredients for accountability and second order governance. We shall now go back to the theoretical roots of these two central concepts here, to gain further insight on them.

2.2 The meta-level: Interpretation requires parsing Cybernetic orders in purposeful human action (Martin and Wilson 2009)

To explore in more detail which contextual information should be integrated, such as actor roles and responsibilities, and other second-order concepts, we can examine how humans reason about change management and identify the necessary information for these tasks. Therefore, in Section 2,1 we have referred readers to the description (in Jacucci 2024b) of four significant approaches, conceived for organizational interventions in socio-technical arenas (Ciborra & Lanzara 1994; Jacucci 2007, Cattani & Jacucci 2007, Jacucci *et al.* 2007; Martin *et al.* 2009; Bednar and Welch 2014), addressing change management in complex enterprise settings, including multicultural contexts involving multiple agencies. All these approaches are related among themselves, in philosophical basis and practical objectives, enhancing aspects of *relating, communication, participation, and learning,* beyond achieving *functional rationality.* They all engage with second-order cybernetic concepts, embracing deutero (Bateson 1973) or double-loop (Argyris and Schön 1974) learning, advocating for therapeutic co-construction in order to achieve it in each and every social situation and context.

Let us see more closely the epistemic implications of the cybernetic second order, central to Martin and Wilson's Social Informatics Intervention (2009). In our familiar, first order, rational development process, we traverse phases of vision, plan, execution, and evaluation:

- Purposeful behaviour starts (logically, in post-hock rationalisation) with the conception of a vision of the desired state of affairs.
- Next we must construct a plan based on what we believe is possible and effective, this is strategy.
- This leads to the execution of the plan which involves deploying and consuming resources that are available and appropriate.
- The evaluation of our progress in relation to the plan and the continued relevance of the plan to our vision involves comparing observations, measurements and the use of appropriate criteria
- And this results in learning and the conception of new visions where learning involved the deepening and broadening of our knowledge. It is cumulative.

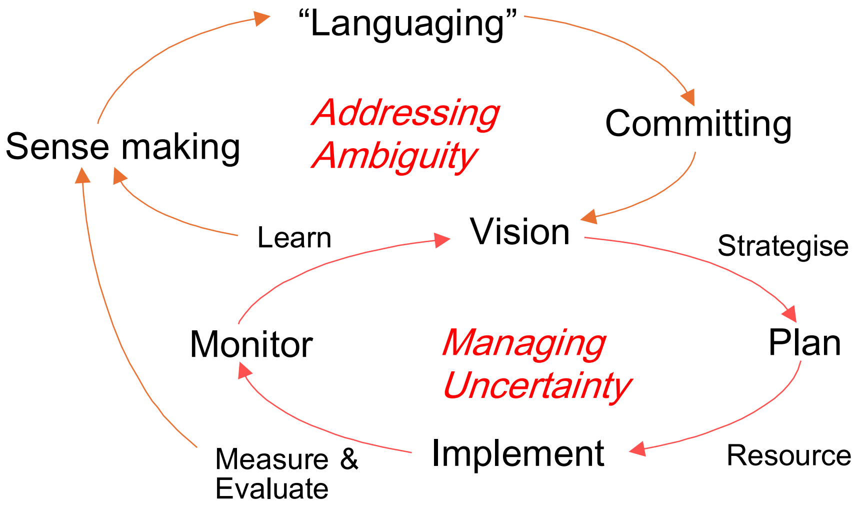


Fig. 1. The second order model of the enterprise (from Martin et al. 2009)

The purpose of this process is to manage uncertainty which represents risk and it is the basis for all of our standard project management approaches. However, every so often something different happens and, when we look back, what we see is that we have started doing things that we previously thought impossible and have stopped doing things that we thought were essential. What we once thought of as our resources have become impediments and what we thought were barriers are now opportunities. We have re-evaluated what we need to evaluates and what we nmean by evaluation and our learning has involved forgetting (!).

We cannot account for these changes within our first order loop; somehow we must have broken out of it. What seems to have happened is that contractions, inconsistencies and paradoxes, as well as discoveries, have built to a point where we have been forced into a different mode of sense making. This is equivalent to entering the inter-subjective mode of conversation where we open ourselves up to the co-construction of new meanings and values. One of the signs that this is taking place is that we start adopting new terms and usages, this is languaging. For the outsider this often appears strange and threatening and is dismissed as jargon but, for the participants, it leads to new commitments and new shared visions. The purpose of this second order loop is the handling of ambiguity rather than the management of uncertainty.

2.3 In socio-technical contexts and systems, Interpretation requires parsing “epistemic registers” (Martin, Welsh and Wilson 2020)

How do humans reflect and reason about their activities, in socio technical contexts? The enterprise systems paradigm and methods, within which we build and deploy integrated solutions, which has evolved over the last four decades, is based on the assumption of a clear demarcation between an inside and an outside. Within, we assume a set of operational norms and expectations which are coherent with the purposes and objectives of the enterprise. Any contradictions or problematic deviations are assumed to be faults which will be rectified by means of internal control mechanisms. As a consequence of these assumptions of coherence and rationality, we further assume that the requirements on the system can be fully expressed in terms of use cases and business logics, that is to say, purely in terms of functional behaviours. Such systems are defined by what they do, their purposes remain implicit because they are taken to be obvious and given, completely expressed and embodied in explicit rules, logics and procedures. The concepts and language associated with the acquisition, processing, distribution and storing of data can express all that needs to be expressed about these systems. When we consider inter-organisational systems, where the relationships between the members are transactional and delimited by explicit contracts and protocols, these assumptions remain more or less valid. Thus, supply chain and customer relationship management systems, and the like, are defined and implemented within the data processing and distribution (DPD) paradigm we have outlined.

When the relationships supported by inter-organisational information systems becomes more relational and, as a consequence, less predefined and predictable, the DPD systems paradigm begins to exhibit some limitations. Note that we are now talking about the nature of the relationships between the organizations themselves which are supported by an inter-organisational information system; this has two very significant implications: Firstly, we have to consider issues of:

- infrastructural capacities, by which we mean shared reusable and, indeed, re-purposable resources,
- the structural systems that makes use of infrastructure to communicate and manage information to coordinate and interwork. and
- super-structural systems that govern this use and communication.

In the face of these horizontal distinctions, and this complexity, we can no longer assume that purpose and intention can remain implicit. We see that the original, central, inside / outside cut, is here replaced by the meta-level type, above / below cuts.

The second and related implication is that, with this shift to this more inclusive view of inter-organizational systems, we have transitioned from a technical to a socio-technical conception of our subject. It is clear that the complexity and ubiquity of automation and information technologies has changed radically since the original mechanization context in which Emery and Trist originally coined the term “socio-technical” in the mid twentieth century. Even the developments of the earlier phases of informatization of the economy and of some aspects of wider society in the 1970s and ’80 through separate developments in the information processing, tele-communications and mass communication/media sectors have now been radically superseded, so it is important that we re-establish a concept of the socio-technical which is able to take into account the complexities of our current, and foreseeable, situation. To do this we must go back to basic principles.

Socio-Cultural View	Individual and Collective Values and Principles	New meanings and values come into being
Conversational View	Roles, relationships and responsibilities	Meanings include intentions
Informatics View	Codes, terms and objects	Meanings are pre-defined and concrete
Engineering View	Bits – terra-bytes, channels and bandwidth	Measurements but no meanings

Table 1. The four epistemic registers (Martin 2024a).

The epistemic stances required to deal with a world of objects and mechanisms is not the same or even commensurable with one required to deal with a world of conversational relationships (Martin 2020; Martin and Wilson 2020). The former are handled in the DPD we have discussed, the latter are not. The DPD paradigm requires augmentation to one of Information Communications (IC) in which information is generated and interpreted by entities that are defined by their purposes and intentions rather than simply by their functions, generating the requirement for second order governance and accountability.

3 The new, conversational, Information Communications (IC) paradigm (Martin and Wilson 2020) on information platforms

One of the main context of multi-organizational systems' construction and deployment, which has provided the context for the development of these concepts of the neo-socio- technical (Martin and Wilson 2020), has been the planning, coordination and delivery of health and social care in communities. In particular it has been concerned with how these systems respond to complex, long term conditions that involve multiple problems and pathways.

The complexities of these contexts and the failure of conventional Data Processing and Distribution (DPD) paradigm approaches, such as the development of shared electronic records at the national or regional level and attempts to develop joint assessments of need across different organizational and care settings, have resulted in critiques of this approach. We have examined the alternative approach proposed by Mike Martin and Rob Wilson (2020) based on what they have called the Information Communications (IC) paradigm which implies the operation of both first and second order governances, and the regulation of the relationships between them.

4 Accountability in the new conversational IC paradigm

It is because of the complexities and ambiguities which are inherent in the domain of practice that is being considered, that Martin and Wilson had to abandon the structural approaches, based on the pre-defined pathways, processes and protocols of DPD systems, and adopt the infrastructural approach of IC services. This shift recognises the need to be explicit about the conversations and relationships between role holders, who are defined by their shared intentions and commitments, that is being supported rather than limiting the definition of the system to sets of predefined pathways or process logics.

The core concept of the ensuing architectural discourse of the neo-socio-technical is that of *conversation*. This is in contrast with the conventional starting point of business process or function. This change is necessary because the concept of conversation implies and includes the fundamental link between intention and extension, of doing something, on the one hand, and of meaning and intending something, on the other. This is an essential aspect of the social and, therefore of the socio-technical, which is characterised by networks of purposeful interactions. It also introduces the implication of the inter-subjectivity of conversationalists and of the means of incorporating the normative as well as the descriptive into the processes of specification and design.

In the next Section we address an initial model of conversation as purposeful interaction. We shall have here the occasion to show, in a worked out example, the significance and expressive power of epistemic registers.

5 A model for purposeful human interaction capable of handling Accountability

In an initial model of conversation as purposeful interaction, an *actor* is something that is capable of behaving. The range of possible behaviours that an actor can perform is defined in terms of a repertoire of *actions* or operations. We can depict the entities and relationships in a diagram representing *purposeful action*.

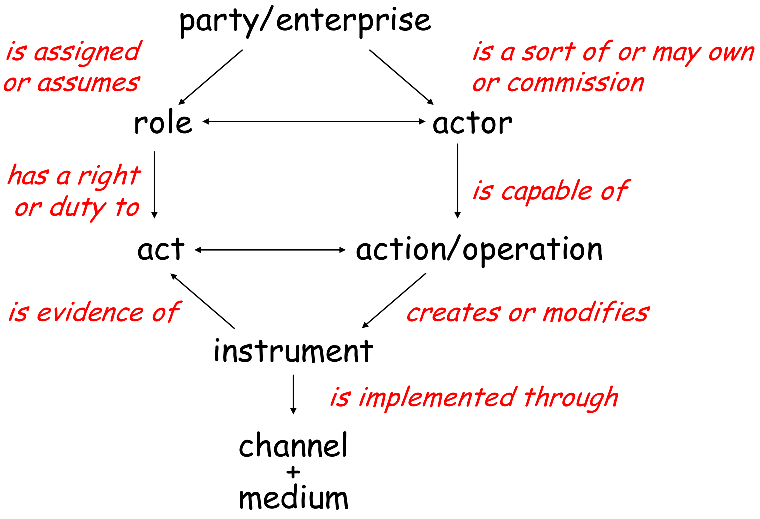


Figure 2. Purposeful action.

The next stage in the development of the model of conversation involves the extension of purposeful action to *purposeful interaction*. We achieve this by inking two purposeful actors producing the conversational equivalent of Shannon and Weavers transmitter-receiver diagram.

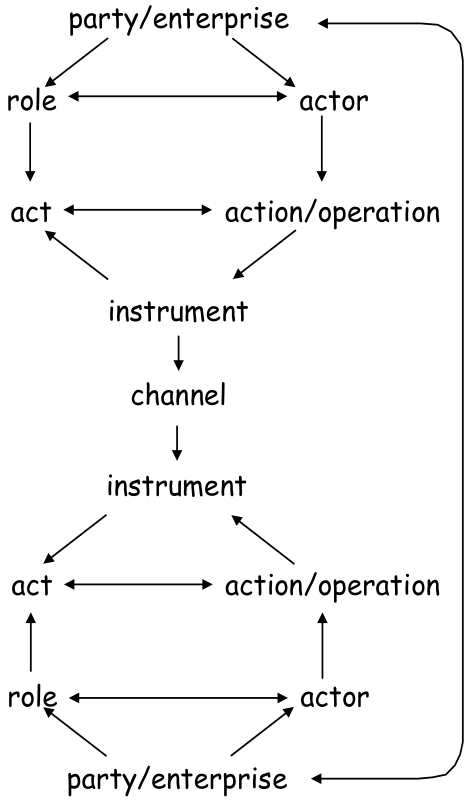


Figure 3. Purposeful interaction.

In the context of partnership based community care and complex long term conditions, the term "constellation of care" has emerged to represent the group of different formal or informal care relationships that are currently relevant and involved with the subject. These are represented by Roles C, D, E, and F in Fig. 4 below.

It is clear that, while the members of a care partnership have shared intentions, they retain their individual identities and must have their own distinct relationships with the subject of care. To account for these particular mutualities between carers the exchanges between their respective systems must correspond to conversations. They cannot be safely reduced to the data exchanges of the DPD paradigm.

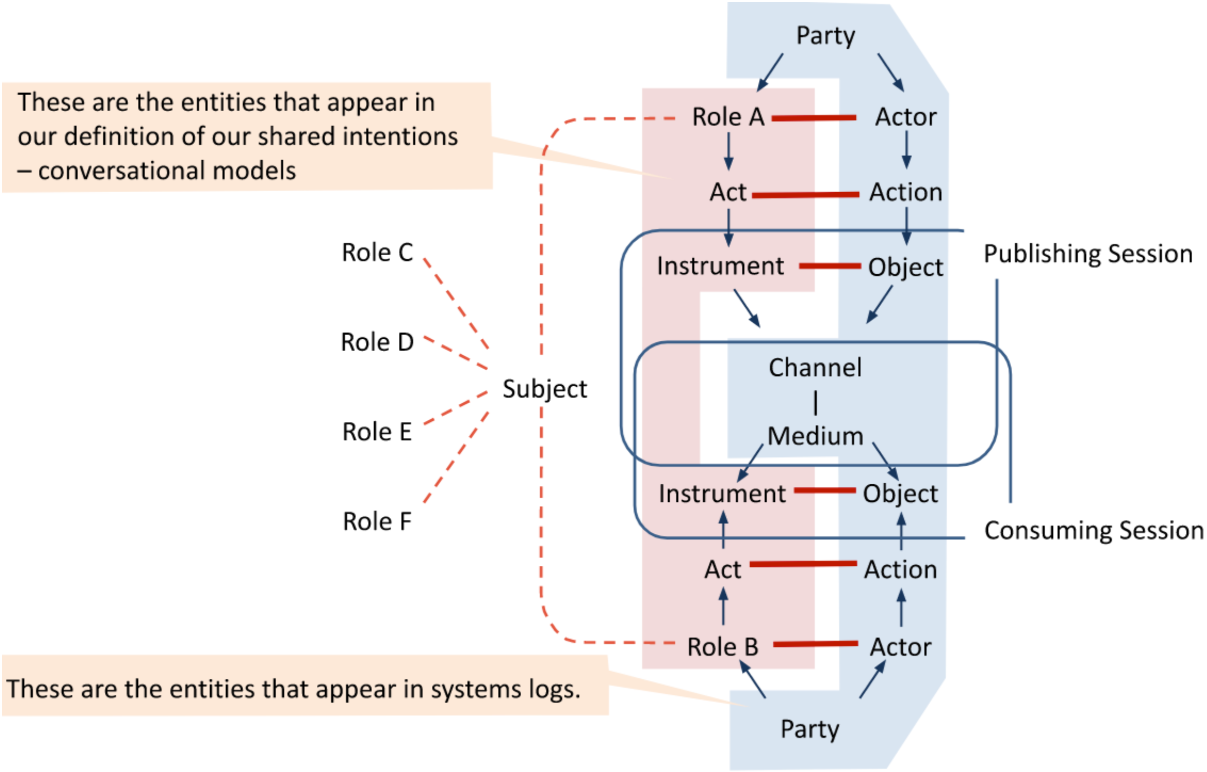


Fig 4. **Deconstructing the purposeful interaction.**

Returning to the model of purposeful interaction, Fig. 4, we observe that it contains two categories of entity: the "blue stuff" (for the digital version showing colours) at the right, which corresponds to the concrete extensional entities and events which we detect in the operation of the system and which will be captured and appear in the systems logs. The other category is the "pink stuff", at the left, which are the concepts that appear in our definitions and specifications of the norms and intentions of the system at the conversational level. For example, the term "Doctor" in the blue representation is an individual person with certain accredited capabilities performing certain actions. In the pink representation, clinician is a role defined in terms of a set of responsibilities to, and relationships with, other roles.

To summarise at this stage in the argument: our contention is that, in multi-organisational systems of complex care, intentionalities have to be made explicit so that the responsibilities and processes of governance can be supported. These involve asking and answering the questions: "Is this what we intended? Do we still intend this?" where the evidence corresponds to the "blue stuff" which can be found in the system logs and the criteria by which it is interpreted and judged corresponds to the “pink stuff” of Fig. 3 which, as provenance, is the explicit record of the shared intentions and ethos of the care community.

6 Questioning the ability of different AI approaches to address Accountability issues

Again, current AI routines, appear to lack hermeneutic interpretation capabilities. Let us consider this point in more detail, by considering explicitly the different characters of various, current AI approaches, with particular attention to those involving a human companion to help put in context AI operations.

Within qualitative research in human sciences, there are three the main epistemological stances, underpinning different philosophical approaches and ensuing respective methodologies: empiricism, hermeneutics, and phenomenology. (Jacucci 2023, Jacucci 2024a) Although, at the most abstract level, all three share the same objective: to extract disciplinary meanings from human recounts of worldly facts, each one of them exhibits a different, specific vulnerability and shortcoming in the results it generates. Respectively: the incipient theoretical interpretation in e.g. practice based theorising, of the empirical stance, the outstanding choice of interpretation of the hermeneutic stance, and the adherence to reality of facts in the conscience of the phenomenological stance. (Jacucci 2023, Jacucci 2024a) I making this meta-methodological analysis, we must observe that the three approaches do different sorts of work and, from within any of them, what the others undertakes does not count as work. The architectural approach which has been referenced as the background to the considerations of this paper, recognises that, in the architectural discourse of the socio-technical system in co-evolution with its techno-socio-cultural context, we cannot remain within a single, narrow epistemic register but must travers all of them: engineering, interpretations and enculturation are all involved.

By reflecting on the possibility of extension to AI of our considerations about such human qualitative research methods, we can try to characterise the ability of different AI approaches in addressing Accountability issues, by discerning the specific epistemology which is closer to the characteristics of each AI approach, and assigning to the approach the vulnerability and shortcoming specific of that epistemology. We provide hints for directions of future work, in linking different AI architectural styles to the different methods of the three foundational philosophies.

6.1 Hints for Empiricism

Empiricism vulnerability: the arbitrary choice of the initial practice based theory.

This is the slot for AI applied to natural science type cases, where "facts in the conscience" are not relevant. The AI literature is mostly filled with this type of cases. But we are interested in the life of organisations made up of people. If we seek to identify empiric type cases where subjectivity is relevant, we should look to AI DL algorithms and architectures (Shrestha *et al.* 2019) and also to other architectures, searching for analogies with an empiric method, e.g., the practice based theorising method of *empiricist* grounded theory, to point to eventual biases, for a specific choice of initial theory to be built up. This concept corresponds to the context of asking questions to a chatbot like ChatGPT in the frame of a conversation. (Pries-Heje & Cranefield 2024)

6.2 Hints for Phenomenology

Phenomenology vulnerability: adherence to reality.

Traditional AI architectures may be anyway closer to the unbiased, virgin intuition of *descriptive phenomenology:* producing unbiased data. This corresponds to chatbots like ChatGPT, responding to repeated questions with different answers, as if they were data produced by a measuring instrument affected by a random error (Pries-Heje & Cranefield 2024). Different data, in principle all good, albeit in contrast. This creates the opportunity for the human operator of the AI system to combine answers, take mean values of data, or throw away unacceptable, unrealistic data.

6.3 Hints for Hermeneutics below

Hermeneutics vulnerability: the arbitrary choice of the "correct" interpretation.

The situation is a rather complex one, however, for the *hermeneutics* item. A new trend of Augmented Intelligence focusses on the enhancement of human decision-making through the collaborative efforts of humans and AI systems rather than AI as a replacement for human intelligence, aiming at a harmonious integration that amplifies human capabilities without supplanting them. According to Harfouche, Quinio, & Bugiotti (Harfouche A. 2023a) organisations are increasingly adopting collaborative decision-making frameworks that exploit the strengths of both humans and AI, leading to more informed and effective outcomes. So, the comparison we are trying to perform is not limited to human and machine learning, but includes mixed situations.

This is reminder of the fact that one may ask questions, to a chatbot like ChatGPT, in the context of a conversation in which subjective info has been added on the researchers objectives, and other issues, thus influencing or biasing or even addressing the answer of the AI system. (Pries-Heje & Cranefield 2024) Informed AI (IAI) (Johnson et al. 2022), exhibits the feeding by humans of additional information, to operate with a possibly more comprehensive awareness and integration of relevant data, context, and knowledge to make decisions or provide insights, i.e., with a specific interpretation. It resembles a *hermeneutic* approach, and it is biased by the choice of human interpretation. We start seeing light towards the ability of handling the concept of *accountability* of actions by human actors involved in the situations.

Furthermore, we are moving towards a Human-Centric Explainable AI system architecture, where Explainable AI (X-AI) refers to AI systems designed to provide insights into how and why they arrived at a particular decision or output. (Mohseni *et al.* 2021) The goal is to make AI's decision-making process transparent, understandable, and interpretable for humans. This approach, however, raises another layer of consideration of the distribution of responsibility between the human and what we might refer to as the "AI service" and the AI service provider producing. Considering AI as a service with a set of service definitions and responsibilities is applying the hermeneutic and phenomenological accountability considerations we have been considering to AI itself rather than limiting these considerations merely to the observation of performance, underlining the recursive complexity of this domain.

7 Conclusion

In this reflections paper, we note that the concept of accountability in AI raises significant questions that need thorough discussion, which have not yet been fully explored. Additionally, generative AI brings numerous challenges, and its cybernetic implications remain largely unexplored.

We have set out by addressing the role of two relevant concepts, interpretation and cybernetics, in emergent and evolving social service and enterprise contexts, e.g. related to the recently started transition from transactional to relational. (Martin 2024a, 2024b, Martin *et al.* 2024) We continue revisiting the higher order cybernetics aspects of purposeful human actions, and the limitations on this respect of the current data processing and distribution paradigm of present institutional information platforms. With these crucial elements we have examined the affordance of present AI routines and have observed their inadequacy in tackling accountability issues. To address the cybernetics aspect, we have observed that a new social interaction model, inserted in a more appropriate, innovative conversational informational paradigm, is required. For the interpretation aspect, we have considered the respective appropriateness of different AI approaches, as suitable candidates to support hermeneutic interpretation of roles and responsibilities, in considering actions and consequences in socio technical contexts. Given the inability of present AI routines to tackle interpretation on their own, we examine current approaches involving operational AI-human pairs, where human interpretation can help upgrade AI performance.

The key idea of our conclusion has two aspects: connecting AI with accountability in service and enterprise presents both a *challenge* and a *fundamental issue*. The *challenge* is that current institutional ICT platforms are inadequate for meeting accountability needs, because they rely on a data processing approach, limited to first order functional data. The *fundamental issue* is that accountability is inherently interpretive, meaning it is about understanding intentional actions and relationships, and it cannot be addressed by just automating the handling of behaviours and transactions ("Don't automate wrong processes").

The consequence of these considerations is that AI approaches cannot support inquiry and accountability as mere "automation." Instead, they need facilitating interpretive understanding. Furthermore, they require that we integrate beforehand advanced conversational paradigms into information and communication infrastructures, incorporating contextual information like roles, responsibilities, and other second-order concepts.

On the other hand, the capability of Interpretation by AI based systems, Informed AI (IAI) (Johnson et al. 2022), exhibits the feeding by humans of additional information, to operate with a more comprehensive awareness and integration of relevant data, context, and knowledge to make decisions or provide insights, i.e., with a specific interpretation. We start seeing light towards the ability of handling by AI systems the concept of *accountability* of actions by human actors involved in the situations. A direction to be closely followed from the perspective of using the power of Generative AI to support human accountability and second order governance in complex social, relational service and enterprise settings.

References

Argyris, C. and Schön, D. (1974) Theory in practice: Increasing professional effectiveness, San Francisco: Jossey-Bass.

Bartels, K., & Turnbull, N. (2020). Relational public administration: A synthesis and heuristic classification of relational approaches. Public Management Review, 22(9), 1324–1346. https://doi.org/10.1080/14719037.2019.1632921

Bateson, G. (1973). Steps to an Ecology of Mind. London: Paladin Books.

Bednar P. and Welch C. (2014) Contextual Inquiry and Socio-Technical Practice, In Kybernetes 43 (9/10).

Cattani, C., Jacucci, G. (2007) From software development service provider – helas, a captive resource! - to one's own products and brand. AIS eLibrary Proceedings of MCIS2006 in Venice, It: http://aisel.aisnet.org/mcis2007/18/.

Ciborra C. and Lanzara G. F. (1994) *Formative contexts and information technology: understanding the dynamics of innovation in organisations,* Accounting, Management and Information Technology 4, 2: 61-86; reprinted in "Bricolage, Care and Information: Claudio Ciborra's Legacy in Information Systems Research" by C. Avgerou, G. F. Lanzara and L. P. Willcocks (2009) Palgrave, Macmillan.

Faggella D. *AI in Law and Legal Practice – A Comprehensive View of 35 Current Applications.* Online resource, last updated on September 7, 2021. https://emerj.com/ai-sector-overviews/ai-in-law-legal-practice-current-applications/

Habermas, J. (1984) Theory of Communicative Action, trans. Thomas McCarthy, Boston: Beacon Press.

Harfouche, A., Quinio, B., and Bugiotti, F. (2023a) Human-Centric AI to Mitigate AI Biases: The Advent of Augmented Intelligence. Journal of Global Information Management, 2023, 31 (5), pp.1-23. 10.4018/JGIM.331755. hal-04263509 https://hal.science/hal-04263509

Harfouche, L., Nakhle, F., Harfouche, A. H., Sardella, O. G., Dart, E., and Jacobson, D. (2023b), A primer on artificial intelligence in plant digital phenomics: embarking on the data to insights journey, Feature Review Volume 28, ISSUE 2, P154-184 DOI:https://doi.org/10.1016/j.tplants.2022.08.021

Ichien, N., Stamenković, D. and Holyoak, K. J. (2024) Interpretation of Novel Literary Metaphors by Humans and GPT-4. In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), Proceedings of the 46th Annual Conference of the Cognitive Science Society.

Jacucci G. (2007) Social Practice Design, pathos, improvisation, mood, and bricolage: the Mediterranean way to make place for IT? AIS eLibrary Proceedings of MCIS2007 in Venice, Italy http://aisel.aisnet.org/mcis2007/19/.

Jacucci G., Tellioglu H., Wagner I. (2007) Design Games as a part of Social Practice Design: a case of employees elaborating organizational problems. Proceedings of MCIS2007 in Venice, and ECIS 2008 in Limerick, AIS eLibrary, http://aisel.aisnet.org/ecis2008/207/.

Jacucci G. (2023) Reflecting on Scientific Rigour in Socio-Technical Research, spotlighting Husserl's Phenomenology and Amedeo Giorgi's Descriptive Phenomenological Method. The 9th International Conference on Socio-Technical Perspectives in IS (STPIS'23) 27.–28.10.2023, Portsmouth, UK https://ceur-ws.org/Vol-3598/paper2.pdf

Jacucci G. (2024a) Applying Epistemology to AI in the Social Arena: Exploring Fundamental Considerations. CCPPE 2024 : International Conference on Continental Philosophy, Phenomenology and Existentialism. Nicosia, Cyprus. November 04-05, 2024 https://publications.waset.org/abstracts/189199/applying-epistemology-to-artificial-intelligence-in-the-social-arena-exploring-fundamental-considerations.

Jacucci G. (2024b) The SPD Story. Network resource, https://www.academia.edu/121776249/

Johnson, M., Albizri, A., Harfouche, A., and Fosso-Wamba, S. (2022) Integrating human knowledge into artificial intelligence for complex and ill-structured problems: Informed artificial intelligence, International Journal of Information Management, Volume 64, 102479.

Krippendorff, K. (2009) Ross Ashby's information theory: a bit of history, some solutions to problems, and what we face today. International Journal of General Systems, Vol 47, page 204, Published online: 30 Jan 2009 . https://www.researchgate.net/publication/49128560_Ross_Ashby%27s_Information_Theory_A_Bit_of_History_Some_Solutions_to_Problems_and_What_We_Face_Today

Langefors, B. (1966) Theoretical Analysis of Information Systems. Lund: Studentlitteratur.

Langefors, B. (1995) Essays on Infology - Summing up and Planning for the Future. Lund: Studentlitteratur.

Lave, J. and Wenger, E. (1991) Situated learning: Legitimate peripheral participation. Cambridge U. P.

Martin M., Walsh S., Wilson R. (2009) *A Social Informatics Intervention: theory, method and practice*" KITE Research Group, Newcastle University: http://www.woa.sistemacongressi.com/web/woa2009/papers/Martin_Walsh_Wilson.pdf

Martin M. (2020) "*Inter-organisational systems: a personal history*".UK Academy for InformationSystems Conference Proceedings 2020. 23. https://aisel.aisnet.org/ukais2020/23

Martin M. and Wilson R. (2020) "*Inter-organisational systems:a neo-socio-technical perspective. a neo-socio-technical perspective*.". UK Academy for Information Systems Conference Proceedings 2020. 22. https://aisel.aisnet.org/ukais2020/22.

Martin, M. (2024a) How and why relationality demands a shift in information systems paradigm. In 'Futures in Public Management: The emerging Relational approach to public services', R. Wilson *et al.* Emerald Publishing. Draft typescript preprint.

Martin, M. (2024b) The Trustworthy, Governable Platform: supporting accountability and governability in complex, multiparty enterprise. Submitted to ITAIS2024 conference.

Martin, M., Wilson, R., and Jamieson, D. (2024) Moving towards Relational Services – the role of digital service environments and platforms? Chapter 11 in S. Baines *et al*., CO-CREATION IN PUBLIC SERVICES FOR INNOVATION AND SOCIAL JUSTICE. Draft typescript preprint.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 11(3-4), 1-45.

Montagnani, Maria L. and Cavallo, Mirta (2021) "Liability and Emerging Digital Technologies: An EU Perspective," Notre Dame Journal of International & Comparative Law: Vol. 11 : Iss. 2 , Article 4. Available at: https://scholarship.law.nd.edu/ndjicl/vol11/iss2/4, p. 217

Myllyaho, L., Raatikainen, M., Männistö, T., Mikkonen, T., & Nurminen, J. K. (2021). Systematic literature review of validation methods for AI systems. Journal of Systems and Software, 181, 111050.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., ... & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Computing Surveys, 55(13s), 1-42.

Pries-Heje J., & Cranefield J. (2024) HOW GENERATIVE AI IS ALTERING IS PROJECT MAN- AGEMENT, IRIS 2024, Preprint

Sai, A. B., Mohankumar, A. K., & Khapra, M. M. (2022). A survey of evaluation metrics used for NLG systems. ACM Computing Surveys (CSUR), 55(2), 1-39.

Shrestha, A., and Mahmood, A. (2019) Review of Deep Learning Algorithms and Architectures, Digital Object Identifier 10.1109/ACCESS.2019.2912200.

Skylar Mastro, O. (2024) Human vs. Machine: Behavioral Differences between Expert Humans and Language Models in Wargame Simulations. Preprint https://doi.org/10.48550/arXiv.2403.03407

Wagner, H. (1970) editor: Alfred Schutz on Phenomenology and Social relations: U. of Chicago P., Chicago.

Wilson, R., French, M., Hesselgreaves, H., Lowe, T., & Smith, M. (2024). New development: Relational public services—reform and research agenda. Public Money &amp; Management, 1–6. https://doi.org/10.1080/09540962.2024.2344902

Parental paradox: navigating the tightrope between digital surveillance and children's social privacy

Marie Griffiths
Salford Business School

Oliver Kayas
Liverpool Business School

Rachel MacLean
Liverpool Screen School & Liverpool School of Art and Design

Abstract

This developmental paper examines the impact parental surveillance has on the social privacy of children in the digital age, where monitoring has shifted from overt check-ins to constant covert tracking through devices like smartwatches and apps. Through a series of semi-structured interviews with 18 parents of children aged 0-21, this developmental paper addresses the question: how does parents' use of surveillance technology affect privacy boundaries socially negotiated between parents and children? This paper presents preliminary findings on how technology intended for protection may intrude on the social privacy of a child as well as other parents' children.

Keywords: parental surveillance, digital society, social privacy, children, safety

1.0 Introduction

Digital parental surveillance

Surveillance is an everyday practice which people routinely undertake, often without thinking about it (Lyon, 2001). This includes a parent watching over their child to ensure their safety, security, or wellbeing (Ema & Fujigaki, 2011; Ghosh et al., 2018; Mols et al., 2023). While parents have watched over their children for millennia, the practice of parental surveillance has changed in recent decades, as human-orientated approaches to surveillance have been augmented with digital technology (Ema & Fujigaki, 2011; Leaver, 2017; Lyon, 2022; Mavoa et al., 2023; Mols et al., 2023). In this context, surveillance is defined as "any collection and processing of personal data, whether identifiable or not, for the purposes of influencing or managing those whose data have been garnered" (Lyon, 2001, p. 2). Crucially, this definition elucidates how surveillance technologies that allow parents to watch over their children have social implications (Lyon, 2009). While these social implications are extensive (e.g., Ema &

Fujigaki, 2011; Ghosh et al., 2018; Leaver, 2017; Mavoa et al., 2023; Mols et al., 2023), this paper focuses on the impact parental surveillance has on children's social privacy.

2.0 Conceptual Framework

As new dangers emerge with the development of digital technologies, parents find themselves confronted by new risks concerning the safety of their children who are navigating the digital society (Mols et al., 2023). Indeed, research suggests that parents perceive the physical and digital world as an unsafe place (Gür & Türel, 2022). For example, children exposed to violent and pornographic imagery online, cyberbullying, privacy invasions, and psychological vulnerability with social media to name a few. As a result of these dangers, parents face a paradox: adopt surveillance technologies to monitor their children's digital activities in the hope that it will mitigate their concerns about this unsafe society, while knowing that it could simultaneously impact their children's privacy.

Previous research has examined the reasons why parents surveil their children (e.g., Ema & Fujigaki, 2011; Mavoa et al., 2023; Sukk & Siibak, 2021), the impact parental surveillance has on the privacy of either infants, pre-teens, teens, or adolescents (e.g., Akter et al., 2022; Leaver, 2017; Sukk & Siibak, 2021), and the impact specific technologies have on children's privacy, including the internet (e.g., Akter et al., 2022; Leaver, 2017; Steeves & Regan, 2014), social media (e.g., Leaver, 2017), and location tracking (e.g., Ema & Fujigaki, 2011; Mavoa et al., 2023). Crucially, extant research tends to draw on informational conceptualisations of privacy, while failing to account for the social elements of privacy.

Accordingly, this study not only focuses on the impact parental surveillance has on the privacy of children aged 0 to adulthood but it also considers the impact different surveillance technologies have on privacy. It also moves beyond informational conceptualisations of privacy through the adoption of the social theory of privacy, to elucidate how parents and children negotiate personal privacy boundaries in intersubjective relations (Steeves, 2009). Given this background, this project's research question seeks to answer *how parents' use of surveillance technology affects privacy* *boundaries socially negotiated between parents and children?* This developmental paper begins to answer this question through the initial findings presented from a qualitative study.

Social privacy

Unlike the dominant informational conceptualisations of privacy, emphasising antisocial perspectives concerned with an individual's ability to control information about themselves (Westin, 1967), this study draws on arguments that privacy is in fact a dynamic process of social boundary control, as children change degrees of openness and degrees of disclosure depending on their privacy needs (Regan, 1995; Steeves, 2009; Steeves & Regan, 2014). Based on George Herbert Mead's work on social interactionism and Irwin Altman's work on territoriality, Steeves (2009) conceptualises privacy as a social practice involving actors (i.e., parents and children) negotiating personal boundaries through intersubjective communication. She recognises that instead of positioning privacy and social interaction as opposites, Altman juxtaposes openness and closedness to others, meaning privacy becomes the negotiated line between the two. In this sense, territories are bounded areas that children perceive as their own, and within which they may place objects and information. While a child can invite a parent into their territory, trespassing is unacceptable.

Steeves incorporates this view into her social theory of privacy while extending Mead's understanding of the self as a social construction. This frees privacy from the claim that parents and children act in isolation of others. Instead, she argues that privacy arises through a process of socialisation mediated through language that is intersubjectively constituted through communication. Thus, enabling a child "to see itself as a social object" that can "negotiate appropriate levels of openness and closedness to others" (Steeves, 2009, p. 205). It can be argued that trust between parents and children is crucial when negotiating levels of openness and closedness because it is at the heart of social relationships. It is a product of social negotiation that is important when establishing privacy boundaries; for without trust, there is no chance for reciprocity or mutuality of social negotiation (Steeves & Regan, 2014).

3.0 Methodology

This study employed a qualitative approach to explore parents' perspectives on using technologies to monitor and track their children's activities in the digital society. We conducted semi-structured interviews with 18 parents/guardians of children under the age of 21. The semi-structured interview format allowed for flexibility, enabling participants to discuss their insights and experiences while ensuring key topics were covered consistently across the interviews (Kvale, 1996). The interviews were conducted via MS Teams. Each interview lasted 60 minutes on average. During the interviews, participants were asked about their families' general use of technologies such as laptops, tablets, social media, gaming devices, home surveillance devices like Ring Doorbell, and mobile applications and technologies that are designed to monitor or track their children's activities. We also asked about how, and if, the use of these technologies was socially negotiated in family units, with discussions focusing on the perceived benefits, concerns, and privacy implications of such technologies.

The interview recordings were automatically transcribed by MS Teams and then reviewed and edited by the researchers to resolve any errors. The analysis of the interview data has started and is ongoing, following a thematic approach (Braun & Clarke, 2006). Thematic analysis is a widely recognised method for identifying, analysing, and reporting patterns (themes) within qualitative data. The transcribed interviews are being systematically coded by the research team to identify recurring themes, which are being categorised and refined through an iterative process. This method is allowing us to explore the nuances of parents' attitudes toward monitoring technologies, providing deeper insights into their children's privacy. Given the ongoing analysis and wordcount constraint, this developmental paper presents a few select findings to highlight elements of social privacy. Pseudonyms are used to protect participants' identity.

4.0 Findings

Privacy negotiation

Privacy boundaries are often negotiated between parents and children through intersubjective interactions. Indeed, parents and children discuss the extent to which technology can be used to monitor their activities:

> *When it comes to watching what my kids are doing, I've established values and expectations. So, when I tell them, 'I want to see what you're watching and what you're doing,' they open up to me. They're able to confide in me without putting any strain on our relationship. (Ella)*

Parents recognise that there are some personal matters they should not be able to monitor using technology. Consequently, privacy boundaries are often established without direct social interaction:

> *The girls have all been talking about the boys they fancy at school, and my daughter's quite private about that. I've got certain rights to know certain things, but I shouldn't know everything. She's got to have some of her own secrets and some things that she just has to bond with her friends. (Becky)*

> *You've got to trust that your kids are where they say they are. You've got to trust that they're using the tablet appropriately. Because you can't control everything. (John)*

In other instances, parents reported that there is no privacy negotiation. Parents use technology to invade their children's privacy boundaries.

> *Google have an app that you install on your phone to control your child's phone. You can track their messages and location… There wasn't much of a conversation [about us using the app]. It was more, 'We can do this. We can check what you're doing, what you're saying, and who you're messaging' as a deterrent dinner conversation. (John)*

Invading the social privacy of other children

Parents using technology to monitor their children's digital activities can not only breach the privacy of their children but also other people's children. This occurs when a child has negotiated a privacy boundary with a friend, for example, and one of their parent's peers into their negotiated boundary using technology without consent. This intrusion not only undermines their child’s negotiated privacy but also that of another child.

> *We regularly police her iPad to see who she’s messaging. We’ve actually got access to see who she’s messaging and who she’s talking to. We can see both sides of the conversation. (Janet)*

> *It does feel a bit weird when you’re looking through your daughter’s phone because they are quite personal items… I am concerned that it's her own private conversations with her friends that I wouldn’t normally be party to, but because it's on her phone, I can get insight into that... It hasn't stopped me checking her phone, but I think this will get much more complicated when she goes to high school. As she gets older, her right to privacy will increase and I've got to be respectful of that. (Becky)*

5.0 Conclusions

This developmental paper has presented some initial findings from the preliminary analysis of empirical data pertaining to how parents’ use of surveillance technology affects privacy boundaries socially negotiated between parents and children. While further analysis is required, the initial findings suggest new insight into the privacy negotiations parents and children undertake in a digital context. Specifically, how parents and children socially negotiate privacy boundaries in the context of digital parental surveillance and how acts of parental surveillance not only intrude on the privacy of their children but other people’s children. These initial insights will be later extended to drive the literature on parental surveillance.

References

Akter, M., Godfrey, A. J., Kropczynski, J., Lipford, H. R., & Wisniewski, P. J. (2022). From Parental Control to Joint Family Oversight: Can Parents and Teens Manage Mobile Online Safety and Privacy as Equals? Proceedings of the ACM on Human Computer Interaction, Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *32*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Ema, A., & Fujigaki, Y. (2011). How far can child surveillance go?: Assessing the parental perceptions of an RFID child monitoring system in Japan. *Surveillance & Society*, *9*(1/2). https://doi.org/10.24908/ss.v9i1/2.4105

Ghosh, A. K., Badillo-Urquiola, K., Guha, S., LaViola, J. J. J., & Wisniewski, P. J. (2018, 21 April 2018). Safety vs. surveillance: what children have to say about mobile apps for parental control. Conference on Human Factors in Computing Systems, Montreal.

Gür, D., & Türel, Y. K. (2022). Parenting in the digital age: attitudes, controls and limitations regarding children's use of ICT. *Computers & Education*, *183*(104504). https://doi.org/10.1016/j.compedu.2022.104504

Kvale, S. (1996). *Interviews: An Introduction to Qualitative Research Interviewing*. Sage Publications Ltd.

Leaver, T. (2017). Intimate Surveillance: Normalizing Parental Monitoring and Mediation of Infants Online. *Social Media + Society*, 1-10. https://doi.org/10.1177/2056305117707192

Lyon, D. (2001). *Surveillance society: monitoring everyday life*. Open University Press.

Lyon, D. (2009). Surveillance, power, and everyday life. In C. Avgerou, R. Mansell, D. Quah, & R. Silverstone (Eds.), *The Oxford Handbook of Information and Communication Technologies* (pp. 449-468). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199548798.001.0001

Lyon, D. (2022). Surveillance. *Internet Policy Review*, *11*(4). https://doi.org/10.14763/2022.4.1673

Mavoa, J., Coghlan, S., & Nansen, B. (2023). “It’s About Safety Not Snooping”: Parental Attitudes to Child Tracking Technologies and Geolocation Data. *Surveillance & Society*, *21*(1), 45-60. https://doi.org/10.24908/ss.v21i1.15719

Mols, A., Campos, J. P., & Pridmore, J. (2023). Family surveillance: understanding parental monitoring, reciprocal practices, and digital resilience. *Surveillance & Society*, *21*(4), 469-484. https://doi.org/10.24908/ss.v21i4.15645

Regan, P. (1995). *Legislating Privacy*. University of North Carolina Press.

Steeves, V. (2009). Reclaiming the social value of privacy. In I. Kerr, V. Steeves, & C. Lucock (Eds.), *Lessons from the identity trail: anonymity, privacy and identity in a networked society* (pp. 191-208). Oxford University Press.

Steeves, V., & Regan, P. (2014). Young people online and the social value of privacy. *Journal of Information, Communication and Ethics in Society*, *12*(4), 298-313.

Sukk, M., & Siibak, A. (2021). Caring dataveillance and the construction of "good parenting": Estonian parents' and pre-teens' reflections on the use of tracking technologies. *Communications*, *46*(3). https://doi.org/10.1515/commun-2021-0045

Westin, A. F. (1967). *Privacy and Freedom*. Atheneum.

Dancing with Synthetic Data: AI Educational Research using an AR Ballet

Genevieve Smith-Nunes[1] and Alex Shaw[2]
[1] University of Roehampton, London, UK
[2] Glastonbridge Software, Edinburgh, Scotland
ges52@cantab.ac.uk

Research In Progress

Abstract

Synthetic data (SD) is becoming an increasingly important tool in artificial intelligence (AI) research, particularly in domains where real-world data can be difficult or costly to obtain. In this research-in-progress paper, we explore the use of SD derived from brainwave and movement data to power an augmented reality (AR) episodic ballet experience. The goal of this WIP is to prompt discussions around the ethical use of body data in computing education through immersive technologies and to explore new technologies for teaching and learning within computing education. By leveraging SD rather than real user data, we aim to create an immersive AR experience that allows exploration of the relationship between physical movement, cognition, and artistic expression, while avoiding potential privacy and consent issues associated with the use of personal biometric data. Through this WIP, we investigate the technical challenges and opportunities in using SD to enable novel educational experiences, as well as the broader implications for the role of synthetic data in AI-powered educational research and applications. Our findings have the potential to inform best practices around the ethical development of data-driven educational technologies that respect individual privacy and autonomy.

Keywords: Synthetic Data, Biometrics, Data Ethics, AI, Computing Education, XR

1.0 Introduction

As artificial intelligence (AI) applications grow across diverse fields, the need for accessible, cost-effective data is more pressing than ever. Synthetic data (SD) has emerged as a valuable tool to address this need, especially in scenarios where obtaining real-world data is challenging, costly, or raises ethical concerns. In this research-in-progress paper, we investigate the use of SD generated from brainwave and movement data to create an augmented reality (AR) episodic ballet experience. The educational research from this project is in the field of creative computing through the following: programming for SD generation and AR development, ethical discourse on data ethics and AI.

This work in progress (WiP) is an exploration of synthetics for computing education and an opportunity to reflect on the ethical implications of using biometric data within immersive learning environments. By substituting synthetic for real biometric data, we aim to create a secure and engaging AR experience and learning resources that invites users to explore connections between programming, movement, cognition, and expression, while safeguarding privacy. This study seeks to spark discussion on the responsible integration of AI approaches in creative computing education, outlining both technical challenges and opportunities of using SD in AI-driven educational applications. Highlight the potential for SD to reshape data ethics in immersive learning technologies

1.1 What Is Synthetic Data

SD in simple terms refers to artificially generated data that replicates the statistical properties and patterns of real-world datasets without exposing sensitive or personally identifiable information (PII) (Rubin 1993). SD potentially offers solution to privacy, data scarcity, and financial challenges of real-world data generation. Recent research emphasises it role in domains beyond education such as Health Care (Goncalves et. Al., 2020), Governmental Census data (Abowd & Hawes, 2020), and Finance (Altman et al, 2024).

1.2. Why Use Synthetic Data?

SD can be a valuable tool in cases where using real-world data raises ethical concerns. For example, in research involving minors or other vulnerable populations, SD enables researchers to conduct analyses without directly involving these groups. Particularly relevant in educational research, where privacy concerns are paramount, and the use of real-world data involving minors requires strict ethical protocols (Adams et al., 2023). We aim to explore and develop a publicly available product without compromising participant privacy or violating ethical guidelines. SD approach(s) allow research to move forward while respecting the rights and well-being of individuals represented in the original data. It should be noted that there are no specific guidelines on the best use-case for SD (Dankar & Ibrahim, 2021), especially in education. In this WiP we use movement and brainwave datasets to generate the synthetic data. We purposefully selected python for SD generation aligning with computing education practices in England (Hadwen-Bennett & Kemp, 2024, p.10).

2. Project Overview / Background

This paper focuses on the processes, data, and creative computing techniques that will be developed into computing education resources aimed at pre-university students and non-technical artists. It forms part of a larger project, a data-driven AR ballet. The AR experience consists of five episodes, built using real and synthetic biometric datasets. The AR storyline, see table 1, follows four astronauts as they make the first ever manned journey into the vastness beyond our solar system. We follow them as they train, blast off, and explore, visualised through a web of personal journeys, biometrics, and digital simulations. It explores how our data-driven society shifts the way we perceive reality, each other, and presents the contention between enriching and dehumanising ourselves through data measurement. Only one episode will rely entirely on synthetic data.

2.1 AR Episodes

Episode	Story	Pedagogical
1 Training for Space	Focuses on the astronauts' physical and cognitive preparation using only synthetic biometric data.	Illustrate the foundational relationship between biometric analysis and performance optimisation.
2. The Launch	Simulates the physical and emotional intensity of leaving Earth, combining real and synthetic data for immersive effects.	Highlight SD's role in simulating extreme scenarios.
3. Space Station Alpha	A pause in the journey: data testing, communications, and preparation for interstellar travel	Practical computing education of data transmission, latency, and biometric data. Facilitate ethical SD discussion applications.
4. **Interstellar Travel**:	Vast distance from earth, consent monitoring.	Recursion, iteration. Data less predictable, outliers
5. **Personal Journeys:**	Astronauts feeling so far removed from 'home', detached and less able to see the importance of their data at this distance from earth. Comms glitches	Connect data to storytelling and emotional expression. Bridge real-world and SD augmentation with unknown future horizons.
6. **Signal Failure**:	Concludes with signal failure – unknown ending, loss of data, catastrophe, for astronaut agency of personal data.	Designed to be unknowable and potentially uncomfortable for discussion. Communication at this point is reduce the raw 1's and 0's

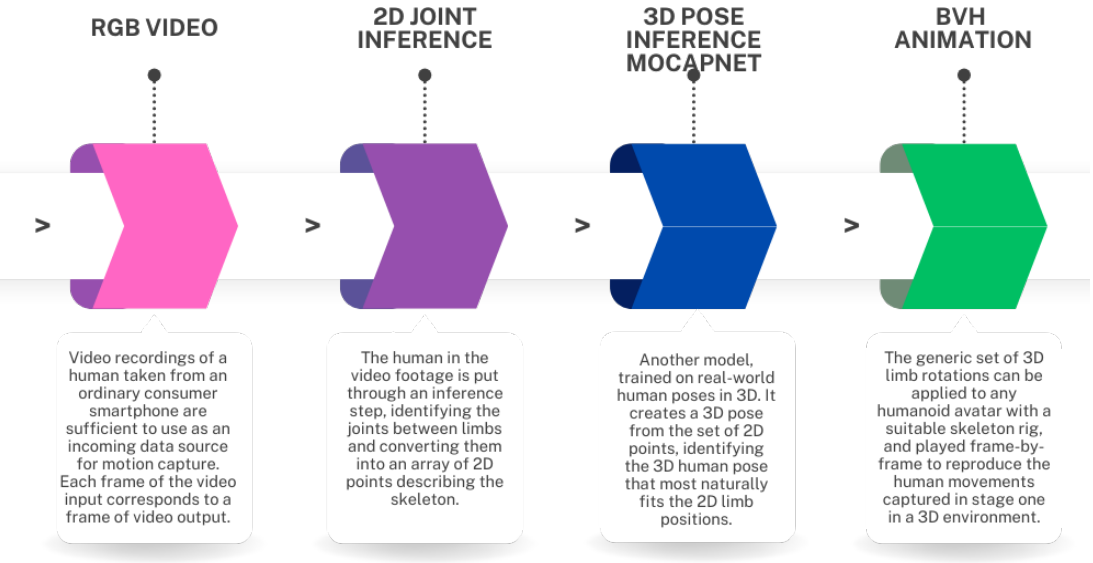
Table 1. AR Episode Overview and Pedagogical Intent.

3. Method

For the AR ballet, dancers' biometric datasets were (i) motion-captured using an extended pipeline, fig 1. Motion data, using a markerless motion capture system and MocapNET (Qammaz & Argyros, 2019, 2023) to create 3D avatars in BVH (BioVision Hierarchy) format. (ii) CSV format of Electroencephalography (EEG) data for augmenting the graphic and sound effects in the AR ballet, not for neuroscience purposes. Note that these techniques are solely to produce mathematical representations of our dancers' movement and are not related to the current trend of AI and ethics that repurposes existing creative work.

3.1 Movement Data: Real > Synth

Our process is based around MocapNET (Qammaz & Argyros, 2019, 2023) , a research motion capture project relying on a single RGB video stream to generate a 3D pose estimation of a human dancer. MocapNET uses two inference stages, the first one identifies 2D joint positions from an image, and the second estimates the 3D pose of the human from 2D joint positions. This final inference result can be exported as BVH and applied to rigged 3D models.



Video recordings of a human taken from an ordinary consumer smartphone are sufficient to use as an incoming data source for motion capture. Each frame of the video input corresponds to a frame of video output.

The human in the video footage is put through an inference step, identifying the joints between limbs and converting them into an array of 2D points describing the skeleton.

Another model, trained on real-world human poses in 3D. It creates a 3D pose from the set of 2D points, identifying the 3D human pose that most naturally fits the 2D limb positions.

The generic set of 3D limb rotations can be applied to any humanoid avatar with a suitable skeleton rig, and played frame-by-frame to reproduce the human movements captured in stage one in a 3D environment.

Figure 1. **Motion Capture Pipeline**

RBG video - 2D joint data – 3D pose – BVH 3D skeleton

3.2 EEG: Real > Synth

EEG (brainwave) data was recorded using Emotiv's Insight 5-channel headset and exported as unprocessed data to a CSV format. The data is anonymised and synthesised to create new SD sets. This process is simpler than synthesising movement data due to data structure.

3.3 Teaching Synthesising Data with Ethics

Google Colab with SynthDataVault (Patki et al., 2016) served as an interactive tool for SD generation. Pedagogical methods emphasized interdisciplinary learning by integrating Python-based programming, ethical discussions, and real-world applications in XR. For example, datasets such as dancer biometric capture stories facilitated ethical debates, while programming exercises focused on SD generation principles.

4. Pedagogical Innovation

Pedagogical innovation aims to integrate synthetic data generation, programming, and extended reality (XR) development through creative computing education. Educational objectives (i) developing students' AI competencies, (ii) generation and use of SD (iii) in XR applications, and (iv) foster critical thinking about ethical AI approaches. Building understanding of essential concepts like data privacy including differential privacy (Wood et al, 2018), algorithmic bias, and ethical considerations in synthetic data generation. Additionally, illustrates the contention between data synthesis and the intellectual rights of the person who created the training data, and how we can treat artists fairly in the AI age.

The project aligns with established computing education frameworks, which emphasise the importance of AI competencies, ethical discourse, and practical programming skills (Hadwen-Bennett & Kemp, 2024). It incorporates XR to support diverse learning outcomes and enhancing educational experiences through emerging technologies.

5. Ethics and Limitations

The three main areas of concern: bias, information loss, and ethical implications.

SD may reinforce and potential enhance existing bias form the original dataset. Loss of subtle nuances could manifest misrepresentation. The need for transparent practices for obtaining and processing biometric data are vital. Data subjects must fully understand and explicitly consent to the use of their data, even when presented in synthetic form, to uphold ethical standards and maintain trust. Ethical implications for artists, developers, and participants are significant concerns. Data ownership and intellectual property (IP) of contributions *must* be recognised and protected.

6. Next Steps

Through developing the process and learning resources for SD design for creative computing purposes we aim to test the functionality of the process of both (i) creating synthetic dataset and (ii) building XR experience with those datasets. Alongside ethical discussion of body data ethics.

Research and Design Iterations:

- Test and validate process for creating SD designed for creative computing education. (April - July with pre-service secondary computing teachers and four CS undergraduates)
- Evaluate the effectiveness of using these SD in building XR (Extended Reality) experiences. (June - Sept)
- Examine the ethical implications of collecting, generating, and using body-related data (July onwards)

We are currently generating all the synthetic biometric datasets for use in the AR ballet. It is very difficult to create synthetic ballet movement data. We believe that synthetic data will not be as effective as data generated by real dancers. However, this research will help us understand the limitations of synthetic movement data and develop teaching materials to explain the process and ethics of data synthesis.

References

Abowd, J. M., & Hawes, M. B. (2023). Confidentiality Protection in the 2020 US Census of Population and Housing. *Annual Review of Statistics and Its Application*, *10*(1), 119–144. https://doi.org/10.1146/annurev-statistics-010422-034226

Adams, C., Pente, P., Lemermeyer, G., & Rockwell, G. (2023). *Ethical principles for artificial intelligence in K-12 education.* Computers and Education: Artificial Intelligence, 4, 100131. https://doi.org/10.1016/j.caeai.2023.100131

Altman, E., Blanuša, J., Egressy, B., Anghel, A., & Atasu, K. (n.d.). *Realistic Synthetic Financial Transactions for Anti-Money Laundering Models*.

Dankar, F. K., & Ibrahim, M. (2021). Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. Applied Sciences, 11(5), 2158. https://doi.org/10.3390/app11052158

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, *20*(1), 108. https://doi.org/10.1186/s12874-020-00977-1

Hadwen-Bennett, A., & Kemp, P. (2024). Programming in Secondary Education in England: Technical Report. King's College London. https://www.kcl.ac.uk/ecs/assets/programming-in-secondary-education-in-england-full-technical-report.pdf

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 399–410. https://doi.org/10.1109/DSAA.2016.49

Qammaz, A., & Argyros, A. (2019). MocapNET: Ensemble of SNN Encoders for 3D Human Pose Estimation in RGB Images. https://users.ics.forth.gr/~argyros/mypapers/2019_09_BMVC_mocapnet.pdf

Qammaz, A., & Argyros, A. (2023). A Unified Approach for Occlusion Tolerant 3D Facial Pose Capture and Gaze Estimation using MocapNETs. In IEEE/CVF International Conference on Computer Vision Workshops (AMFG 2023 - ICCVW 2023), Paris, France, October 2023 (pp. 3178-3188). IEEE. https://users.ics.forth.gr/~argyros/mypapers/2023_10_AMFG_Qammaz.pdf

Rubin, D. (1993). Discussion Statistical disclosure limitation. Journal of Official Statistics, 9(2), 461–468. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf

Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., O'Brien, D., Steinke, T., & Vadhan, S. (2018). Differential Privacy: A Primer for a Non-Technical Audience. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3338027

Challenges Inhibiting Personalisation of the Learning Process Within an African Open and Distance Learning Institution

Godfrey M. Haonga
University of Cape Town, godfrey.m.haonga@gmail.com
Lisa F. Seymour
University of Cape Town, lisa.seymour@uct.ac.za

Completed Research

Abstract

The personalisation of the learning process has gained increasing attention in education, particularly due to new advancements in technology. Personalised learning (PL) refers to the customisation of the learning process to address the diverse needs and characteristics of individual learners. While PL has gained momentum in enhancing students' online experiences and providing personalised learning experiences, numerous challenges continue to hinder its effective implementation. Furthermore, its application in the context of Open Distance Learning (ODL) institutions within the African continent remains underexplored. In understanding its exploration within the African context, this study describes the underlying challenges that inhibit its implementation within an African ODL institution. The study involved the collection of empirical data through qualitative interviews and data analysis using an inductive approach. The study's major findings highlighted limited infrastructure, restrictive policies, the digital divide, and teachers' reluctance to adapt to new and online teaching methods.

Keywords: Open and Distance Learning, ODL, Personalised Learning, Africa.

1.0 Introduction

Personalised Learning (PL) has gained popularity in education as a means to improve student learning outcomes in distance learning environments (Zhang et al., 2020). The concept of PL has long been discussed; however, its adoption has been spurred by the development of new technologies, such as Artificial Intelligence (AI) and learning analytics, the growth of data, and the need for learner-centred approaches (Schmid et al., 2022). Technology has influenced a paradigm shift in learning from a teacher-centred approach to a learner-centred approach, allowing learners to study at their own time and in their preferred environments (Kong & Song, 2015). This shift is transforming educational pedagogies to facilitate self-regulated learning and PL based on learners' interests, preferences, and needs (Troussas et al., 2019), thereby enabling the development of 21st-century skills and placing the learner at the centre of the learning process (Kong & Song, 2015).

PL allows for the customisation of the learning process based on individual needs and characteristics (Shemshack & Spector, 2020). It actively engages students by offering them the freedom to learn in a way that aligns with their strengths, needs, interests, knowledge level, and prior learning experiences (Kamardeen & Samaratunga, 2020). This approach allows learners to engage with the material through flexible learning paths and at their own pace. With this flexibility, learners can create their own PL pathways, thereby increasing ownership of their learning (Staikopoulos et al., 2014). PL also helps mitigate challenges in online education, such as lack of motivation, poor performance, difficulty in understanding, and high dropout rates (FitzGerald et al., 2018). By individualising instruction to meet the unique needs of each learner, PL enhances the online learning experience by improving satisfaction, engagement, motivation, and performance (Shemshack & Spector, 2020), ultimately increasing student retention (Salloum et al., 2024) and improving learning experiences (Torres Kompen et al., 2019). Moreover, PL represents a significant shift in education, moving from a “traditional generic and inactive learning environment to a more personalised and active learning environment” (Kamardeen & Samaratunga, 2020, p. 106).

However, despite the potential benefits of PL, its development in educational settings remains a global challenge (Alshammari, 2020; Shemshack et al., 2021) and is particularly understudied in African Open Distance Learning (ODL) institutions. Many African ODL institutions face numerous challenges, including high student dropout rates, resource limitations, inadequate infrastructure, and outdated teaching methods (Musingafi et al., 2015). The majority of these institutions still adopt a “one-size-fits-all” approach, which fails to cater to the diverse needs of distant learners and lacks personalisation (Magomadov, 2020; Ntaba & Jantjies, 2019). This approach treats all learners equally and focuses on offering the same materials and assessments to everyone without considering their individual learning styles, preferences, abilities, and skills (Ntaba & Jantjies, 2019). The absence of PL has been linked to low learner performance and high student dropout rates

(Li & Wong, 2021). Towards addressing this problem, this study investigates the following research question: "*What are the challenges that inhibit the personalisation of the learning process within African open and distance learning institutions?".* Therefore, this paper is structured as follows: an introduction, a brief literature review, the method, key findings, and the conclusion.

2.0 Short Literature Review

This section of the literature review introduces the concept of PL, some of the PL strategies employed, and the challenges underpinning its implementation. Furthermore, it explores the context of African ODL institutions, highlighting current practices.

2.1 Personalised Learning

PL has its roots tracing back to the 20th century (Zhang et al., 2023). Its meaning varies depending on the context in which it is applied (Barrera Castro et al., 2024), largely due to its broad and often ambiguous definition (FitzGerald et al., 2018). Sabani et al. (2023) define PL as "strategies that use educators' and students' learning experiences and philosophies to create a learning environment that aligns with students' previous knowledge, learning styles, interests, cultural backgrounds and experiences" (p. 282). Similarly, Raj and Renumol (2022) define PL as "the pedagogy where the pace of learning, the instructional preferences and the learning objects are optimised as per the needs of each learner" (p. 17).

2.2 Personalised Strategies

Various strategies and methods have been employed to integrate PL into educational settings. These include proposed personalised models for customising learning content and learner preferences (Alshammari, 2020), recommendation systems (Aeiad & Meziane, 2019), and learning paths (Cavanagh et al., 2020). Additionally, adaptative e-learning systems (Alshammari, 2020) and data-driven approaches, such as learning analytics and machine learning techniques, have been used to optimise personalisation (Meacham et al., 2020; Qi, 2018). The integration of these approaches has been crucial for enhancing online learning.

2.3 Personalised challenges

The implementation of PL in educational settings still faces numerous challenges. Some barriers stem from the complexity of designing PL itself (Imran et al., 2024; Shemshack & Spector, 2020). Additionally, integrating and adapting various methods into PL requires significant time and resources (Imamah et al., 2024). Key resource constraints include a lack of digital tools and technologies (Xu et al., 2024), inadequate infrastructure such as limited internet connectivity (Dhananjaya et al., 2024), and technological challenges like outdated hardware and software (Bingham et al., 2018), all of which hinder the successful implementation of PL. Moreover, the lack of a universally accepted definition of PL complicates implementation efforts (Xu et al., 2024). Similarly, the absence of a standard framework that can be utilised across various contexts and platforms presents another challenge (Apoki & Crisan, 2022; Trushin & Ermakova, 2024). Additional barriers include educators' and instructional designers' limited skills and knowledge in utilising PL digital tools for PL implementation (Imran et al., 2024), financial constraints (Kucirkova et al., 2021), data issues (Bin et al., 2024), and the need for ongoing staff training and support (Zhang et al., 2023).

While numerous studies have explored the challenges of integrating PL strategies and methods in educational settings, research on PL challenges in ODL contexts, particularly African ODL environments, remains limited. Studies such as Xu et al. (2024) and Zhang et al. (2023), explore PL challenges in traditional classroom settings. Existing research has predominantly focused on elementary, middle and high school environments (Bingham et al., 2018; Kucirkova et al., 2021; Mukhamadiyeva & Hernández-Torrano, 2024; Robinson & Sebba, 2010), leaving a gap in understanding PL challenges in ODL institutions. Moreover, the African ODL context presents distinct characteristics and challenges that warrant further study, underscoring the need for research to identify barriers to PL implementation in this context. Therefore, this study explores the challenges of implementing PL in an African ODL institution.

2.4 ODL and African Context

ODL institutions provide a distinct approach to teaching and learning using technology. Unlike conventional universities, which primarily deliver education in physical classrooms, ODL institutions use technology to facilitate online learning. The Commonwealth of Learning (2023) defines ODL as "the provision of distance education opportunities in ways that seek to mitigate or remove barriers to access, such as finances, prior learning, age, social, work or family commitments, disability, incarceration or other such barriers" (p. 5). The key benefits of ODL include the ability to access education anytime, anywhere, the flexibility of learners to study at their own pace, and the provision of education to a broader and more diverse population through overcoming barriers of distance and time (Isaacs & Mohee, 2020). Many African ODL institutions have adopted technologies such as e-learning platforms. Some African ODL institutions are the University of South Africa, the Open University of Tanzania, and the National Open University of Nigeria. The majority of ODL learners in these institutions are geographically dispersed, come from diverse cultural backgrounds and have varying learning needs.

3.0 Method

This study is part of a Design Science Research (DSR) project, employing DSR methodology to assess the problem identification phase. This phase examined the challenges hindering the personalisation of learning process within an ODL institution, which is essential for developing a framework as a DSR artefact to assist educators in personalising learning within an African context. To understand the problem domain, which forms the first phase of the DSR methodology, data were collected using an interpretive approach and analysed qualitatively. Interviews were conducted, and the data were analysed inductively, allowing themes to emerge directly from the data. A purposive sampling technique was used to recruit 18 participants, including 10 academic staff, 3 instructional designers, and 5 students comprising both postgraduate (2) and undergraduate (3). Participants were selected based on their relevant expertise in distance learning and their knowledge relevant to the study's focus. The duration of the interviews ranged from 35 minutes to 1 hour. Table 1 presents the participants with the coded representation to protect their anonymity. The study received ethical approval from both the ODL institution and the researcher's university. The data were thematically analysed following the process outlined by Braun and Clarke (2006). This process began with familiarisation with the data, followed by coding, categorization, and the generation of initial themes. The next step involved a peer review by the supervisor, followed by refinement of themes to establish the final themes, which formed the basis of the main findings.

To explore the challenges within the African ODL institutions, this study identifies the Open University of Tanzania (OUT) as the research context and a key source of requirements during the development and evaluation of the DSR artefact. Established in 1992, OUT is an accredited ODL public university that adopts the Moodle e-learning platform for teaching and learning, with most of its programs offered online. The university provides a range of academic programs, including Doctor of Philosophy, Masters, Undergraduate, Diploma, and Certificate programs. Its headquarters are in Dar es Salaam, with centres in 30 regions in Tanzania and coordination centres in countries such as Kenya, Uganda, Rwanda, and Namibia.

Participant	Code	Number
Teacher (Academic Staff)	**TU**	10
Student	**STU**	5
Instructional designer	**ID**	3
Total		18

Table 1. Research Participants

4.0 Findings and Discussion

In answering the research question, the following themes emerged as challenges that hinder the personalisation of the learning process within the African ODL institution.

4.1 Technological and Infrastructure Challenges

Technological and infrastructure limitations emerged as significant barriers to implementing PL in the ODL institution, with six subthemes. One major challenge is the lack of **adequate infrastructure and technology** to support PL. The institution lacks reliable internet connections and power supply, particularly in rural areas. This unequal distribution of infrastructure and technological development among areas hinders PL implementation, as learners cannot fully engage in the learning process. The findings align with the literature, which emphasizes that establishing effective PL environments requires substantial investment in infrastructure (Dhananjaya et al., 2024).

> *"There are some students who come from the interior…where there is probably no access to the internet, but also, on the side of power too, sometimes it happens in the interior, they lack electricity for some students ~ (**ID01**). There is a breakdown of the internet. There may be some problems with the internet; sometimes it is not available. It is a very big challenge ~ (**ST04**). ... the challenges that we are facing are in terms of developing and maintaining the infrastructures for ODL ... implementing personalised or personalisation ... requires very sophisticated technology ~ (**TU07**)."*

Another challenge pertains to the inability of the existing **e-learning platform** to adapt to and support PL. This finding aligns with the literature, which confirms that many e-learning platforms lack the personalisation and adaptability capabilities needed to support PL (Abhirami & Devi, 2022). Alshammari (2020) further supports this, noting that these platforms often overlook the learners' requirements and lack functionalities to facilitate the design and integration of PL. As a result, they primarily support the delivery of uniform resources and assessment methods (Arsovic & Stefanovic, 2020). This study reveals that the current e-learning platform lacks the necessary features for PL integration, as it predominantly delivers the same material to all learners.

> "*... it is the system that provides an unequal service to all and similar service to all students ~ (**TU04**)."*

In addition, there is a limited **availability of data** for addressing the needs of learners. The analysis reveals a lack of sufficient data to inform the creation of PL experiences within the institution. Data plays a crucial role in developing PL, as it helps identify learners' characteristics and needs (Boelens et al., 2017). According to Barrera Castro et al. (2024), data is essential for enabling automated PL solutions through technologies like AI and learning analytics. However, data also presents multifaceted concerns, particularly regarding **privacy and security**, which pose additional challenges in implementing PL (Imran et al., 2024; Xu et al., 2024). This study highlights concerns about protecting sensitive information, further hindering the implementation and integration of PL.

> *"We don't have enough data that could inform us about our students and like how they engage online, how they behave ... and how that links to their past experiences and their interests ~ (**TU10**). ... we have issues of data privacy concerns.*

> *Personalisation in most cases, it relies on collecting and analysing a significant amount of student data...this can also raise issues of data privacy and security and on how this person's information is protected and used ~ (**TU07**).*"

Another challenge is the **adaptation to new technologies**, which makes it difficult for both students and teachers to keep up with constant changes. The rapid emergence of new technologies requires both teachers and learners to be proficient in using new tools and updates. As noted by Shemshack et al. (2021), the presence of numerous personalised technologies creates challenges in adaptation due to the lack of a unified implementation approach, making their effective use in creating PL experiences more complex. Interview responses indicate that these technologies can be overwhelming for teachers, posing a significant barrier to the successful implementation of PL.

> *"Sometimes also adaption of new technology; change of technologies, as we understand that the technology changes in every time ~ (**TU01**). ... being conversant with such technologies again is another obstacle ~ (**TU03**)."*

A further challenge is the **digital divide** among teachers and students. The findings indicate that some learners and teachers within the ODL lack access to digital devices, which limits their ability to participate and fully engage in PL environments. As noted by Chitanana (2024), the inequality in access to essential devices results in unequal PL benefits, with disadvantaged students being particularly affected. The responses attributed this challenge to the high cost of acquiring digital devices and related expenses, such as internet data bundles.

> *"Not all have smartphones, and if they have a smartphone, not all that they have a bundle for them to be able to use ~ (**ID02**). ... most of them cannot afford to purchase their own digital devices like mobile phones and laptops and so on ~ (**ID03**). ... the absence of these devices can hinder their participation and engagement in an online learning ~ (**TU07**)."*

4.2 Teacher Challenges in Adapting to New and Online Teaching Methods

One of the key findings pertains to the reluctance of teachers to **adapt to new teaching methods** that are student-centred, hence hinders their engagement in the PL process. This reluctance is attributed to the struggle teachers face in transitioning from a traditional, teacher-centred mode to more personalised, student-centred approaches. As noted in the literature, the traditional mode of teaching emphasizes 'one-size-fits-all' instruction, which does not consider the needs and abilities of individual learners (Magomadov, 2020). Our findings indicate that teachers' attachment to conventional practices and resistance to embracing new strategies impede the effective implementation of PL within the ODL institution.

> *"Most of our academic staff are from conventional mode of delivery. So, it is very hard for them to change it and adapt to the new system ~ (**ID02**). ... they do not want to use this student-centred mode ~ (**TU06**). ... most of the lecturers do not take that into consideration [needs of learners]. They just put the content so as to make sure that the content is there, but they do not think of who's trying to use the content to the essence that if somebody comes with different requirement from what the lecture has thought about, then it is going to be not easier for them ~ (**TU06**)."*

Another challenge identified is the reluctance of teachers to **engage in the online environment**, which poses a significant obstacle to PL implementation. The findings reveal that some teachers are accustomed to traditional face-to-face teaching methods; therefore, adapting to online teaching presents challenges for them. This aligns with Alserhan and Yahaya (2021), who emphasise the importance of teachers being able to effectively use technology like the internet and possess competencies in e-learning approaches. Some teachers do not fully adopt online environments, which becomes an obstacle towards the efforts to implement PL effectively.

> *"Reluctance of some of the academic staff to engage in the online environment. Some of them came from the universities where they were being taught in the conventional mode ~ (**ID03**)."*

4.3 Restrictive Curriculum and Policies

The findings of this study indicate that **restrictive policies** hinder the personalisation of the learning process within the ODL environment. The analysis reveals that restrictive institutional policies and regulations limit teachers' autonomy in adapting their teaching methods and personalising the learning content to meet the needs of diverse learners. As a result, teachers are bound by the established regulatory bodies and institutional policies which prioritize standardised teaching and lack flexibility. This inflexibility undermines their ability to implement adaptable PL experiences in their teaching, making it difficult to create PL that fosters individualisation. The findings align with the literature, which emphasizes the importance of establishing flexible educational policies that foster learning environments designed to promote PL (DeMink-Carthew et al., 2020).

> *"I believe it starts with the institution itself because I believe the barrier is that the instructors, the lecturers are not given much choice on how to deliver, how to personalise the learning ... so there's no flexibility within an institution ~ (**TU06**). We have institutional policies and regulations that may not be flexible ~ (**TU07**). ... we don't have the flexibility of accommodating these changes or to implement ...personalisation ~ (**TU04**)."*

Another aspect of the challenge reported is the **rigidity of the curriculum**. The findings reveal that the curriculum within the ODL institution lacks flexibility and does not accommodate the diverse needs of the learners. As supported by literature, most curricula focus on static content delivery and support standardised teaching methods based on 'one-size-fits-all', which overlooks the unique needs of individual learners (Alawneh et al., 2024). Thus, teachers within the ODL are bound by the existing curriculum, which limits their ability to implement PL and address the varying needs of their learners. This leads to their reliance on uniform methods, such as providing the same learning outcomes for all learners, with little flexibility in the learning process.

> *"... we usually provide content based on the curriculum or the course description. The course description is what guides that every student will study the course the same way ... ~ (**TU02**). The system that is being prepared is the same, which will follow the same ~ (**STU05**). ... most of our curriculums, they are not updated to meet this kind of requirements ~ (**TU07**)."*

4.4 Lack of Institutional Guidelines and Support

The **lack of institutional guidelines** is another challenge noted to affect PL efforts within the ODL institution. Our findings align with Apoki and Crisan (2022), who note that the absence of a standard framework or design guidelines poses a significant challenge for implementing PL. Most respondents assert that the absence of guidelines hampers PL efforts, as PL creation has instead been an individual effort due to the lack of guidelines. According to Trushin and Ermakova (2024), the availability of guidelines is essential for providing methodological support for PL design and integration.

> "*There's no guideline which we are using as the institute...It is very important because we need it when it comes to implementation. It is difficult because we have never seen any guideline which we have come across to help us or instruct us, so we do as ourselves ~ (**TU01**). ... normally, we do it individually. It's just an*

> *individual effort as an instructor, but there's no guideline or method or a framework or any strategy within an institution now ~ **(TU05)**."*

Another barrier to PL implementation is the **lack of management and institutional support**. This study's findings align with the literature, which suggests that transforming the current learning process toward PL requires both institutional and management backing to allocate the necessary funds (Zhang et al., 2023). Our analysis reveals that the absence of management support for PL initiatives hinders teachers' efforts in creating PL experiences within the ODL environment. Thus, implementing PL requires sufficient resources and funding. The respondents emphasize the need for management support, adequate funding, and incentives for teachers as crucial to successfully implementing PL in the ODL institution.

> *"... support from the management is very low ~ **(ID02)**. ... To cater to individual needs of students. It means you need resources. The facilitators [and] lecturers have to sit down, maybe in workshops and design their lecture materials, and then write them. All this needs funding. But, if there are no funds, I think it is a difficult task ... we need to have incentives to motivate these facilitators to sit down, design and write study materials ~ **(TU09)**."*

4.5 Lack of Awareness, Understanding and Training on Personalisation Strategies

The **lack of awareness of PL and its requirements** among teachers is another challenge identified. Similar findings were reported in the study by Bingham et al. (2018). As Barrera Castro et al. (2024) noted, a lack of clarity on how to implement and achieve PL effectiveness is a significant challenge. This study found that some teachers are unaware of PL due to their unfamiliarity with the concept itself, and they lack the knowledge of its design and implementation within the ODL environment.

> "*… the awareness of it, I think to get that education, how to use it, to prepare those study materials, to be able to meet that personalised learning ~ **(ID01)**. I think most of us, perhaps, we don't understand that concept clearly. Personalised learning: what it means and what it entails, what it requires us to do. I think most of us, we don't understand ~ **(TU03)**."*

A **Lack of training** necessary for the design and implementation of PL within the ODL is another barrier. This study confirms with the literature, which emphasizes the importance of training to help teachers master the new technologies and tools essential for PL and adapt the learning process to meet the unique characteristics and needs of each learner (Trushin & Ermakova, 2024). Our analysis reveals that the lack of training within the ODL institution limits teachers' ability to create PL experiences for their students.

> *"... institution does not provide a scheduled workshop or scheduled training for the academic staff, to make sure that, all staff are being trained in that manner ~ **(ID02)**. ... the lecturers are not well equipped ~ **(TU06)**."*

Another challenge reported is the **insufficient pedagogical knowledge and design skills** required for the implementation of PL in both teaching and learning. According to Xu et al. (2024), the design and implementation of PL demand a substantial understanding of instructional design principles to effectively structure PL experiences. Our study reveals that teachers lack the competencies required to facilitate the creation of PL within the ODL institution. The absence of these foundational skills and knowledge hinders PL implementation, as targeted training is essential to equip them with the competencies needed for successful PL design and integration.

> *"We're just teachers, instructors, but we don't have actual instructional design [skills]. It's a course, it's a skill, it's a knowledge that you need to have ~ **(TU05)**. ... they do not have the skills to design these materials to make them personalised ~ **(TU09)**."*

4.6 Limited Digital Literacy

The lack of **digital literacy among learners** presents a significant obstacle to their ability to engage in PL environments, as highlighted in the literature (Xu et al., 2024). Our findings indicate that there are variations in digital skills among learners within the ODL, which poses a challenge to their participation in the online environment and involvement in PL. Some learners lack the necessary digital skills, which restricts their engagement in the learning process and limits their ability to personalise their learning to reflect their needs. Additionally, students come from diverse educational backgrounds with varying levels of familiarity with IT, particularly in the use of electronic devices. As a result, the lack of digital literacy affects the ability of some students to fully engage in PL.

> *"The issue of variations in digital literacy level among learners ~ **(TU07)**. They normally have inadequate digital skills. That could be one of the challenges that, sometimes, they don't want to participate in the online environment because to be able to do that, then they need to have literacy in digital skills ~ **(ID03)**. ... some students come from diplomas and some students come direct from Form 6, so their*

> *knowledge level on electronic devices; computers will always be different ~ **(ST01)**."*

Another challenge pertains to **inadequate digital skills among teachers**. In alignment, Alserhan et al. (2023) note that the creation of PL requires teachers' ability to engage with and adapt various digital tools that support personalisation. This study found that a lack of digital skills among teachers as a barrier, as some lack the necessary digital abilities to design materials that inform the PL experiences of their students. For teachers to fully participate in PL, they need to have skills in integrating the required tools and features essential for effective PL implementation (Xu et al., 2024).

> *"Inadequate knowledge of digital skills among academic staff has been one of the challenges hindering personalisation ~ **(ID03)**."*

4.7 Time and Complexity Challenges

Time constraint is also identified as a challenge, as teachers often have significant workloads, limiting the amount of time they can dedicate to PL. Most teachers indicate that developing tailored materials to meet individual learning needs is time-consuming and adds to their already busy schedules. As reinforced by the literature, creating PL environments requires a significant amount of time and intensive teacher involvement (Barrera Castro et al., 2024). Torres Kompen et al. (2019) noted that constructing PL is a complicated task that consumes a significant amount of time. In this study, teachers reported having limited time to engage in the PL process while balancing other demanding tasks assigned to them, hence, the barrier to the PL process.

> *"... preparing study materials that reflect that personalised learning, means it takes time. A teacher has a lot of lessons, and the way to do this, I think, it needs a lot of time to be able to prepare these things and personalise ~ **(ID01)**. Personalising the same content to different groups of students. It can be quite hard work for a lecturer ~ **(TU03)**. ... there is a busy schedule of academic [activities] that it is difficult for them ~ **(ID02)**."*

Another challenge reported is the **perceived difficulty and tediousness of the PL process**. The implementation of PL within the ODL environment is described as a challenging task that requires a substantial amount of work and effort, especially when managing many students. This aligns with findings from Bayounes et al. (2022), that personalising the learning process to cater to the needs of each learner is challenging due to the complexity of the task. This complexity hinders teachers' ability to effectively engage with PL process in their teaching.

> *"Personalised learning; it can also be quite tedious, especially with the large number of students, personalised learning can be a bit tedious ~ (**TU03**). ... the process of knowing your students, their backgrounds, their interests; It can be quite challenging ~ (**TU10**)."*

5.0 Recommendation and Conclusion

The establishment of effective and efficient PL environments requires substantial investment in ICT infrastructure. Thus, the current technological advancements necessitate the continuous upgrading of essential infrastructure to support these environments. Therefore, there is a need to establish basic infrastructure, such as reliable internet connection and power supply, particularly in rural areas, to ensure equal access and reduce the digital divide among learners. This will be essential for bringing PL benefits to all distance learners, regardless of their location. Furthermore, there is a need to establish supportive environments spanning from management to technical levels. This involves fostering motivation, securing necessary resources, and implementing capacity-building programs. Training plays a key role in PL by equipping educators and students with the skills and knowledge needed to engage effectively in PL environments. It also facilitates the development of digital competencies and enhances digital literacy, which are crucial for the success of PL initiatives. Additionally, the development of policies that foster personalisation is crucial for the successful implementation of PL initiatives. Adaptive policies are essential in driving transformational changes in teaching and learning practices, aligning them with 21st-century skills that emphasize student-centred approaches. Therefore, educational institutions should adopt flexible policies that empower educators to adapt curricula and teaching methods to support PL. This includes curriculum revision, and the integration of personalised teaching strategies tailored to the unique characteristics of distance learners, thereby enhancing their PL experiences.

This study highlights the challenges that inhibit the personalisation of the learning process in the context of an African ODL institution. Many African ODL institutions continue to adopt a “one-size-fits-all” approach that fails to accommodate the diverse needs of learners and lacks personalisation. The continued use of this method has been associated with low learner performance and high dropout rates. Despite the potential benefits offered by PL in addressing these challenges, its implementation has remained slow. Therefore, this study addressed the following question: *What are the challenges that inhibit the personalisation of the learning process within African ODL institutions?* The key challenges identified include infrastructure and technological disparities due to uneven rural-urban development, digital illiteracy among teachers and learners, the digital divide among learners, teachers’ difficulties in adapting to new and online teaching methods, and restrictive institutional curricula and policies.

One limitation of this study is that it was conducted within a single ODL institution which limits its generalisability. To gain more comprehensive insights, future research could expand the research to include multiple ODL institutions. Additionally, further research could assess the impact of PL within African ODL institutions, particularly with respect to improvements in learner performance and the enhancement of PL experiences.

References

Abhirami, K., & Devi, M. (2022). Student Behavior Modeling for an E-Learning System Offering Personalized Learning Experiences. *Computer Systems Science and Engineering*, *40*(3), 1127–1144. https://doi.org/10.32604/csse.2022.020013

Aeiad, E., & Meziane, F. (2019). An adaptable and personalised E-learning system applied to computer science Programmes design. *Education and Information Technologies*, *24*(2), 1485–1509. https://doi.org/10.1007/s10639-018-9836-x

Alawneh, Y. J. J., Sleema, H., Salman, F. N., Alshammat, M. F., Oteer, R. S., & ALrashidi, N. K. N. (2024). Adaptive Learning Systems: Revolutionizing Higher Education through AI-Driven Curricula. *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, 1–5. https://doi.org/10.1109/ICKECS61492.2024.10616675

Alserhan, S., Alqahtani, T. M., Yahaya, N., Al-Rahmi, W. M., & Abuhassna, H. (2023). Personal Learning Environments: Modeling Students’ Self-Regulation

Enhancement Through a Learning Management System Platform. *IEEE Access*, *11*, 5464–5482. https://doi.org/10.1109/ACCESS.2023.3236504

Alserhan, S., & Yahaya, N. (2021). Teachers' Perspective on Personal Learning Environments via Learning Management Systems Platform. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(24), 57–73. https://doi.org/10.3991/ijet.v16i24.27433

Alshammari, M. (2020). Design and evaluation of an adaptive framework for virtual learning environments. *International Journal of Advanced and Applied Sciences*, *7*(5), 39–51. https://doi.org/10.21833/ijaas.2020.05.006

Apoki, U. C., & Crisan, G. C. (2022). A Modular and Semantic Approach to Personalised Adaptive Learning: WASPEC 2.0. *Applied Sciences*, *12*(15), 7690. https://doi.org/10.3390/app12157690

Arsovic, B., & Stefanovic, N. (2020). E-learning based on the adaptive learning model: Case study in Serbia. *Sādhanā*, *45*(1), 266. https://doi.org/10.1007/s12046-020-01499-8

Barrera Castro, G. P., Chiappe, A., Becerra Rodriguez, D. F., & Sepulveda, F. G. (2024). Harnessing AI for Education 4.0: Drivers of Personalized Learning. *Electronic Journal of E-Learning*, *22*(5), 01–14. https://doi.org/10.34190/ejel.22.5.3467

Bayounes, W., Saadi, I., & Kinsuk. (2022). Adaptive learning: Toward an intentional model for learning process guidance based on learner's motivation. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00215-9

Bin, Q., Zuhairi, M. F., & Morcos, J. (2024). A Comprehensive Study On Personalized Learning Recommendation In E-Learning System. *IEEE Access*, *12*, 100446–100482. https://doi.org/10.1109/ACCESS.2024.3428419

Bingham, A. J., Pane, J. F., Steiner, E. D., & Hamilton, L. S. (2018). Ahead of the Curve: Implementation Challenges in Personalized Learning School Models. *Educational Policy*, *32*(3), 454–489. https://doi.org/10.1177/0895904816637688

Boelens, R., De Wever, B., & Voet, M. (2017). Four key challenges to the design of blended learning: A systematic literature review. *Educational Research Review*, *22*, 1–18. https://doi.org/10.1016/j.edurev.2017.06.001

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cavanagh, T., Chen, B., Lahcen, R. A. M., & Paradiso, J. R. (2020). Constructing a design framework and pedagogical approach for adaptive learning in higher education: A practitioner's perspective. *International Review of Research in Open and Distance Learning*, *21*(1), 153–171. Scopus. https://doi.org/10.19173/irrodl.v21i1.4529

Chitanana, L. (2024). The sustainability of blended learning in Zimbabwean state universities in the post-COVID-19 era. *E-Learning and Digital Media*, 20427530241277913. https://doi.org/10.1177/20427530241277913

Commonwealth of Learning. (2023, May). *Open and Distance Learning: Key Terms and Definitions*. https://oasis.col.org/items/7dc20f7c-4901-433a-90f1-6274f5ce53dd

DeMink-Carthew, J., Netcoh, S., & Farber, K. (2020). Exploring the potential for students to develop self-awareness through personalized learning. *The Journal of Educational Research*, *113*(3), 165–176. psyh. https://doi.org/10.1080/00220671.2020.1764467

Dhananjaya, G. M., Goudar, R. H., Kulkarni, A. A., Rathod, V. N., & Hukkeri, G. S. (2024). A Digital Recommendation System for Personalized Learning to Enhance Online Education: A Review. *IEEE Access*, *12*, 34019–34041. https://doi.org/10.1109/ACCESS.2024.3369901

FitzGerald, E., Jones, A., Kucirkova, N., & Scanlon, E. (2018). A literature synthesis of personalised technology-enhanced learning: What works and why. *Research in Learning Technology*, *26*(0). https://doi.org/10.25304/rlt.v26.2095

Gunathilaka, T., Fernando, M., & Pasqual, H. (2018). Individual learning path personalization approach in a virtual learning environment according to the dynamically changing learning styles and knowledge levels of the learner. *International Journal of Advanced and Applied Sciences*, *5*(5), 10–19. https://doi.org/10.21833/ijaas.2018.05.002

Iatrellis, O., Kameas, A., & Fitsilis, P. (2019). EDUC8 ontology: Semantic modeling of multi-facet learning pathways. *Education and Information Technologies*, *24*(4), 2371–2390. https://doi.org/10.1007/s10639-019-09877-4

Imamah, Yuhana, U. L., Djunaidy, A., & Purnomo, M. H. (2024). Development of Dynamic Personalized Learning Paths Based on Knowledge Preferences and the Ant Colony Algorithm. *IEEE Access*, 1–1. https://doi.org/10.1109/ACCESS.2024.3442312

Imran, M., Almusharraf, N., Ahmed, S., & Mansoor, M. I. (2024). Personalization of E-Learning: Future Trends, Opportunities, and Challenges. *International Journal of Interactive Mobile Technologies (iJIM)*, *18*(10), 4–18. https://doi.org/10.3991/ijim.v18i10.47053

Isaacs, S., & Mohee, R. (2020). *Baseline Situational Analysis on Open Distance Learning (ODL) in Southern African Development Community (SADC) Member States*.

Kamardeen, I., & Samaratunga, M. (2020). DigiExplanation driven assignments for personalising learning in construction education. *Construction Economics and Building*, *20*(3). https://doi.org/10.5130/AJCEB.v20i3.7000

Kong, S. C., & Song, Y. (2015). An experience of personalized learning hub initiative embedding BYOD for reflective engagement in higher education. *Computers & Education*, *88*, 227–240. https://doi.org/10.1016/j.compedu.2015.06.003

Kucirkova, N., Gerard, L., & Linn, M. C. (2021). Designing personalised instruction: A research and design framework. *British Journal of Educational Technology*, *52*(5), 1839–1861. psyh. https://doi.org/10.1111/bjet.13119

Kurilovas, E. (2016). Evaluation of quality and personalisation of VR/AR/MR learning systems. *Behaviour & Information Technology*, *35*(11), 998–1007. psyh. https://doi.org/10.1080/0144929X.2016.1212929

Li, K. C., & Wong, B. T.-M. (2021). Features and trends of personalised learning: A review of journal publications from 2001 to 2018. *Interactive Learning Environments*, *29*(2), 182–195. https://doi.org/10.1080/10494820.2020.1811735

Li, K., & Wong, B. (2023). Personalisation in STE(A)M education: A review of literature from 2011 to 2020. *Journal of Computing in Higher Education*, *35*(1), 186–201. https://doi.org/10.1007/s12528-022-09341-2

Magomadov, V. S. (2020). The application of artificial intelligence and Big Data analytics in personalized learning. *Journal of Physics: Conference Series*, *1691*(1). Publicly Available Content Database; Technology Collection. https://doi.org/10.1088/1742-6596/1691/1/012169

Meacham, S., Pech, V., & Nauck, D. (2020). AdaptiveVLE: An Integrated Framework for Personalized Online Education Using MPS JetBrains Domain-Specific Modeling Environment. *IEEE ACCESS*, *8*, 184621–184632. https://doi.org/10.1109/ACCESS.2020.3029888

Mukhamadiyeva, S., & Hernández-Torrano, D. (2024). Adaptive Learning to Maximize Gifted Education: Teacher Perceptions, Practices, and Experiences. *Journal of Advanced Academics*, *35*(4), 652–670. https://doi.org/10.1177/1932202X241253166

Musingafi, M. C. C., Mapuranga, B., Chiwanza, K., & Zebron, S. (2015). Challenges for Open and Distance learning (ODL) Students: Experiences from Students of the Zimbabwe Open University. *Journal of Education and Practice*.

Ntaba, A., & Jantjies, M. (2019). Open Distance Learning and Immersive Technologies: A Literature Analysis. *Proceedings of the 16th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2019)*, 51–60. https://doi.org/10.33965/celda2019_201911L007

Perisic, J., Milovanovic, M., & Kazi, Z. (2018). A semantic approach to enhance moodle with personalization. *Computer Applications in Engineering Education*, *26*(4), 884–901. https://doi.org/10.1002/cae.21929

Premlatha, K., & Geetha, T. (2015). Learning content design and learner adaptation for adaptive e-learning environment: A survey. *Artificial Intelligence Review*, *44*(4), 443–465. https://doi.org/10.1007/s10462-015-9432-z

Qi, Z. (2018). Personalized Distance Education System Based on Data Mining. *International Journal of Emerging Technologies in Learning (iJET)*, *13*(07), 4. https://doi.org/10.3991/ijet.v13i07.8810

Raj, N., & Renumol, V. (2022). An improved adaptive learning path recommendation model driven by real-time learning analytics. *Journal of Computers in Education*. https://doi.org/10.1007/s40692-022-00250-y

Razali, F., Sulaiman, T., Ayub, A. F. M., & Majid, N. A. (2022). Effects of Learning Accessibility as a Mediator between Learning Styles and Blended Learning in Higher Education Institutions during the Covid-19 Pandemic. *Asian Journal of University Education*, *18*(2), 569–584. https://doi.org/10.24191/ajue.v18i2.18189

Robinson, C., & Sebba, J. (2010). Personalising learning through the use of technology. *Computers & Education*, *54*(3), 767–775. https://doi.org/10.1016/j.compedu.2009.09.021

Sabani, N., Jimmie, A., & Salleh, S. Mohd. (2023). Factors Influencing the Use of Digital Learning Personalisation. *2023 11th International Conference on Information and Education Technology (ICIET)*, 282–287. https://doi.org/10.1109/ICIET56899.2023.10111345

Salloum, S. A., Salloum, A., Alfaisal, R., Basiouni, A., & Shaalan, K. (2024). Predicting Student Adaptability to Online Education Using Machine Learning. In A. Basiouni & C. Frasson (Eds.), *Breaking Barriers with Generative Intelligence. Using GI to Improve Human Education and Well-Being* (Vol. 2162, pp. 187–196). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-65996-6_16

Sarwar, S., Ul Qayyum, Z., Garcia-Castro, R., Safyan, M., & Munir, R. (2019). Ontology based E-learning framework: A personalized, adaptive and context aware model. *Multimedia Tools and Applications*, *78*(24), 34745–34771. https://doi.org/10.1007/s11042-019-08125-8

Schmid, R., Pauli, C., Stebler, R., Reusser, K., & Petko, D. (2022). Implementation of technology-supported personalized learning—Its impact on instructional quality. *The Journal of Educational Research*, *115*(3), 187–198. https://doi.org/10.1080/00220671.2022.2089086

Shemshack, A., Kinshuk, & Spector, J. M. (2021). A comprehensive analysis of personalized learning components. *Journal of Computers in Education*, *8*(4), 485–503. https://doi.org/10.1007/s40692-021-00188-7

Shemshack, A., & Spector, J. (2020). A systematic literature review of personalized learning terms. *Smart Learning Environments*, *7*(1). https://doi.org/10.1186/s40561-020-00140-9

Solari, M., Vizquerra, M. I., & Engel, A. (2022). Students' interests for personalized learning: An analysis guide. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-022-00656-3

Sridharan, S., Saravanan, D., Srinivasan, A. K., & Murugan, B. (2021). Adaptive Learning Management Expert System with Evolving Knowledge Base and Enhanced Learnability. *Education and Information Technologies*, *26*(5), 5895–5916. https://doi.org/10.1007/s10639-021-10560-w

Staikopoulos, A., O'Keeffe, I., Rafter, R., Walsh, E., Yousuf, B., Conlan, O., & Wade, V. (2014). AMASE: A framework for supporting personalised activity-based learning on the web. *Computer Science and Information Systems*, *11*(1), 343–367. https://doi.org/10.2298/CSIS121227012S

Torres Kompen, R., Edirisingha, P., Canaleta, X., Alsina, M., & Monguet, J. M. (2019). Personal learning Environments based on Web 2.0 services in higher education. *Telematics and Informatics*, *38*, 194–206. https://doi.org/10.1016/j.tele.2018.10.003

Troussas, C., Chrysafiadi, K., & Virvou, M. (2019). An intelligent adaptive fuzzy-based inference system for computer-assisted language learning. *Expert Systems With Applications*, *127*, 85–96. https://doi.org/10.1016/j.eswa.2019.03.003

Trushin, S. M., & Ermakova, V. I. (2024). Application of Adaptive Assessment Methods in Digital Educational Technologies: Modeling and Statistical Analysis of Student Work Outcomes. *2024 4th International Conference on Technology Enhanced Learning in Higher Education (TELE)*, 214–218. https://doi.org/10.1109/TELE62556.2024.10605650

Vagale, V., Niedrite, L., & Ignatjeva, S. (2020). Implementation of Personalized Adaptive E-leaming System. *Baltic Journal of Modern Computing*, *8*(2), 293–310. https://doi.org/10.22364/bjmc.2020.8.2.06

Walkington, C., & Bernacki, M. L. (2020). Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. *Journal of Research on Technology in Education*, *52*(3), 235–252. https://doi.org/10.1080/15391523.2020.1747757

Xu, X., Li, Z., Hin Hong, W. C., Xu, X., & Zhang, Y. (2024). Effects and side effects of personal learning environments and personalized learning in formal education. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12685-0

Yang, Z., Zheng, Y., & Yang, Y. (2024). Recommendations for Personalised Learning Paths. *2024 13th International Conference on Educational and Information Technology (ICEIT)*, 36–40. https://doi.org/10.1109/ICEIT61397.2024.10540724

Zhang, L., Basham, J. D., & Yang, S. (2020). Understanding the implementation of personalized learning: A research synthesis. *Educational Research Review*, *31*, 100339. https://doi.org/10.1016/j.edurev.2020.100339

Zhang, Y., Xu, X., Zhang, M., Cai, N., & Lei, V. N.-L. (2023). Personal Learning Environments and Personalized Learning in the Education Field: Challenges and Future Trends. In C. Hong & W. W. K. Ma (Eds.), *Applied Degree Education and the Shape of Things to Come* (pp. 231–247). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-9315-2_13

Back to the Roots: Investigating the Drivers of Nostalgia in the Retro Gaming Experience

Simos Chari
Alliance Manchester Business School
Savvas Papagiannidis
Newcastle University Business School
Davit Marikyan
Newcastle University Business School
Dinara Davlembayeva
Newcastle University Business School

Research In progress

Abstract

In recent years, there has been a growing interest in classic games, highlighting nostalgia's role in enhancing the emotional appeal of retro games. Despite this, limited research has focused on what triggers nostalgia in players. To explore this further, we used grounded theory and inductive reasoning to investigate psychological traits linked to nostalgia in retro gaming. Through 19 interviews with retro game players, our study identified eight key drivers of game-induced nostalgia: social disconnectedness, the need for social stability, belongingness, feelings of meaninglessness, self-discontinuity, boredom, a desire for competence, and the tendency to escape reality. These findings expand existing theory by offering a deeper understanding of retro game consumption and distinguishing consumer profiles, thereby informing the retro gaming market.

Keywords: Digital games, retro games, nostalgia.

1.0 Introduction

There has been resurgence of interest in classic games, such as *Pokemon Go*, *Donkey Kong* and *Rayman* (Wulf et al., 2018), as evidenced by the relaunching of hardware gaming devices such as the Sega Genesis Mini (Malen, 2023) and higher demand for old games (Gilbert, 2021). The rekindled interest in bygone experiences draws on the feeling of nostalgia; "*a sentimental longing or wistful affection for the past*" (Pearsall & Hanks, 1998, p. 1226). However, the role of evoked nostalgia in gaming consumption remains unclear; we still lack understanding regarding the broader psychological conditions and the situational prerequisites that induce nostalgia, the implications of game-evoked nostalgia for game-related behaviour, and the implications of gamers' psychological differences for businesses.

To address these challenges, this study employs grounded theory to investigate and uncover more about game-evoked nostalgia before moving towards a solution (i.e., testing the resulting theory on the antecedents and implications of game-evoked nostalgia) (Ancillai et al., 2019). Through our rigorous discovery process, our study explores consumers' differences underpinning retro gaming experiences and provides a comprehensive model of the psychological characteristics and individual differences of gamers experiencing game-evoked nostalgia.

2.0 Discovery-oriented approach

The research adopted a discovery-oriented approach (e.g., Morgan et al., 2005), which allowed us to synthesise secondary findings with field-based interviews with retro gamers and conceptualise the model about the drivers of game-evoked nostalgia for testing. Our approach comprised three stages; each stage in the process provided an output that informed the subsequent stage of the approach (cf. Fattoum et al., 2024). The *first stage* comprised a literature review, aiming to identify the antecedents and outcomes of game-evoked nostalgia. The literature review resulted in six game-related nostalgia antecedents, namely *extrinsic self-focus*, *low self-esteem*, *boredom*, *self-discontinuity*, *meaninglessness* and *social disconnectedness*. Using the output of stage 1 we developed an interview protocol to deploy in stage 2. The protocol included open-ended questions aimed at uncovering gamers' emotions and thoughts related to retro-gaming, the psychological states during nostalgic episodes and behaviours resulting from game-evoked nostalgia. The *second stage* involved 19 interviews with retro gamers. These participants had over 15 years of gaming experience and played a variety of retro games, including titles such as Resident Evil, Bubble Bobble, Wizball, Mario, Crash Bandicoot, and the Pokémon series, among others. The interviews generated a dataset of approximately 18,000 words. By the conclusion of data collection, findings began to recur, indicating theoretical saturation (Strauss & Corbin, 1998). The interviews were transcribed, and the data were analysed using an open and axial coding system, following the framework proposed by Strauss and Corbin (1998). That process revealed four additional (i.e., *need for social stability*, *need for belongingness*, *need for competence* and *tendency to escape reality*) and two drivers (i.e., *extrinsic self-focus* and *low self-esteem*) that were irrelevant to retro gamers. The *third stage* synthesised outputs 1 and 2, and resulted in developing the study's conceptualisation of the drivers of nostalgia.

2.1. Stage 1: Nostalgia and retro gaming

Nostalgia works as a regulatory mechanism directed at coping with threats and helping individuals reach physiological and psychological homeostasis (Sedikides et al., 2008; Wildschut et al., 2011). When a feeling of nostalgia arises, the perceptions of threats get attenuated and individuals' defensive reactions to behaviour get weaker. The regulatory mechanism of nostalgia works for three life domains, namely the self-oriented, existential and sociality domains (Sedikides, Wildschut, Routledge, Arndt, et al., 2015).

When it comes to the sociality domain, nostalgia motivates socially oriented behaviour, such as social relationship building and social support (Jiang et al., 2021). The conditions triggering the sociality domain are social disconnectedness (Jiang et al., 2021), which is "*the discrepancy between individuals' desired and perceived social relationships"* (Wu et al., 2020, p. 3).

In the existential domain, nostalgia can increase the feeling of the meaningfulness of life and the preservation of personal identity, by derogating the meaning-undermining factors (Sedikides, Wildschut, Routledge, Arndt, et al., 2015). As such, the existential function is activated by a sense of meaninglessness, that is, an individual's subjective perception that life is without meaning (Routledge et al., 2011; Steger et al., 2006). Meaninglessness can be attributed to boredom (Van Tilburg et al., 2013) or to a feeling of self-discontinuity—a sense of disconnection between one's past and present self (Sedikides, Wildschut, Routledge, & Arndt, 2015).

The self-oriented regulation mechanism refers to the role of nostalgia in enhancing one's self-concept (Sedikides & Wildschut, 2020). Nostalgia "*sets up cognitive (i.e., self-positivity) barricades, both explicit and implicit, upon which threats to self may be deflected*" (Sedikides, Wildschut, Routledge, Arndt, et al., 2015). The activation of this mechanism is conditioned by low self-esteem—the situation when an individual loses a sense of personal control and mastery of their environment (Lyubomirsky et al., 2006). Low self-esteem depends on extrinsic self-focus, which refers to the tendency of a person to compare him/herself against external standards and expectations and underrate personal achievements (Lasaleta & Loveland, 2019).

Nostalgia is an integral aspect of retro games (Wulf et al., 2018). Retro games transfer individuals back to the past and stimulate consumption (Suominen, 2008). However, empirical investigation has only established links between the need for competence, relatedness and nostalgia (Wulf et al., 2020), and has found that age and in-game achievements define the level of nostalgia induction (Bowman et al., 2023). As such, an array of psychological factors within the self-oriented, existential and sociality domains, which could potentially facilitate nostalgia while gaming, remain under-explored.

2.2 Stage 2: Field-based insights

Drawing on the literature-based insights of stage 1, we developed an interview protocol/script. Using the Prolific platform, we recruited 19 gamers who had previous experience playing retro games. Each interview was recorded and lasted on average 30 minutes. The collected interview data amounted to approximately 18,000 words. Findings started to repeat themselves towards the end of data collection, signalling theoretical saturation. We transcribed the interviews using the methodology developed by Strauss and Corbin (1998), whereby we coded the words and concepts spoken by the interviewees, then coded the patterns of text sharing similar meanings and then grouped the codes denoting underlying relationships.

2.3 Stage 3: Synthesis of the literature and field-based insights

Our field-based insights demonstrate that the sociality function of nostalgia is activated under the conditions of social disconnectedness, the need for belongingness and the need for social stability. Social disconnectedness manifested itself as the feeling of loneliness, yearning for memories when participants were close to their family and friends and for people who were not alive anymore. As interviewee 3 mentioned: “*It's good memories, as I don't speak to my dad anymore. We've had quite a difficult relationship for a few years, so we don't really have any contact anymore. So it's remembering. They were good times.*” Also, based on field-based insights it appears that the realisation of one’s deficiency in social relationships is accompanied by the need to reconnect with people and share gaming experiences, e.g.: “*We all have the same similar experiences in the sort of games we played when we were growing up… I can talk to pretty much anyone my own age and they'll know of Pokémon*.” [Interviewee 19]. The need for social stability refers to individuals’ inner motivation to strive for structure and clarity in most situations and avoid ambiguity (Thompson et al., 2013). Gamers feel nostalgic about the games they played many years ago because they want to return to simpler times, as the current social system seems complicated. The following quote is illustrative of the opinions expressed: *“I think there is that level of association again in my mind of going back to a simpler time when I had* *fewer responsibilities and kind of more free time, more free money, and all the rest of that.* [Interviewee 5].

The factors underpinning the existential function of retro game-related nostalgia are meaninglessness, boredom and self-discontinuity. Our findings show that there is a sense of being lost and a sentiment that today's culture deprives us of spending meaningful time with families, e.g., interviewee 1: "*No meaning. I think it is a culture now where we're so busy all the time. It's just that bills need paying. I've got to go to work, that you forget to sit down and take some time and do stuff as a family*". Routine life makes participants long for an innocent childhood when they had free time and did not have a burden of responsibility. Interviewee 16 shares the following: *"I got a lot more responsibility, but if I think back to when I was nine years old, I'll go to school, come back from school and play video games... Seeing just old games pop up ... bring back some happy memories of me playing Sonic back in the day".* In addition, game-evoked nostalgia was stipulated by a situational state when individuals found themselves in a dull environment and felt unchallenged. To that end, interviewee 4 posited: "*I guess being bored of all the games that I've played recently… If anything, remembering how good they were playing the games, so I want to experience them again*."

Two factors underpinning the self-oriented function of retro game-related nostalgia include the need for competence and a tendency to escape reality while playing retro games. The need for competence reflected in gamers' desire to play a game that they had mastered before, as mentioned by interviewee 9: "*In one session, you can set a new high score, whereas with some of the modern games today, you have to play for many weeks and months just to get half-decent score*." The tendency to escape reality is manifested as a response to the hardships of present life. Playing retro games is believed to help alleviate what feels wrong and take the mind off problems, as interviewee 11 put it: "*Yeah. Sometimes I go and play video games, and I'm like, OK, I can tackle the rest of the day, or I can go back to the problem I had before ... it helps alleviate everything that feels wrong or feels imbalanced.*"

In conclusion, in line with the theory on the regulatory functions of nostalgia, the role of self-discontinuity, meaninglessness, boredom, social disconnectedness and the need to belong contribute to the arousal of retro game-evoked nostalgia. These literature-based findings were also validated by field-based insights. In addition, the need for competence, the need for social stability and a tendency to escape reality, emerged from our field-based insights as the key driving factors of retro game-evoked nostalgia. Even though the current body of literature identifies self-esteem (e.g., Sedikides et al., 2022; Sedikides, Wildschut, Routledge, Arndt, et al., 2015) and extrinsic self-focus (e.g., Baldwin et al., 2015; Lasaleta & Loveland, 2019) as key drivers of nostalgia, these were not found to be relevant for retro gamers. A possible reason is that self-esteem and extrinsic self-focus are influenced by socio-cultural norms and dynamics. Within the gaming context, these factors might have minimal impact on emotions, as retro-gaming behaviour is primarily motivated by a hedonic pursuit of personal enjoyment rather than a need to conform to socio-cultural standards. Hence, they were excluded from our theorisation. Thus, as shown, a synthesis of the existing knowledge and our field based data resulted in eight drivers of nostalgia related to sociality (i.e., social disconnectedness, the need for social stability and belongingness), self-oriented (i.e., meaninglessness, and self-discontinuity, and boredom) and existential domains (i.e., the need for competence, and a tendency to escape reality) (Figure 1).

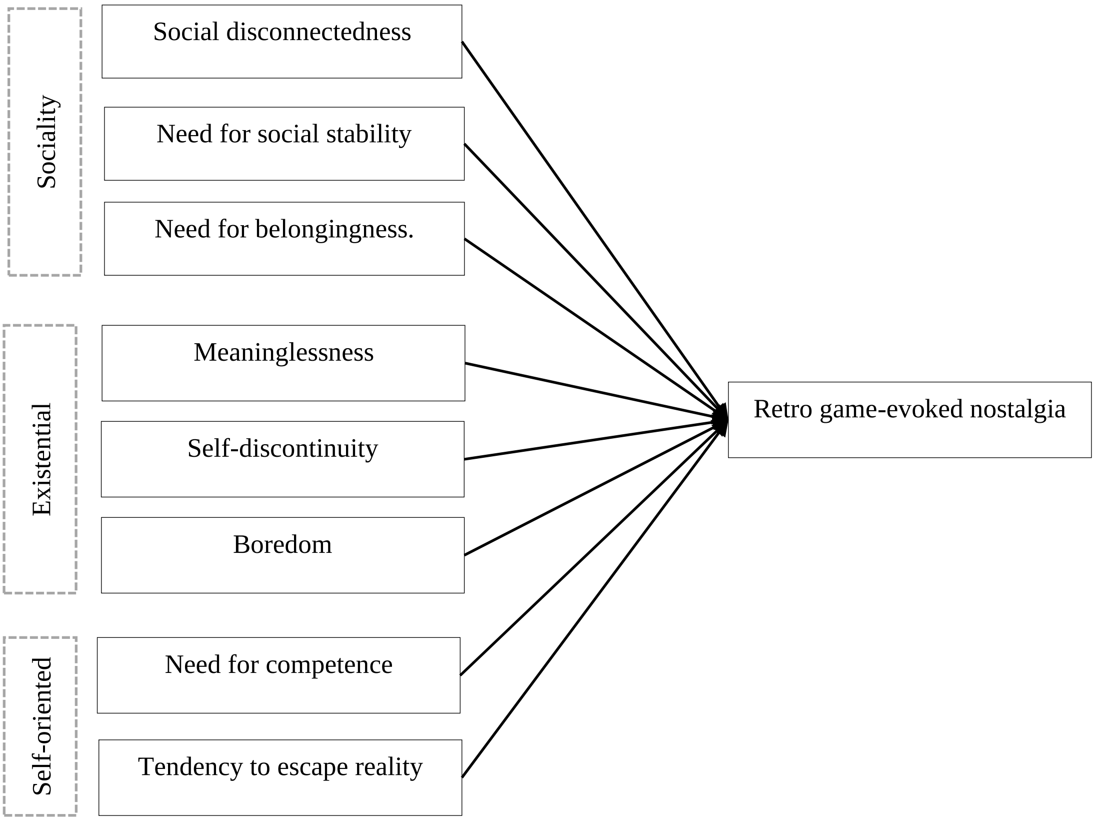


Figure 1. **The antecedents of retro game-evoked nostalgia**

3.0 Conclusions, further steps and potential implications

The results from the exploratory stage of the study reveal that game-evoked nostalgia is driven by eight factors pertaining to sociality, self-oriented, and existential domains. Subsequent stages of the study aim to validate the research model on the drivers of game-evoked nostalgia and moderating conditions that may influence the relationships between psychological factors and nostalgia with a sample of approximately 500 retro game players. Using the same sample, we also aim to examine the impact of game-evoked nostalgia on players' affective states (e.g., game satisfaction) and behaviours, such as intentions to purchase retro games, participate in retro-gaming communities, and watch others play retro games.

This study offers multiple contributions to the literature. First, it provides insights into the underlying factors of game-evoked nostalgia, contributing broadly to the literature on nostalgia (Reid et al., 2023), which has previously lacked evidence on the psychological conditions of nostalgia in the gaming context. Second, this research aims to enrich literature about the affective and behavioural outcomes of game-evoked nostalgia, shifting focus beyond psychological aspects of gaming (Wulf et al., 2020). Third, the study investigates behaviours beyond gameplay, acknowledging that gaming experiences may include passive experiences (e.g., watching others play) or active experiences, such as participating in user-driven communities or providing walkthroughs.

The findings have numerous practical applications: game developers, gaming platforms, and hardware manufacturers should consider not only how to launch the next AAA title but also how to serve and target a growing market segment that values retro gaming experiences.

References

Ancillai, C., Terho, H., Cardinali, S., & Pascucci, F. (2019). Advancing social media driven sales research: Establishing conceptual foundations for B-to-B social selling. *Industrial Marketing Management*, *82*, 293-308.

Baldwin, M., Biernat, M., & Landau, M. J. (2015). Remembering the real me: Nostalgia offers a window to the intrinsic self. *Journal of personality and social psychology*, *108*(1), 128.

Bowman, N. D., Velez, J., Wulf, T., Breuer, J., Yoshimura, K., & Resignato, L. J. (2023). That bygone feeling: Controller ergonomics and nostalgia in video game play. *Psychology of Popular Media*, *12*(2), 147.

Fattoum, A., Chari, S., & Shaw, D. (2024). Configuring systems to be viable in a crisis: The role of intuitive decision-making. *European Journal of Operational Research*, *317*(1), 205-218.

Gilbert, B. (2021). *Boosted by a record $2 million 'Super Mario Bros.' sale, the retro video game collector's market is being overrun by speculators looking to cash in*. Business Insider. https://www.businessinsider.com/retro-gaming-market-being-overtaken-by-speculators-2021-9?r=US&IR=T

Jiang, T., Cheung, W.-Y., Wildschut, T., & Sedikides, C. (2021). Nostalgia, reflection, brooding: Psychological benefits and autobiographical memory functions. *Consciousness and Cognition*, *90*, 103107.

Lasaleta, J. D., & Loveland, K. E. (2019). What's new is old again: Nostalgia and retro-styling in response to authenticity threats. *Journal of the Association for Consumer Research*, *4*(2), 172-184.

Lyubomirsky, S., Tkach, C., & DiMatteo, M. R. (2006). What are the differences between happiness and self-esteem. *Social Indicators Research*, *78*, 363-404.

Malen, K. (2023). *The Resurgance of Retro Gaming: Nostalgia and the Return of Classics*. Game Space. https://www.gamespace.com/all-articles/news/the-resurgence-of-retro-gaming-nostalgia-and-the-return-of-classics/

Morgan, N. A., Anderson, E. W., & Mittal, V. (2005). Understanding firms' customer satisfaction information usage. *Journal of Marketing*, *69*(3), 131-151.

Pearsall, J., & Hanks, P. (1998). *The new Oxford dictionary of English*. Oxford University Press.

Reid, C. A., Green, J. D., Buchmaier, S., McSween, D. K., Wildschut, T., & Sedikides, C. (2023). Food-evoked nostalgia. *Cognition and Emotion*, *37*(1), 34-48.

Routledge, C., Arndt, J., Wildschut, T., Sedikides, C., Hart, C. M., Juhl, J., Vingerhoets, A. J., & Schlotz, W. (2011). The past makes the present meaningful: nostalgia as an existential resource. *Journal of personality and social psychology*, *101*(3), 638.

Sedikides, C., Leunissen, J., & Wildschut, T. (2022). The psychological benefits of music-evoked nostalgia. *Psychology of Music*, *50*(6), 2044-2062.

Sedikides, C., & Wildschut, T. (2020). The motivational potency of nostalgia: The future is called yesterday. In *Advances in motivation science* (Vol. 7, pp. 75-111). Elsevier.

Sedikides, C., Wildschut, T., Arndt, J., & Routledge, C. (2008). Nostalgia: Past, present, and future. *Current directions in psychological science*, *17*(5), 304-307.

Sedikides, C., Wildschut, T., Routledge, C., & Arndt, J. (2015). Nostalgia counteracts self-discontinuity and restores self-continuity. *European journal of social psychology*, *45*(1), 52-61.

Sedikides, C., Wildschut, T., Routledge, C., Arndt, J., Hepper, E. G., & Zhou, X. (2015). To nostalgize: Mixing memory with affect and desire. In *Advances in Experimental Social Psychology* (Vol. 51, pp. 189-273). Elsevier.

Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The meaning in life questionnaire: assessing the presence of and search for meaning in life. *Journal of counseling psychology*, *53*(1), 80.

Strauss, A., & Corbin, J. (1998). Basics of qualitative research techniques.

Suominen, J. (2008). The past as the future? Nostalgia and retrogaming in digital culture. Proceedings of perthDAC2007. The 7th International Digital Arts and Cultures Conference. The Future of Digital Media Culture, Thompson, M. M., Naccarato, M. E., Parker, K. C., & Moskowitz, G. B. (2013). The personal need for structure and personal fear of invalidity measures: Historical perspectives, current applications, and future directions. *Cognitive social psychology*, 25-45.

Van Tilburg, W. A., Igou, E. R., & Sedikides, C. (2013). In search of meaningfulness: nostalgia as an antidote to boredom. *Emotion*, *13*(3), 450.

Wildschut, T., Sedikides, C., & Cordaro, F. (2011). Self-regulatory interplay between negative and positive emotions: The case of loneliness and nostalgia. *Emotion regulation and well-being*, 67-83.

Wu, A. F.-W., Chou, T.-L., Catmur, C., & Lau, J. Y. (2020). Loneliness and social disconnectedness in pathological social withdrawal. *Personality and Individual Differences*, *163*, 110092.

Wulf, T., Bowman, N. D., Rieger, D., Velez, J. A., & Breuer, J. (2018). Video games as time machines: Video game nostalgia and the success of retro gaming. *Media and Communication*(2), 60-68.

Wulf, T., Bowman, N. D., Velez, J. A., & Breuer, J. (2020). Once upon a game: Exploring video game nostalgia and its impact on well-being. *Psychology of Popular Media*, *9*(1), 83.

Exploring Trust Dynamics in Higher Education: A Comprehensive Analysis of Educators' Perceptions of Students' Ethical Adoption of Generative AI

Ishan Vats
UCL Centre for Systems Engineering, University College London, UK
ishan.vats.23@alumni.ucl.ac.uk

Chekfoung Tan
UCL Centre for Systems Engineering, University College London, UK
chekfoung.tan@ucl.ac.uk

Completed Research

Abstract

Generative Artificial Intelligence (GAI) present both transformative opportunities and complex ethical challenges in the evolving Higher Education (HE) landscape. This research explores the crucial aspect of trust among educators in HE regarding the ethical use of GAI, avital factor for its successful integration into teaching and learning environments. Through a survey research approach, this study combines quantitative and qualitative analysis to assess the levels of trust educators place in students' ethical use of GAI. The research examines key constructs such as transparency, reliability, accountability, cultural contexts, trust, and ethical alignment through descriptive and thematic analysis. This study explores two interrelated aspects: educator's trust in students' ethical use of GAI and educator's trust in GAI technology itself for teaching practices. The research posits that educator's trust in GAI may influence their trust in student's ethical use of the technology. By clarifying these distinct yet connected focuses, the findings reveal that trust is a critical lever in the adoption and effective use of GAI in HE. The research highlights how various dimensions of trust affect educators' engagement with GAI. These insights pave the way for the development of targeted guidelines aimed at strengthening trust and promoting an ethical framework for GAI in HE.

Keywords: Generative Artificial Intelligence, Trust, Ethical Use, Cultural Context, Higher Education, ChatGPT

1.0 Introduction

The genesis of Generative Artificial Intelligence (GAI) in educational contexts, dating back to the early 20th century, has evolved significantly with the advent of modern machine learning models like GPT-3 (Haenlein & Kaplan, 2019). These advancements, while facilitating personalised learning experiences, have sparked concerns regarding their potential to compromise academic integrity (Michel-Villarreal et al., 2023). Despite these concerns, UNESCO posits that GAI, when employed responsibly, can significantly enhance educational outcomes while adhering to ethical standards (UNESCO, 2023). The ethical use of GAI by students is particularly concerning, as it involves considerations of academic integrity, the appropriateness of GAI interactions, and the long-term implications of GAI on learning outcomes (Michel-Villarreal et al., 2023). Educators play a crucial role in this context, as students' trust in the ethical use of GAI directly impacts the adoption and effective integration of these technologies in educational practices. In addition, research by Tan et al. (2024) highlights a lack of trust in students' responsible use of GAI in their summative assessments. The integration of GAI in educational setting has raised significant questions around trust, both in the technology itself and in how students ethically use it. This study aims to dissect these dynamics by focusing on two interrelated but distinct aspects — educators' trust in students' ethical use of GAI and educators' trust in GAI as a teaching tool. This study proposes that educator's trust in the technology may directly influence their perception of students' ethical use, creating a layered trust dynamic. This perspective is grounded in recent studies highlighting the importance of educator's confidence in GAI as a precursor to trusting students responsible use of these tools (Lucas et al., 2024; Nazaretsky, Ariely, et al., 2022). Addressing these challenges requires a comprehensive understanding of the factors that influence educators' trust in the ethical use of GAI by students. This includes exploring how educators perceive the risks and benefits associated with GAI, their experiences with GAI technologies, and their attitudes toward the ethical implications of such tools in educational settings. Furthermore, as GAI continues to evolve, there is a need for ongoing research to develop robust frameworks and guidelines that can support educators in fostering an environment of trust and ethical responsibility in the use of GAI (Moorhouse et al., 2023). Consequently, the research question addressed in this paper is: *How does educators' trust in students' ethical use of GenAI influence the integration and efficacy of these technologies within higher education settings?*

This paper aims to explore the dynamics of trust between educators and students concerning the ethical use of GAI in HE. The research identifies key factors that affect trust, assesses their impact of GAI adoption, and proposes recommendations to foster a trust-rich environment that supports the ethical use of GAI. This paper is structured as follows. Section 2 covers the related work, followed by the research methodology in Section 3. Section 4 presents the results and Section 5 provides the relevant discussion. The paper concludes with research implications, limitations, and future work in Section 6.

2.0 Related Work

The literature review informing this study focuses on key themes around trust in GAI within HE. Relevant academic sources are identified through searches in databases such as Scopus, Web of Science, and Google Scholar using terms including “trust in AI,” “ethical AI in education,” “Generative AI,” “educators’ trust,” and “student use of AI.” The review focuses on peer-reviewed journal articles and conference papers exploring factors shaping trust in technology, particularly GAI, and its ethical considerations in educational settings. Inclusion criteria focus on studies examining trust in technology, ethical considerations in AI, and educator-student dynamics in HE settings. Empirical studies and theoretical frameworks are prioritised to ensure a balances and comprehensive review.

2.1 Generative Artificial Intelligence (GAI)

GAI has appeared as a transformative force within HE, marked distinctly by the introduction of OpenAI's ChatGPT in late 2022 (Bengio et al., 2000; Hinojo-Lucena et al., 2019; Radford, 2018). ChatGPT's widespread adoption has sparked broad discussions, extending the debate on GAI's impact far beyond academic circles. These discussions often reveal sharply divided opinions on technology's role in education, with some viewing it as potentially damaging ("doomsters"), while others ("boosters") see it as revolutionary (Selwyn, 2014). A systematic review found predominantly positive assessments of GAI's potential to enhance educational practices, with few addressing ethical concerns (Zawacki-Richter et al., 2019). Conversely, critiques focus on issues like the potential for corporate overreach, exemplified by concerns over automated plagiarism detection (Popenici & Kerr, 2017). Pre-ChatGPT discourse analyses by Bearman et al. (2023) highlight an urgent need for educational institutions to adapt, reflecting the shifting power dynamics GAI introduces into the learning environment.

In HE, GAI tools such as automated content generators, adaptive learning systems, and personalised assessment engines can significantly enhance the learning experience. While GAI can offer personalised learning experiences, there is a risk that these technologies could also worsen existing disparities in educational access and quality if not implemented thoughtfully. Ensuring that GAI tools are accessible to all students and do not favor certain groups over others is essential for their ethical integration into educational systems (Lacey & Smith, 2023).

2.2 Trust and Ethical Considerations in GAI

Trust is a foundational aspect of effectively integrating new technologies, particularly AI, in educational settings. Trust in technology is influenced by many factors, including reliability, predictability, and ethical considerations (Faulkner, 2010; Tschannen-Moran, 2014). Faulkner (2010) emphasises that trust in technology is not solely based on its functional reliability but also on its ethical design and transparency. Within educational setting, trust in GAI is a multifaceted issue. It encompasses not only the reliability and performance of technology but also the adherence to ethical standards, crucial for fostering a productive educational environment. In the context of GAI, ethical concerns such as data privacy, algorithmic bias, and the potential for misuse are paramount, affecting stakeholder trust (Dunn et al., 2021). Empirical research underlines several trust-enhancing factors specific to GAI in education, such as system reliability, user experience, and alignment with educational goals (Batista et al., 2024; Mogavi et al., 2023; Shahzad et al., 2024). Studies also suggest that educators' trust in AI-powered educational technology can influence their trust in students' ethical use of such tools and educators who trust the transparency and reliability of these tools are more likely to believe that students will use them ethically (Lucas et al., 2024; Nazaretsky, Ariely, et al., 2022). The integration of GAI into HE has initiated a profound transformation in pedagogical methods and student engagement. As these GAI systems evolve, characterised by their ability to generate content autonomously based on extensive data sets, they are increasingly deployed to personalise learning and streamline educational processes. However, the rapid advancement and integration of GAI raises significant trust and ethical considerations (Haenlein & Kaplan, 2019). While GAI offers potential enhancements in educational outcomes, the dynamics of trust between educators and students regarding the ethical use of these technologies require careful consideration (Jobin et al., 2019).

2.3 Conceptual Framework for Exploring the Dynamics of Trust

Figure 1 shows a conceptual framework that provides a structured approach to understanding the dynamics of trust and ethical considerations in the use of GAI within HE settings. Trust in GAI is the overarching theme that encapsulates educators' overall confidence in students' ethical use of GAI tools. Educators' trust in students' ability to use GAI ethically involves several layers of confidence and expectation, ranging from students' technical competence to their moral judgment and adherence to ethical guidelines. According to Nguyen et al. (2023), trust in technologies like GAI extends beyond its functional capabilities to include how it manages data and maintains integrity. This aspect is supported by Stahl (2021), who emphasises the need for transparency and accountability in GAI systems to secure educators' trust. This framework is developed through the literature review process, comprises six constructs, which are discussed in the sub-sections below. The conceptual framework is developed based on a thorough review of existing literature on trust in technology, ethical considerations in GAI, and educator-student dynamics within higher education.

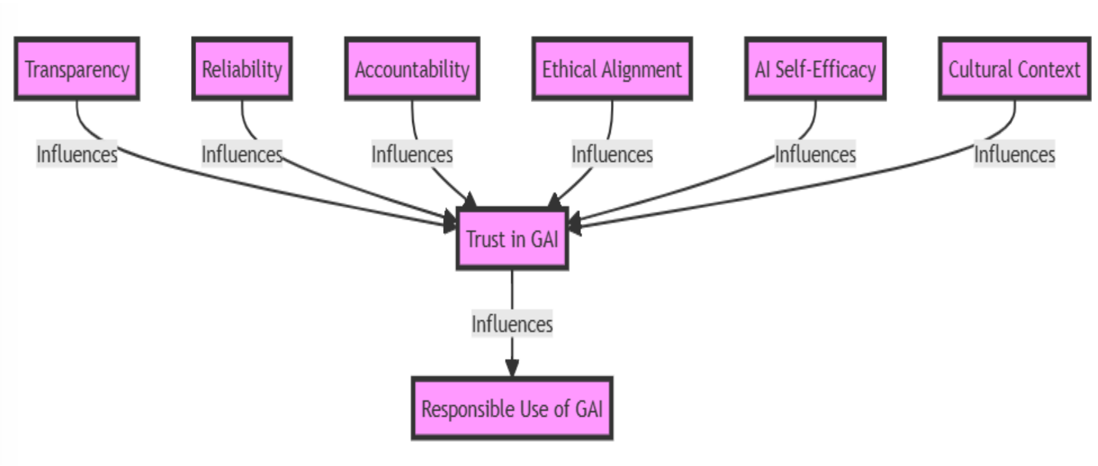


Figure 1. Conceptual Framework

Six key constructs are identified as central to understanding educators' trust in GAI. These constructs are selected due to their consistent presence in the literature. Each construct is further elaborated in the subsections below, detailing its relevance and role within the framework.

2.3.1 Transparency (T)

Ensuring that educators and students are aware of and understand these aspects can lead to more informed and confident use of GAI tools, thereby enhancing educational outcomes. It has been revealed through studies that algorithmic transparency supports ethical use and improves trust, which are considered indispensable factors for the effective adoption of GAI technologies (Lacey & Smith, 2023).

2.3.2 Reliability (R)

Reliability refers to the consistent performance of GAI systems, crucial for their trustworthiness. In HE, reliable GAI ensures that tools used for student assessments and personalised learning are dependable and accurate. Studies show that reliability influences educators' willingness to integrate GAI into their teaching practices significantly (Haenlein & Kaplan, 2019; Lacey & Smith, 2023). Reliable GAI systems contribute to an environment where both students and educators feel secure in their interactions with the technology, leading to more positive educational outcomes and a greater acceptance of GAI as a beneficial educational tool. The consistent performance and reliability of GAI systems plays a pivotal role in fostering trust (Stolpe & Hallström, 2024). Reliable systems are crucial for gaining educators' trust, as they need assurance that the technology will function as expected without frequent errors or failures.

2.3.3 Accountability (A)

Accountability in GAI supports the ethical use of technology by providing clear pathways for addressing misuse and managing ethical breaches (Turilli & Floridi, 2009). These structures help in maintaining a balance between innovation and ethical responsibility, ensuring that GAI tools benefit educational environments without compromising integrity or fairness. Thus, accountability is not just about having reactive measures but also about proactively ensuring that GAI systems operate within agreed ethical parameters, fostering a culture of trust and responsibility. Strong accountability frameworks are likely to enhance trust by providing assurances that the systems are under responsible oversight.

2.3.4 Ethical Alignment (EA)

Ethical alignment is about ensuring that GAI systems adhere to ethical norms and values, particularly in sensitive areas like education. It involves aligning GAI functionalities with societal and educational standards to ensure that their deployment enhances learning without compromising ethical standards (Celik et al., 2022; Matthias, 2004). The importance of ethical alignment is also emphasised by Eason

(2007), who argues that trust in technology is greatly enhanced when the technology demonstrably aligns with societal ethical standards. Furthermore, Nazaretsky, Cukurova, et al. (2022) discuss how ethical alignment influences educators' perceptions of GAI, suggesting that educators are more likely to adopt GAI technologies that transparently uphold ethical standards. Systems that are ethically aligned are presumed to foster greater trust among educators, as these systems reflect broader societal and educational standards.

2.3.5 GAI Self-Efficacy (SE)

GAI self-efficacy is crucial for fostering a proactive and positive interaction with GAI technologies in educational settings. It reflects the confidence educators and students have in their ability to effectively understand and use GAI tools. When educators possess high GAI self-efficacy, they are more likely to explore advanced features of GAI systems, apply them creatively in their pedagogy, and adjust their instructional strategies based on GAI feedback and analysis (Celik et al., 2022). Moreover, GAI self-efficacy extends beyond personal competence, impacting the overall educational ecosystem by promoting a culture of innovation and continuous improvement. Training programs that enhance GAI self-efficacy can significantly improve the adoption rates and effective use of GAI in educational settings, leading to better learning outcomes and more personalised educational experiences (UNESCO, 2023). Higher GAI self-efficacy among educators is expected to increase their trust in GAI, as they feel more competent and in control of the technology.

2.4.6 Cultural Context (CC)

The literature also addresses how cultural and institutional factors influence trust in AI. Research by Hofstede et al. (2014) provides a framework for understanding how cultural differences impact technology adoption and trust in GAI systems. Institutional trust, on the other hand, is shaped by educational policies, leadership attitudes, and the overall organisational culture surrounding technology use. Cultural norms significantly influence the acceptance and effectiveness of GAI in education. Yu et al. (2023) emphasise that understanding cultural differences is crucial in designing GAI tools that are sensitive to the diverse backgrounds of students. This sensitivity can enhance the relevance and usability of GAI applications, making them more effective across various cultural contexts. In addition, institutional policies and educators' attitudes toward technology significantly influence the extent to which these innovations are embraced (Bottery, 2004). Cultural context considers the influence of cultural norms and values on the perception and adoption of technology. It explores how cultural differences affect educators’ trust in AI, reflecting the diverse settings in which GAI is implemented (Holmes et al., 2022). Adapting GAI systems to align with local cultural norms requires a deep understanding of the specific educational ecosystem. Therefore, differences in cultural context can affect educators' trust in GAI. CC is included as a core construct at the same level as the others due to its significant influence on trust in AI technologies. Cross-cultural studies have demonstrated that cultural norms and values shape how individuals perceive and trust AI systems. Kaplan et al. (2023) conducted a meta-analysis revealing that trust in AI varies considerably across cultures, with German participants displaying higher trust levels compared to Japanese participants. Similarly, Agrawal et al. (2023) explored cross-cultural differences between OECD countries and India, finding notable variations in perceived trust, responsibility, and reliance on AI systems versus human experts. These findings shows that CC is not merely a moderating factor but a fundamental component which shapes the educators' overall trust in student's ethical use of GAI. Recognizing the diverse cultural backgrounds of educators is crucial for understanding how trust in GAI develops and how it influences the ethical use of these technologies in higher education.

3.0 Research Methodology

This research adopts a survey research approach from Check and Schutt (2011), which has been previously applied to study the use of technology in HE by Tan et al. (2023). This approach is designed to investigate educators' perceptions and ethical considerations of GAI in HE. Survey research gathers information from a sample of individuals through both quantitative responses, using numerical rated items, and qualitative insights via open-ended questions. The survey was developed based on the key constructs identified in the conceptual framework (see Figure 1), incorporating both a five-point Likert scale (1-Strongly Disagree; 2-Disagree; 3-Neutral; 4-Agree; 5-Strongly Agree) and open-ended questions to ensure a comprehensive data collection process. Administered via Microsoft Forms, the survey ran for four weeks, with responses collected anonymously. Departmental research ethics approval was obtained prior to data collection. The survey instrument was developed based on the conceptual framework, ensuring that each of the six constructs—Transparency, Reliability, Accountability, Ethical Alignment, GAI Self-Efficacy, and Cultural Context—was represented. For each construct, specific Likert-scale questions were designed to measure perceptions and attitudes. The items were formulated through an iterative process involving a review of relevant literature and expert input to ensure content validity and clarity.

For this study, a purposive sampling strategy was employed to target a specific subset of the population that possesses unique characteristics relevant to the research questions, the educators in HE who are engaged with or have perspectives on the use of GAI in their teaching environments. Participants were selected based on their involvement with educational technologies, including those who have either used GAI tools in their teaching practice or participated in workshops and seminars on GAI applications in education. The selection was facilitated through direct invitations sent via academic networks and professional social media platforms, such as LinkedIn and academic listservs related to educational technology. Additionally, snowball sampling techniques were utilised, where initial respondents were encouraged to recommend the survey to eligible colleagues, thus expanding the reach effectively within the academic community.

Following Tan et al. (2024), descriptive analysis was adopted to establish a baseline understanding of the data. This process included:

- *Mean Calculation*: Provided an average score for each question, indicating the overall trend or inclination of the respondents towards certain viewpoints on GAI.
- *Median Calculation:* Identified the middle value in the distribution of responses, which is particularly useful in understanding the central position of data in skewed distributions.
- *Standard Deviation:* Quantified the amount of variation or dispersion in responses, offering a clear picture of consensus or diversity in opinions among participants.

A thematic analysis was conducted on the open-ended responses gathered for each construct within the conceptual framework. This process involves steps such as immersing in the data to gain a deep understanding of the content and systematically coding the data in segments that highlight key features relevant to the research questions. Section 4 discusses the results.

4.0 Results

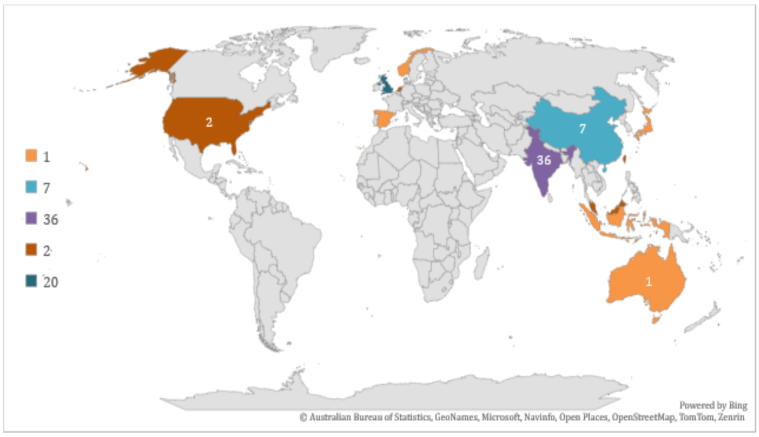
4.1 Demographics

The demographic breakdown of the 77 survey participants is presented in Table 1. The results indicate a diversity of age groups represented vary, providing a wide lens on the generational attitudes towards GAI. Young educators (21-30 years) make up 29.87% of the sample, suggesting a significant engagement from this demographic in GAI discussions. The 31-40 age group is the most represented at 37.07%, bringing a blend of youthful vigour and mature professional insight into the mix. With 68.83% male and 31.17% female participants, the gender distribution points towards a higher male engagement which might reflect broader trends in technology uptake and interest areas within academia. Table 1 also indicates that participants hail from both STEM (58.44%) and non-STEM (41.56%) fields, providing a balanced perspective from both technical and non-technical domains. This diversity is critical in evaluating the interdisciplinary implications of GAI tools. The experience levels among participants range from less than 2 years (23.37%) to over 10 years (24.67%), highlighting a mix of fresh insights and seasoned understandings within the educator community. The survey captured responses from educators in 11 different countries, with a notable majority from India (46.75%), followed by participants from China (9.09%), and a spread across other countries including the United States, United Kingdom, Australia, and several European and Asian nations. This global diversity is pivotal in assessing the cultural and regional nuances that might influence perceptions of GAI. Findings from underrepresented countries, like the single response from Norway, are included only when offering unique insights but not generalised.

The demographic diversity within the survey participants allows for a rich, multilayered analysis of the data. As shown in Figure 2, each demographic variable such as age, gender, professional background, teaching experience, and geographic location contributes to a more nuanced understanding of the factors that influence educators' trust in and use of GAI. The global spread of participants underscores the universal relevance of GAI discussions and the need for culturally aware educational technologies.

Characteristics	**Count (n)**	**%**
Age Group		
21-30	23	29.87
31-40	20	37.03
41-50	20	58.82
51-60	11	28.94
61 and older	3	3.75
Gender		
Female	24	31.16
Male	53	68.83
Teaching Domain		
Non-STEM	32	41.55
STEM	45	58.44
Teaching Experience		
<2 years	18	23.37
>10 years	19	24.67
2-5 years	19	24.67
5-10 years	21	27.27
Country		
Country	Count (n)	%
Australia	1	1.29
China	7	9.09
India	36	46.75
Indonesia	1	1.29
Japan	1	1.29
Malaysia	2	2.59
Netherlands	2	2.59
Norway	1	1.29
Spain	1	1.29
Taiwan	2	2.59
United Kingdom	20	25.97
United States of America	2	2.59

Table 1. Demographic Breakdown of Survey Participants



© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, Open Places, OpenStreetMap, TomTom, Zenrin

Figure 2. Global Distribution of Survey Respondents

The data gleaned from this demographic analysis not only frames the subsequent findings but also provides key insights into potential biases and areas of focused interest for future studies on the integration of GAI technologies in education. This detailed demographic overview is crucial for contextualising the attitudes and experiences that shape educators' perspectives on ethical GAI usage.

4.2 Descriptive Analysis

The descriptive analysis of the survey reveals nuanced insights into the educators' perceptions of GAI across six key constructs: Transparency (T), Reliability (R), Accountability (A), Self-Efficacy (SE), Cultural Context (CC), and Ethical Alignment (EA). Table 2 shows a detailed breakdown of each construct, focusing on the mean, median, and standard deviation.

The Transparency (T) construct, assessed through statements T1, T2, and T3, reveals a strong preference among educators for clear and understandable GAI decision-making processes. With means close to 4 (T1: 3.99, T2: 3.97, T3: 3.84) and consistent medians of 4, the data suggests that educators place significant value on the transparency of GAI tools. This preference underscores the importance they place on understanding how GAI influences and guides student decisions. The relatively low standard deviations (T1: 0.71, T2: 0.75, T3: 0.86) indicate a general agreement among participants, reinforcing the critical role transparency plays in fostering trust in educational technologies.

Reliability (R) is another crucial factor for educators, as evidenced by their responses to R1, R2, and R3. These items scored means of 3.77, 3.70, and 3.64 respectively, with all maintaining a median of 4, indicating a robust expectation for GAI tools to deliver consistent and accurate assistance. The standard deviations, hovering around 0.90, reflect a slightly broader range of opinions regarding the reliability of GAI, possibly due to varying subjective experiences with the technology. This variation suggests areas where GAI tools need to enhance their reliability to meet educator expectations fully.

In the Accountability (A) construct, statements A1, A2, and A3 explore the expectations for GAI systems to autonomously monitor and correct unethical behaviors. The means for these statements (A1: 3.71, A2: 3.57, A3: 3.74) with medians at 4, reflect a cautiously optimistic view among educators regarding the accountability mechanisms embedded in GAI tools. The somewhat higher standard

Construct	Statements	Mean	Median	Standard Deviation
T1	I trust GAI tools more when I understand how they guide students in making decisions.	3.98	4	0.71
T2	I can trust GAI tools when their influence on student choices is clear and transparent.	3.97	4	0.74
T3	I can accept the integration of GAI tools in educational settings when I am fully aware of the criteria these tools use to generate outputs for students.	3.84	4	0.85
R1	I trust GAI tools that provide consistent and error-free assistance to students.	3.76	4	0.90
R2	I can trust GAI tools when they deliver accurate information to students without fail.	3.70	4	0.88
R3	I am confident in the ethical use of GAI by students when the tools consistently function as intended.	3.63	4	0.95
A1	I can trust GAI tools equipped with effective mechanisms to monitor and correct unethical behaviors such as plagiarism, cheating, and data manipulation by students.	3.71	4	0.95
A2	I trust GAI tools that can report and rectify misuse by students autonomously.	3.57	4	1.00
A3	I am more confident in the ethical use of GAI tools by students when these tools feature audit trails and alerts for misuse.	3.74	4	0.96
EA1	I can trust GAI tools that reflect high ethical standards aligned with educational principles.	3.93	4	0.81
EA2	My trust in GAI tools depends on their capability to reinforce ethical behavior among students.	3.84	4	0.84
EA3	I value GAI tools designed with strong ethical considerations that reflect our educational values.	3.93	4	0.78
SE1	I can trust students using GAI tools ethically when I feel confident in my ability to oversee and understand these technologies.	4.12	4	0.67
SE2	My competence in using GAI tools correlates with my trust in students' ethical use of these technologies.	3.88	4	0.77
SE3	I am more trusting of GAI technologies overall when I am confident in my ability to use them effectively.	3.96	4	0.86
CC1	I can accept students using GAI tools ethically when these tools are consistent with the learning values upheld by my culture.	3.87	4	0.83
CC2	I trust my students to use GAI tools ethically when these technologies are	3.77	4	0.78

	perceived as beneficial for ethical academic practices within my cultural context.			
CC3	My cultural background’s definition of trust, which involves specific ethical behaviors, guides my evaluation of students' use of GAI tools."	3.75	4	0.84

Table 2. Summary of Survey Responses

deviations, especially for A2 (1.01), indicate diverse opinions about the effectiveness of these mechanisms, suggesting that while there is hope for robust accountability, there is also recognition of the challenges it faces.

In the Ethical Alignment (EE) construct, statements E1, E2, and E3 scored the highest means (E1: 3.94, E2: 3.84, E3: 3.94), indicating a strong consensus on the importance of aligning GAI tools with high ethical standards. These scores underscore educators' prioritisation of ethical considerations in GAI applications, reflecting a broad agreement that ethical alignment is paramount for the successful integration of GAI in educational settings.

Responses to the Self-Efficacy (SE) construct through SE1, SE2, and SE3 (means of 4.13, 3.88, and 3.96 respectively) emphasise the strong link between educators' confidence in their ability to oversee GAI and their trust in students using these tools ethically. The relatively low standard deviations indicate a consensus that personal competence in managing GAI technologies is crucial for ethical usage. This suggests that enhancing educator training and familiarity with GAI could further promote ethical practices among students.

Cultural Context (CC), assessed via CC1, CC2, and CC3, highlights how cultural norms and values shape the acceptance and implementation of GAI in education. With all means around 3.8 and medians consistently at 4, there is a clear recognition of the need for GAI tools to align with cultural learning values. The standard deviations suggest moderate variability in how educators from diverse cultural backgrounds perceive these issues, indicating a need for culturally sensitive approaches in the deployment of GAI technologies.

The descriptive analysis of the survey data reveals not just surface-level perceptions but also deeper patterns that relate directly to the research question and conceptual framework. For example, higher means in Transparency and Reliability indicate a general positive perception of GAI's functionality, aligning with the conceptual framework’s emphasis on these factors as foundational to trust. Meanwhile, the variability reflected in the standard deviations, especially within Accountability and Cultural Context, highlights diverse educator experiences and perspectives, suggesting that trust in GAI is influenced by both system attributes and external cultural factors. This complexity supports the inclusion of Cultural Context as a key construct in the framework. Moreover, correlations observed between GAI Self-Efficacy and trust in students' ethical use of GAI underscore the interdependence of educators' confidence in using GAI and their trust in students, directly linking the descriptive statistics back to the core research question. In summary, this analysis sheds light on the complex landscape of educators' perceptions regarding GAI, highlighting the crucial areas of transparency, reliability, accountability, self-efficacy, cultural context, and ethical alignment.

4.3 Thematic Analysis

The thematic analysis of educator responses gathered from the open-ended questions provides deep insights into the various constructs influencing their perceptions of GAI tools in education. The analysis followed an inductive coding approach, where responses were first open-coded to identify recurring ideas. These initial codes were then grouped into broader themes related to trust dynamics and ethical considerations. To ensure consistency, the coding process was reviewed and refined iteratively. Each comment reflects nuanced views that educators hold based on their experiences, expectations, and the theoretical underpinnings of GAI usage. Educators' discussions about the transparency of GAI tools reveal a complex interplay between the desired openness and the practical limitations of technology. The themes identified include *conditional trust*, *ethical implications*, and *practical implementation*. They express a *conditional trust*, encapsulated by remarks such as, "GAI can be trusted to a large extent but should not be 100%," highlighting the impossibility of achieving complete transparency with current technology. This notion is further reinforced by skepticism about whether GAI processes can ever be fully transparent, with one educator noting, "The problem is that this kind of transparency is not currently available."

Responses also delve into the *ethical implications* of transparency in GAI usage. Educators question the impact of transparency on ethical behavior and decision-making processes. Comments like, "There are too many ethical issues around GAI and its outputs for transparency alone to affect my levels of trust in these tools," illustrate the broader concerns regarding how transparency intersects with ethical considerations. Educators are wary of over-reliance on technology, fearing it may lead to complacency or misuse, particularly among students who might exploit the system's transparency.

Educators also expressed concerns about the *practical implementation* of transparency, pointing out that even when GAI tools are designed to be open, they may not fully account for the nuances of human interaction or educational needs. For instance, one comment, "Due to laziness students prone to use GAI tools as they procrastinate to study and have no choice," suggests that transparency in GAI tools is not enough to ensure their effective and ethical use. This sentiment is echoed in broader discussions about the need for GAI tools to be designed in ways that support holistic educational goals and foster genuine understanding and engagement among users.

The reliability of GAI tools is a significant concern for educators, with *accuracy* and *consistency* of these technologies in educational settings emerging as the key themes. One educator captured the essence of this concern, stating, "How accurate are the GAI tools? How many students realise the GAI is not 100% correct." This skepticism underscores a broader apprehension about the potential errors and the unforeseen consequences of relying too heavily on automated systems. The inherent uncertainty in GAI outputs, as noted by educators, challenges the trust educators place in these tools to perform flawlessly.

The accountability construct underpins the theme of *functional reliability*. Educators elaborate extensively on their concerns regarding the accuracy and *functional reliability* of accountability mechanisms within GAI tools. A common theme is the hypothetical scenario where these tools could autonomously detect and manage ethical breaches. As expressed by one educator, "If they were, theoretically, 100% able to spot plagiarism, then I might trust them." This comment highlights the ideal yet currently unattainable standard that would foster greater trust among educators. Educators also delve into the practical challenges of implementing accountability in GAI systems. They voice concerns about the lack of current solutions that could effectively manage the broad spectrum of potential ethical issues. For example, one educator pointed out, "There is nothing that currently works in this space," indicating a significant gap between the desired and actual capabilities of GAI tools. Another comment, "It's very difficult to agree that I would trust these sorts of mechanisms when I have no idea what would be considered inappropriate use," underscores the uncertainty and unease regarding the scope and effectiveness of accountability measures. "They shouldn’t be trusted as much as they pretend to be," one educator remarked, questioning the integrity and reliability of the AI's ethical judgments. This skepticism is rooted in a realistic assessment of current GAI capabilities, coupled with a cautious outlook on the potential future developments in GAI governance. Many educators are concerned about the impact of these accountability issues on educational outcomes. They emphasise the need for robust systems that can not only recognise ethical breaches but also educate and guide students effectively.

As for the Ethical Alignment construct, the thematic analysis reveals a theme advocating for establishing clear *ethical guidelines and frameworks* that guide the development and deployment of GAI tools in education. Educators stress the importance of these frameworks in helping design GAI tools that meet educational goals and adhere to ethical standards. This is seen in discussions about the necessity of GAI tools to be designed with strong ethical considerations that reflect educational values, ensuring that their deployment supports a fair and equitable educational environment.

Educators' responses on GAI Self-Efficacy construct reflect mixed feelings about their capability to effectively use and understand GAI tools. The key themes derived are *confidence* and *capability.* This construct relates to educators’ *confidence* in managing and leveraging GAI in educational settings, which directly influences their trust in these technologies. Many educators’ express uncertainty about their *capability* to keep pace with evolving GAI technologies, with comments like, "It's difficult for me to predict how my trust will go up or down as I get more expert, as I am currently at such a low skill level." Educators feel that as their understanding of GAI grows, so does their awareness of its limitations and potential ethical issues, which in turn affects their self-efficacy. The complexity of GAI systems and the constant evolution of these technologies can be daunting, leading to feelings of inadequacy in fully grasping their implications.

The cultural context construct highlights the *diverse and complexity* theme in how different educational environments perceive and integrate GAI tools. Educators’ responses highlight the *complexity* of applying a uniform technological solution across varied cultural landscapes. Some educators expressed uncertainty and difficulty in defining how cultural context influences trust in GAI, with remarks like, "I'm not sure that my working definition of Trust (i.e., can a person explain their reasoning to me) is a cultural feature, though I'm sure it might be." Responses also delve into the challenges of ensuring that GAI tools align with the cultural values and educational norms of different regions. Educators from various cultural backgrounds bring unique perspectives that influence their trust in and acceptance of GAI tools. This is particularly evident in comments reflecting on how cultural differences can affect the perception of technology's role in education. For example, one UK-based educator remarked, "This as a UK-based academic - learners will see cultural bias in outputs," indicating concerns about how universally GAI technologies can be applied without reinforcing existing biases or creating new disparities. Educators discuss the wide-ranging ethical implications of GAI, which extend beyond just the immediate educational applications. They highlight the complexities involved in aligning GAI tools with ethical standards that reflect broad educational and societal values. Comments like, "Since 'tools' covers a lot of variety, i.e., some that have processes embedded in them, and others that are more neutral, I have difficulty in thinking about their ethical alignment," express the challenges of ensuring GAI tools adhere to diverse ethical expectations. Many educators' express skepticism about the ability of GAI tools to inherently support ethical educational practices. Questions arise about whether the programming of GAI can adequately reflect ethical guidelines or if it merely serves functional purposes without deeper ethical considerations. Comments such as, "I'm not sure how GAI tools are expected to reflect or align with these values. In their programming? Their output?" highlight the ongoing debate about the role of GAI in reinforcing or undermining ethical standards in education.

5.0 Discussions

This research aims to explore how educators' trust in students' ethical use of GAI influences the integration and efficacy of these technologies within HE settings. In this research, descriptive analysis plays a pivotal role in unpacking the statistical dimensions of educators' responses to the students' ethical use of GAI in educational settings. The descriptive analysis of the survey responses reveals insights into educators' trust in the ethical use of GAI by students, closely aligning with themes discussed in the literature. For instance, the high mean scores for transparency-related items (T1, T2, T3) suggest that educators value clarity about how GAI tools operate and influence student decision-making. This finding resonates with literature emphasising transparency as crucial for fostering trust in educational technologies it highlights transparency being a cornerstone for trust, where educators must understand how GAI tools function to fully endorse their use in educational practices (Haenlein & Kaplan, 2019; Jobin et al., 2019). Similarly, the reliability scores (R1, R2, R3) underscore the importance educators place on the consistent and error-free performance of GAI tools, aligning with studies that highlight reliability as a foundational element for trust in technology (Lacey & Smith, 2023). The accountability dimensions (A1, A2, A3) also showed robust mean scores, indicating that features like monitoring and correcting unethical behaviour are critical for educators' trust. This mirrors academic discussions that advocate for robust accountability mechanisms in GAI systems to ensure ethical usage (Tan et al., 2024; UNESCO, 2023). Furthermore, the significant scores related to GAI self-efficacy (SE1, SE2, SE3) highlight that educators' confidence in using GAI tools effectively influences their trust in students' ethical use of these technologies, confirming research that connects self-efficacy with technology adoption (Nazaretsky, Cukurova, et al., 2022).

Addressing the intricacies of the cultural context in GAI use in education, the thematic analysis highlights significant educator confusion and the need for a culturally nuanced understanding, as evidenced in participant comments. The analysis reveals a significant variation in perceptions across different age groups. For instance, participants aged 61 and older exhibit a higher trust in GAI's ethical alignment and self-efficacy, with means peaking at 4.33 and 4.22 respectively, suggesting an optimistic acceptance potentially due to their extensive experience and possibly greater exposure to varied technological transitions. Conversely, younger educators, particularly those between 21-30 years, show higher receptivity towards transparency and self-efficacy in GAI tools, indicated by their respective means of 4.03 and 4.00. This demographic's higher engagement with technological innovations might explain their comfort with and trust in transparent and self-efficacious GAI tools. Culturally, the insights gleaned from international educators point toward significant disparities. Educators in Japan and Spain rate their trust in the ethical alignment of GAI exceptionally high, possibly reflecting cultural nuances that favor technological integration and ethical compliance. In contrast, educators from the Netherlands and Norway present lower averages across constructs like self-efficacy and cultural context, which might stem from different educational priorities or societal values about technology usage in education. On the other hand, the cultural context appeared as a significant factor, with diverse responses showing that the cultural alignment of GAI tools affects their acceptance and effectiveness. This variability across cultural settings shows the importance of designing GAI tools sensitive to the cultural and regional nuances of the educational environments they are intended to serve. Lastly, teaching experience itself modulates trust in GAI, with those having less than two years of experience showing higher enthusiasm for ethical alignment and reliability of GAI tools, possibly due to their recent exposure to and training in newer educational technologies during their formative years. In contrast, educators with over ten years of experience might rely more on traditional methods or exhibit cautious optimism toward innovative technologies. As per the comparative analysis, a notable variation is revealed across different age groups. For instance, participants aged 61 and older exhibit a higher trust in GAI's ethical alignment and self-efficacy, with means peaking at 4.33 and 4.22 respectively, suggesting an optimistic acceptance potentially due to their extensive experience and possibly greater exposure to varied technological transitions. Conversely, younger educators, particularly those between 21-30 years, show higher receptivity towards transparency and self-efficacy in GAI tools, indicated by their respective means of 4.03 and 4.00. Culturally, the insights gleaned from international educators point toward significant disparities. Educators in Japan and Spain rate their trust in the ethical alignment of GAI exceptionally high, possibly reflecting cultural nuances that favor technological integration and ethical compliance. In contrast, educators from the Netherlands and Norway present lower averages across constructs like self-efficacy and cultural context, which might stem from different educational priorities or societal values regarding technology usage in education. Lastly, teaching experience itself modulates trust in GAI, with those having less than two years of experience showing higher enthusiasm for ethical alignment and reliability of GAI tools, possibly due to their recent exposure to and training in newer educational technologies during their formative years. In contrast, educators with over ten years of experience might rely more on traditional methods or exhibit cautious optimism toward new technologies.

6.0 Conclusion

6.1 Research Implications

The contributions of this research are twofold. From a theoretical perspective, this study identifies key constructs that provide a structured approach to understanding various facets of trust in GAI within educational settings, which is essential for its ethical use. Consequently, this research adds to the body of knowledge on responsible AI in information systems, covering both social and technical perspectives in line with Vassilakopoulou et al. (2022). From an empirical perspective, this research proposes specific trust-enhancement guidelines aimed at fostering a deeper understanding and acceptance of GenAI technologies within educational environments (see Table 3).To fully realise the potential of GAI in education, it is critical to establish strategic standards that assure ethical and trustworthy application. These ideas seek to improve the integration of GAI technologies, creating an environment in which both instructors and students may receive help from sophisticated technical resources while following high ethical standards. Trust stays a cornerstone of successfully deploying GAI tools in educational settings, needing the development of comprehensive guidelines to enhance this trust among educators, administrators, and students. These guidelines should focus on improving transparency, reliability, and ethical standards to ensure that GAI tools align with educational values and meet the needs of all stakeholders effectively.

6.2 Limitations and Future Work

One key limitation of this research is the reliance on self-reported data, which introduces potential biases and may not fully capture the complexities of educators' trust dynamics. Additionally, the rapidly evolving nature of GAI technology and its applications in education may affect the longevity of these findings. Future research could address these limitations by incorporating more objective data collection methods and continuously updating the research framework to align with technological advancements. Furthermore, longitudinal studies could track changes in educators' trust in GAI over time, providing deeper insights into the temporal dynamics of trust development. Further exploration of the impact of socio-economic factors on GAI adoption, along with comparative studies across different educational systems and cultural contexts, could enhance understanding of global perspectives on GAI, particularly from the trust perspective.

Guidelines	**Descriptions**
Comprehensive Educator Training	Develop detailed training programs for educators that focus on the ethical use of GAI, understanding potential biases, and the critical integration of GAI tools. This education enhances trust by increasing educators' control over and competence with these technologies.
Transparent Reporting and Feedback Mechanisms	Establish clear transparency guidelines that include mechanisms for educators to provide feedback on GAI tools. This helps in refining the tools based on actual user experiences, thereby improving their reliability and trustworthiness.
Ethical Standards and Regulation Development	Advocate for and help develop ethical standards that address crucial aspects like data privacy, algorithmic transparency, and fairness in GAI outcomes. Setting these standards builds trust by ensuring GAI tools are safe and fair for educational use.
Regular Updates and Continuous Learning	Encourage ongoing updates and learning opportunities about the latest developments in GAI. Keeping educators informed helps maintain their confidence in using these technologies, thus enhancing trust.
Stakeholder Involvement in GAI Development	Include a broad range of stakeholders in the development and evaluation of GAI tools to ensure these technologies are well-suited to the educational contexts they will be used. Participatory design processes increase trust by aligning the tools more closely with user needs and expectations.

Table 3. Trust-Enhancement Guidelines for the Ethical use of GAI in Higher Education

A key limitation of this study is the overrepresentation of respondents from India (46.75%), which may introduce cultural bias into the findings. This skew could affect the generalizability of results, particularly in trust perceptions influenced by cultural norms. Future research should aim for a more balanced sample through targeted outreach and stratified sampling to capture diverse cultural perspectives. Additionally, the age range of respondents (21–30 years) presents a potential limitation, as it is unclear whether all younger participants held formal teaching roles. Without specific validation of respondents’ academic positions, there is a possibility that some responses came from individuals in supporting roles, such as teaching assistants or graduate students involved in instructional activities. This may have introduced variability in the perspectives on GAI use in higher education. Further research should aim for a more balanced sample through targeted outreach and stratified sampling to capture diverse cultural perspectives and ensure clearer respondent validation.

References

Agrawal, V., Kandul, S., Kneer, M., & Christen, M. (2023). From OECD to India: Exploring cross-cultural differences in perceived trust, responsibility and reliance of AI and human experts. *arXiv preprint arXiv:2307.15452*.

Batista, J., Mesquita, A., & Carnaz, G. (2024). Generative AI and Higher Education: Trends, Challenges, and Future Directions from a Systematic Literature Review. *Information*, *15*(11), 676.

Bearman, M., Ryan, J., & Ajjawi, R. (2023). Discourses of artificial intelligence in higher education: A critical literature review. *Higher Education*, *86*(2), 369-385.

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, *13*.

Bottery, M. (2004). Trust: Its importance for educators. *Management in education*, *18*(5), 6-10.

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends*, *66*(4), 616-630.

Check, J., & Schutt, R. K. (2011). *Research methods in education*. Sage Publications.

Dunn, B., Jensen, M. L., & Ralston, R. (2021). Attribution of responsibility after failures within platform ecosystems. *Journal of Management Information Systems*, *38*(2), 546-570.

Eason, K. (2007). Local sociotechnical system development in the NHS National Programme for Information Technology. *Journal of Information Technology*, *22*(3), 257-264.

Faulkner, P. (2010). *Norms of trust*. OUP Oxford.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5-14.

Hinojo-Lucena, F.-J., Aznar-Díaz, I., Cáceres-Reche, M.-P., & Romero-Rodríguez, J.-M. (2019). Artificial intelligence in higher education: A bibliometric study on its impact in the scientific literature. *Education Sciences*, *9*(1), 51.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2014). Cultures and organizations: Software of the mind.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., & Bittencourt, I. I. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 1-23.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human factors*, *65*(2), 337-359.

Lacey, M. M., & Smith, D. P. (2023). Teaching and assessment of the future today: higher education and AI. *Microbiology Australia*, *44*(3), 124-126.

Lucas, M., Zhang, Y., Bem-haja, P., & Vicente, P. N. (2024). The interplay between teachers' trust in artificial intelligence and digital competence. *Education and Information Technologies*, 1-20.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*, 175-183.

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT. *Education Sciences*, *13*(9), 856.

Mogavi, R. H., Deng, C., Kim, J. J., Zhou, P., Kwon, Y. D., Metwally, A. H. S., Tlili, A., Bassanelli, S., Bucchiarone, A., & Gujar, S. (2023). Exploring user perspectives on chatgpt: Applications, perceptions, and implications for ai-integrated education. *arXiv preprint arXiv:2305.13114*.

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, *5*, 100151.

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British journal of educational technology*, *53*(4), 914-931.

Nazaretsky, T., Cukurova, M., & Alexandron, G. (2022). An instrument for measuring teachers' trust in AI-based educational technology. LAK22: 12th international learning analytics and knowledge conference,

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.-P. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies*, *28*(4), 4221-4241.

Popenici, S. A., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and practice in technology enhanced learning*, *12*(1), 22.

Radford, A. (2018). Improving language understanding by generative pre-training.

Selwyn, N. (2014). *Digital technology and the contemporary university: Degrees of digitization*. Routledge.

Shahzad, M. F., Xu, S., & Zahid, H. (2024). Exploring the impact of generative AI-based technologies on learning performance through self-efficacy, fairness & ethics, creativity, and trust in higher education. *Education and Information Technologies*, 1-26.

Stahl, B. C. (2021). *Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies*. Springer Nature.

Stolpe, K., & Hallström, J. (2024). Artificial intelligence literacy for technology education. *Computers and Education Open*, *6*, 100159.

Tan, C., Alhammad, M. M., Long, S. H., Casanova, D., & Huet, I. (2023). Applying the UTAUT2 model to determine factors impacting the adoption of Microsoft Teams as an online collaborative learning tool. *International Journal of Smart Technology and Learning*, *3*(3-4), 300-324.

Tan, C., Alhammad, M. M., & Stelmaszak, M. (2024). Teachers' Perceptions of Students' Use of Generative AI in Summative Assessments at Higher Education Institutions: An Exploratory Study. UK Academy for Information Systems Conf erence (UKAIS2024), Kent Business School, UK.

Tschannen-Moran, M. (2014). *Trust matters: Leadership for successful schools*. John Wiley & Sons.

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*, 105-112.

UNESCO. (2023). *ChatGPT and Artificial Intelligence in Higher Education*. https://www.iesalc.unesco.org/wp-content/uploads/2023/04/ChatGPT-and-Artificial-Intelligence-in-higher-education-Quick-Start-guide_EN_FINAL.pdf

Vassilakopoulou, P., Parmiggiani, E., Shollo, A., & Grisot, M. (2022). Responsible AI: Concepts, critical perspectives and an Information Systems research agenda (Special Issue Editorial). *Scandinavian Journal of Information Systems*, *34*(2).

Yu, X., Xu, S., & Ashton, M. (2023). Antecedents and outcomes of artificial intelligence adoption and application in the workplace: the socio-technical system theory perspective. *Information Technology & People*, *36*(1), 454-474.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 1-27.

Energy use information and traffic light technology: Exploring user engagement with smart sockets

Joanne Swaffield (Newcastle University), Savvas Papagiannidis (Newcastle University) and Diana Gregory-Smith (Newcastle University)

Research in Progress

Abstract

This paper explores how energy consumption feedback can influence user decisions when cost is not an issue. Smart sockets can measure energy use and carbon emissions as well as indicating the source of the electricity coming into the socket via an LED traffic light system. Drawing on a study with university students for whom cost was not a consideration as it was included in the monthly rental fee, the paper investigates how users interact with these devices, focusing on three key areas: (i) drivers for and barriers to adoption, (ii) comprehension of feedback mechanisms, and (iii) potential for behaviour change aimed at reducing carbon emissions. Our findings indicate that participants appreciated the LED feedback,but were hindered by limitations in data granularity and the device's size. Most users were open to adopting such technology, if the cost of the device was not significant, though some confusion around LED and graphical feedback highlighted the need for enhanced user guidance.

Keywords: smart technology; technology adoption; sustainability; user engagement

Introduction

The past two decades have witnessed an unprecedented growth in the availability of smart technologies for the home, which can provide information about levels and patterns of energy use, as well as indicating periods of peak demand (Vasicek *et al.*, 2018). This ability to monitor and manage energy consumption has significant implications for consumers, who are dealing with challenges such as the cost-of-living crisis and the reality of climate change. Indeed, smart devices can be used to assess where and how people can reduce their bills (Balta-Ozkan 2013), facilitating a reduction in carbon emissions (Hledik 2009). Historically, however, research has focused on the technological aspects of the smart home, with only more recent studies considering the user perspective (e.g, Gøthesen *et al.* 2023). Where research has considered the use and effectiveness of feedback, data has generally focused on overall household energy use via smart meters (e.g., Gumz and Fetterman 2024).

The paper will address these gaps by focusing on user engagement and looking at it in the context of a particular type of smart technology, namely smart plug sockets. Specifically, it will assess the potential for this device to reduce carbon emissions by exploring its appeal, comprehensibility, and effectiveness in the absence of cost as a motivating factor. The paper will contribute to the existing literature in two main ways: a) by focusing on user engagement with a specific smart device and b) by considering the impact of information about sustainable energy use. The paper will begin with an outline of the research context, followed by a brief methodology and the initial findings, before concluding with a short discussion.

Research Context

Smart sockets are electrical devices that connect ordinary appliances to the internet and perform a range of different functions. Some of these sockets are used for the sole purpose of controlling one or more appliances remotely (Lin *et al.* 2018), while others deliver information about power consumption and time of use (TOU) (Kang *et al.* 2016). This kind of device can therefore play an integral role in facilitating one of the key functions of a smart home: to reduce energy use and promote environmental sustainability (Chen et al., 2017).

Adoption of Smart Technologies

The importance of sustainability as a factor in technology adoption is a subject of contention. In many studies, *environmental concern* was found to increase willingness to adopt (Stopps and Touchie 2021). However, Girod *et al.* (2017) argue that the most environmentally conscious are likely to hold a post-materialistic worldview and may therefore reject new green technologies. In addition, there are a range of other factors that are reported to influence intention to adopt. These include *perceived usefulness*, *perceived ease of use* (Ande *et al.* 2020) and *familiarity* with the concept, device or system (Baudier *et al.* 2020). In contrast, concerns around *privacy* and *security* are reported to have a negative effect on intention to adopt (Milchram *et al.* 2018). Finally, *cost* could be a driver for adoption where devices are perceived as a tool to help people save money (Larson and Gram-Hanssen 2020), but also a barrier where the upfront price is prohibitive (Sovacool *et al.* 2021).

User Engagement with Smart Technologies

Consistent with research on smart homes more generally, studies of plugs and sockets tend to focus on the technological attributes, while overlooking the users who are expected to engage with the technology (e.g., Rokonuzzaman *et al.* 2022). The limited research on activity specific feedback (cf. Oh 2020) indicates that a smart socket can lead to a reduction in energy use, but that this may be contingent on the provision of detailed information about the functions and features of the device. Similarly, research on smart meters suggests that feedback can reduce consumption, if it is delivered on a regular basis (Fischer 2008), with information that is “accessible, legible and interactive” (Yang *et al.* 2019).

User engagement research to date has focused on the amount of energy consumed and the time of use. The device in this study provides information about the latter as well as indicating the source of the electricity coming into the socket (renewable or non-renewable), thereby providing an additional way to assess the importance of environmental concerns in the context of energy use. There are examples of energy use traffic light systems (Stinson et al. 2015; Ivanov et al. 2013), but there is currently nothing that relays information about the source of energy. The paper will explore the potential for behaviour change where sustainability is the main consideration. It will assess the comprehension and effectiveness of activity specific feedback as well as responses to the energy source LEDs as a hitherto unresearched feature of smart home technology.

Methodology

The study focused on students living in ensuite rooms that share a common kitchen. In such student accommodation the monthly rental is inclusive of utility bills. Hence, students have no cost incentive to reduce their energy consumption, making it possible to focus on the rest of the factors of interest. The project was advertised via existing student communication channels (e.g., weekly emails, social media accounts) and a shopping voucher was offered as an incentive for those who completed the study.

10 women and 11 men from accommodation across the university were invited to take part. They ranged from first year undergraduates to PhD candidates and came from a variety of different disciplines (e.g., law, medicine, business, mechanical engineering).

All participants engaged in a 30-minute introductory interview, exploring their familiarity with smart technology and their views on a range of different issues (e.g., environmental behaviour, the cost-of-living crisis). They were then provided with a smart socket and a router to use in their bedroom for a period of three weeks. The smart sockets themselves are free-standing box shaped devices with four power outlets on each. Wi-Fi routers were also deployed in order to send data directly from the sockets to a dashboard that could be accessed by the research team. The students were asked to leave the socket and router plugged in for the duration of the study and always to use the same device in the same power outlet.

For the first seven days, the socket was the equivalent of a basic extension cord with none of the features activated. At the end of the first week the LEDs were switched on and an email was sent to participants to explain the significance of the different colours (green indicating that energy is coming from renewable sources, red indicating non-renewables and amber signalling a mixture of both.). At the end of the second week the students were sent a basic report on their energy consumption in the form of a graph.

Upon completion of the study, students were invited to engage in a 30-minute follow-up interview to share their thoughts and feedback on the device and their experience. All interview protocols were developed using existing literature on technology adoption and user engagement with smart technology. Due to the number of sockets and routers available, the study was conducted in three different phases over a period of 6 months (February-July 2024).

Initial Findings

This section will assess the appeal of the smart socket, the students' comprehension of the information it provided and its impact on their behaviour.

Motivators and Barriers

Most of the students talked about the appeal of the smart sockets in terms of making life easier, as well as being *interesting* and *cool*. However, many participants felt that the socket was too large and cumbersome. There were also many complaints about the brightness of the LED indicators, with several students getting in touch during the study to ask if and how they could safely cover them during the night.

Some participants referred to environmental motivations for conserving energy. They expressed concern about the environment and noted that sustainability was a consideration for them in their everyday behaviours. Interestingly, however, when participants were asked about environmental behaviours more generally, reducing energy was much lower on the list than actions such as recycling and taking public transport, sometimes not being mentioned at all. Despite the fact they were not paying directly for energy, a number of students highlighted the importance of cost as a motivator for reducing consumption. There was a strong awareness of the price of energy and the general importance of conserving it for this reason, even though it did not apply to them at the time. Many claimed that they would be inclined to purchase the device in the future when they were paying bills and keeping a closer eye on their energy use.

Comprehension of Feedback Mechanisms

Once the LEDs were enabled at the end of the first week, almost all of the participants became aware of them. Some were confused about what the different colours meant, with one wrongly believing that the red light meant they had used too much energy and another worrying that the red light indicated a problem with the device. Generally, however, the participants understood that a green light indicated clean energy. There was also some confusion around the information provided in the graph. One student in particular stated that the graph was easy and obvious to interpret but proceeded to demonstrate a lack of comprehension. However, many participants did understand the graph and discussed high use items and the days/times they were consuming the most energy. There were several calls for more granular data to make the consumption information more detailed and easier to understand.

Potential for Behaviour Change

Generally, participants found the information provided by the LEDs of interest and a number reported waiting for the green light before plugging things in. However, others reiterated the necessity of using their devices and appliances at particular times and claimed that they could not wait for the green light before using energy. One participant stated that he felt better if it was green, but he could not do anything about it. Some participants said that they had tried to work out a pattern to the light changes and felt it would have had a bigger impact on their behaviour if they had known *when* the LEDs would be green.

The graph which was distributed at the end of the second week appeared to have less impact on participant behaviour. Most people stated that the information about their energy use had little impact on the choices they made when it came to plugging in devices and appliances. Several participants remarked that they had expected the socket to do more, specifically, that it should itself contribute to a reduction in energy use, by turning devices off automatically. They felt this would have had a bigger impact on their behaviour. Such functionality was a feature that the smart plugs offered, but in order to avoid inadvertently affecting participants' day to day activities such rules were not applied.

Contributions

This paper makes an important contribution to the growing body of literature on user engagement with smart technologies. Differentiating itself from studies where cost and financial motivations are a consideration, the paper has explored how participants react to feedback about energy use when the only ‘cost’ is the impact on the environment. Notably, it explores responses to how much electricity is being used *and* the source of that electricity. Initial findings suggest that, although sustainability was a concern for the students, they would be much more likely to adopt the technology when cost was once again a factor in their decision making. Moreover, the perceived need for energy use at certain levels and particular times of the day could not be challenged when sustainability was the only concern. Finally, the effectiveness of the device under any circumstances is contingent on a better understanding of the information that is being relayed to the user. Future research might usefully take these qualitative insights as a starting point and explore technology adoption, comprehensibility and behaviour change at a much larger scale.

References

Ande, R. *et al.* (2020) *Internet of things: Evolution and technologies from a security perspective,* Sustainable Cities and Society, 54 1-15.

Balta-Ozkan, N. (2013) *The development of smart homes market in the UK*, Energy, 60 361-372.

Baudier, P. (2020) Smart home: Highly-educated students' acceptance, Technological Forecasting & Social Change, 153 1-19.

Chen, C. et al. (2017) Between the technology acceptance model and sustainable energy technology acceptance model: Investigating smart meter acceptance in the United States, Energy Research & Social Science, 25 93–104.

Fischer, C. (2008) Feedback on household electricity consumption: a tool for saving energy? Energy Efficiency, 1 79–104.

Girod, B. et al. (2017) Economic versus belief-based models: Shedding light on the adoption of novel green technologies, Energy Policy, 101 415–426.

Gøthesen, S. et al. (2023) Empowering homes with intelligence: An investigation of smart home technology adoption and usage, Internet of Things, 24, 1-21.

Gumz, J. and Fetterman, D.C. (2024). User’s perspective in smart meter research: State-of-the-art and future trends, Energy & Buildings, 308 1-75.

Hledik, R. (2009) *How Green Is the Smart Grid?* The Electricity Journal, 22 (3) 29-41.

Ivanov, C. *et al.* (2013) *Enabling technologies and energy savings: The case of Energy Wise Smart Meter Pilot of Connexus Energy,* Utilities Policy 26 76-84.

Kang, B. (2016) *Design and Implementation of Personal ICT Asset Management Service Based on Smart Power Socket*, IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA.

Larsen, S. and Gram-Hanssen, K. (2020) *When Space Heating Becomes Digitalized: Investigating Competencies for Controlling Smart Home Technology in the Energy-Efficient Home*, Sustainability, 12 1-21.

Lin, Y. *et al.* (2018) *MorSocket: An Expandable IoT-Based Smart Socket System,* IEEE Access, 6 1-10.

Milchram, C. *et al.* (2018) *Moral values as factors for social acceptance of smart grid technologies*, Sustainability, 10 1-23.

Oh, J. (2020) *IoT-Based Smart Plug for Residential Energy Conservation: An Empirical Study Based on 15 Months' Monitoring,* Energies, 13 1-13.

Rokonuzzaman, M.D. (2022) *IoT-based Distribution and Control System for Smart Home Applications, IEEE 12th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, Penang, Malaysia.

Sovacool, B.K. *et al.* (2021) *Knowledge, energy sustainability, and vulnerability in the demographics of smart home technology diffusion*, Energy Policy, 153 1-17.

Stinson, J. *et al.* (2015) *Visualising energy use for smart homes and informed users*, Energy Procedia, 78 579-584.

Stopps, H. and Touchie, M.F. (2021) Residential smart thermostat use: An exploration of thermostat programming, environmental attitudes, and the influence of smart controls on energy savings, Energy & Buildings, 238 1-16.

Strengers, Y. *et al*. (2020) *Pursuing pleasance: Interrogating energy-intensive visions for the smart home*, International Journal of Human-Computer Studies, 136 1-14.

Vasicek, D. *et al*. (2018) *IoT smart home concept*, 26th Telecommunications forum *(TELFOR),* Belgrade, Serbia.

Yang, B. *et al.* (2019) Smart metering and systems for low-energy households: challenges, issues and benefits, Advances in building energy research, 13 80-100.

A Value-Based Perspective on Distributed Ledger Technology Adoption

Efstathios Papanikolaou (Durham University) and Spyros Angelopoulos (Durham University)

Completed Research

Abstract

Research on blockchain technology has primarily focused on financial services, yet its broader applications and value drivers remain underexplored. Little attention has been given to identifying the necessary value sources within the Distributed Ledger Technology (DLT) ecosystem for effective composition by its entities. Our study identifies core elements that organizations should consider when adopting DLT to foster value co-creation with customers and partners. This study serves as a foundational step in establishing a value system to enhance the managerial activities within the DLT network. To fill this gap, we propose a conceptual framework that outlines essential Value Generation Objects (VGOs), also referred to as Capitals, which are critical for a well-functioning DLT ecosystem. By breaking down the DLT ecosystem into ecosystem members and the technology itself, we address the question, "What are the VGO elements essential for each DLT ecosystem entity?" Our framework provides a structured map of DLT ecosystem Capitals, offering scholars and managers an initial tool for understanding ecosystem composition. This model also establishes a basis for performance measurement, enabling entities to assess alignment between activities and value generation objectives.

Keywords: blockchain; distributed ledger; value generation; business ecosystem; capital; value system

1. Introduction

The mechanisms organizations use to create value within an ecosystem significantly influence the ecosystem's sustainability. Distributed Ledger Technology (DLT) fosters new modes of inter-firm collaboration by enhancing transparency, security, traceability, visibility, and agility across enterprise boundaries. These synergistic benefits enable organizations to seize specific collaboration opportunities effectively. (Bocek. et al., 2017; Nofer, et al., 2017). Given Distributed Ledger Technology's (DLT) unique features and the network effects from collaboration among loosely connected organizations, (Angelis, Ribeiro da Silva, 2019) we propose a business ecosystem perspective. Under the ecosystem context, DLT powered transactions create value that no single firm could create by itself (Gawer and Cusumano, 2014). Moreover, both DLT ecosystem sustainability and co-produced value depend on actor participation and ecosystem expansion. Within that environment, synergies are created by ecosystem member Capitals held and exchanged.

DLT itself sets the boundaries of the co-created value. We define "Value Generation Objects" (VGOs) as various forms of Capital held by key DLT ecosystem members and the DLT technology itself. Identifying the values held, exchanged, and generated by these two primary entities structures the VGO framework, which helps stakeholders assess their readiness for potential participation in a DLT ecosystem.

Current research has focused primarily on DLT’s implementation, as the technology is still emerging, with limited investigation into the types of capital ecosystem members need to hold or the value sources exchanged among them (Salviotti, Rossi, Abbatemarco, 2018). However, DLT’s benefits are expanding beyond financial services, highlighting the importance for future DLT ecosystem members to prepare for involvement (Papanikolaou, Angelis, Moustakis, 2021). By defining the main VGO elements, DLT ecosystem participants can better evaluate their potential roles and readiness for integration.

Our study identifies the key elements that organizations should consider when adopting DLT and joining the DLT business ecosystem to develop strategies for value co-creation with customers and ecosystem partners. Based on literature this is the first step towards the creation of a value system that will allow DLT interacting actors to support managerial activities in the network, such as value co-creation and performance measurement (Romero, Galeano, Molina, 2010). Value system is the identification, structure, and measurement of a set of values that an actor holds, exchanges and creates for specific purposes (Romero, Galeano, Molina, 2010). Value systems are based on the economical notion that “each product and/or service offered requires a set of activities carried out by a number of actors forming a value creation system, that use tangible and intangible resources for creating value for customers” (Parolini, 1999).

Researching DLT ecosystem value systems provides a comprehensive view of how different components interact and create value. Value system is an important managerial tool, since it helps decision makers to identify the main elements that generate value in a virtual breeding environment (VBE), to focus and adjust their DLT strategy and operation based on those elements. DLT ecosystems involve multiple stakeholders and components interacting in complex ways. Studying the value system mechanisms helps unravel these interactions and how they contribute to overall ecosystem value. We inform better design of DLT ecosystems, including governance structures, incentive systems, and consensus mechanisms that align with value creation goals. Moreover, providing a holistic understanding of ecosystem value system organizations can identify at the eearly stages of DLT adoption potential bottlenecks or inefficiencies, guiding efforts to improve the scalability and performance of DLT networks.

Moreno, Galeano and Molina (2010) conceptualize Value Generation Objects from different stakeholders, to describe a Virtual Breeding Environment (VBE) value system. Virtual Breeding Environments (VBEs) represent networks of disperse organizations, that will exploit specific collaboration opportunities through the creation of Virtual Organizations (VOs) supported by information technologies. Our study draws parallels between Virtual Breeding Environments (VBEs) and Distributed Ledger Technology (DLT) ecosystems, where diverse organizations collaborate in opportunity-driven ecosystems. In this context, DLT serves as the foundational IT infrastructure for ecosystem formation, similar to the role of Virtual Organizations (VOs) within VBEs. We extend blockchain literature by applying insights from value systems within organizational clusters, such as VBEs, to the DLT ecosystem.

Our research addresses a gap in DLT literature, which has largely focused on technical aspects, by exploring Value Generation Objects (VGOs) as the foundational layer in defining a value system framework for DLT ecosystems. We argue that VGOs are represented by both the ecosystem actors and the DLT technology itself, where actors contribute various forms of "Capitals" that DLT interlinks. Different DLT architectures—centralized, decentralized, private, public, and on/off-chain data storage (Xu et al., 2017)—influence the value generated for each actor within the ecosystem.

This study seeks to identify the VGO elements specific to each DLT ecosystem entity, answering the research question, "What are the VGO elements for each entity within the DLT business ecosystem?" Section 2 introduces the concept of value within the DLT ecosystem and distinguishes it from traditional value webs. Drawing on VBE research (Romero, Galeano, Molina, 2007), we argue that VGOs comprise the combined capitals of members and platforms, a notion that applies to the DLT ecosystem. Section 3 synthesizes findings from capitals literature to align with the DLT ecosystem, while Section 4 conceptualizes VGO dimensions based on ecosystem characteristics and details their elements in Section 5.

In conclusion, we assert that DLT ecosystem capitals cannot be adequately captured solely through economic and social lenses. Business value and DLT-specific development capitals play critical roles in co-creating value within the ecosystem, enabling participants to realize both profits and broader benefits.

2. Value in DLT business ecosystem and value webs

In this section, we explain the differences between the DLT ecosystem model and the value web approach, noting that the former better represents DLT's network structure. We also discuss the use of Virtual Breeding Environments (VBEs) and Value Generation Objects (VGOs) as a foundational framework for the DLT ecosystem. Recent reviews of blockchain literature show a focus on value drivers, creation, and propositions, with limited attention to application, value creation, and governance (Becker and Oxman, 2008; Riasanow et al., 2020; Lumineau, et al., 2021). Other studies examine challenges related to blockchain adoption in different industries (Zheng and Lu, 2022; Gad et al., 2020. Overall, research can be categorized into technological and economic studies, with the former focusing on advances, implementation models, and security challenges, and the latter exploring blockchain's value potential in financial services.

Studies commonly use the business model canvas to evaluate blockchain's influence on business and value creation (Sun et al., 2022; Alkhudary et al., 2020). This research emphasizes the types of capital exchanged within DLT ecosystems to foster co-created value. Moore (2006) stressed "space" in ecosystems as essential for generating opportunities. Companies can build value by developing products supporting Industry 4.0, like private DLT setups, or by creating digital assets in open DLT architectures. DLT systems drive ecosystem value by enabling secure data exchange across independent organizations, where interconnected networks replace traditional value chains to enhance collective value creation (Kasper-Fuejrer & Ashkanasy, 2003; Iansiti & Levien, 2004). Value Generation Objects (VGOs), including various capitals held by ecosystem members, are central to DLT value production (Romero, Galeano, Molina, 2010).

VBEs represent interconnected networks of diverse organizations that leverage specific collaboration opportunities facilitated by information and communication technologies (ICTs). The comparison to DLT business ecosystems is profound. DLT serves as the ICT that binds interacting ecosystem participants and encourages the creation of value in the DLT business ecosystem. Analogous to business ecosystems, Virtual Breeding

Environments (VBEs) are comprised of independent, geographically dispersed, and varied organizations in enduring partnerships. While VBEs prepare members for potential collaboration, business ecosystems necessitate collaboration for sustainable value creation. Therefore, we posit that VGOs in DLT ecosystems encompass the collective capital of both members and DLT. Member capital includes resources possessed and exchanged exclusively within the network, impacting ecosystem structure and dynamics, while DLT capital represents the technology's methodologies and attributes, delineating the kinds of value and sculpting the ecosystem.

Value webs, often described as market-based assets or networks, reflect value creation through external, networked resources (Herath, Senaratne, and Gunarathne, 2021). These webs feature interdependent competitors in a cooperative environment, forming alliances in networked markets (Gnanaweera and Kunori, 2018). Despite their similarities to business ecosystems, value webs differ fundamentally; in value webs, value is driven by an organization’s ability to leverage these external networks, rather than the intrinsic product value (Grover, Chiang, Liang, and Zhang, 2018). In networked markets, value arises from customer expectations and complementary products, while in DLT ecosystems, value results from removing bottlenecks and intermediaries across all interacting systems, emphasizing co-created value distinct from value drivers in value webs or networked markets (Kaartemo, Akaka, and Vargo, 2017).

3. Methodology

Research on "capital" across various fields is challenging to systematically review, so we conducted a semi-systematic review to identify recurring themes, theoretical perspectives, and conceptual components. This approach allows for recognizing patterns in studies within specific disciplines or methodologies (Asrar-ul-Haq & Anwar, 2016). We used the PRISMA semi-systematic methodology to search for literature in Web of Science and Scopus databases, focusing on management research. The search included terms like “capital(s)” or "value system," and we filtered results for English-language articles in relevant fields. Starting with 419 publications, we eliminated duplicates, leaving us with 286 unique sources. We then prioritized descriptive and umbrella reviews to identify capital categories with broad agreement. Screening titles and abstracts narrowed down the selection to 217 articles that aligned with our objectives. Using snowballing (Wohlin, 2014), we further reduced the list to 83 publications before ultimately including 36 in our final analysis. Each study was thoroughly reviewed to identify 20 types of capital organized into five categories, with their elements outlined. The PRISMA review process is illustrated in Figure 1.

[insert **Figure 1** about here]

Affinity diagram was selected, as in Table 1, to organize capitals in categories based on their underlying similarity (Widjaja, Takahashi, 2016). Affinity diagram was shared among research authors for review. That was intended to reduce subjectivity, since for certain types of reviews at least two independent reviewers need to be involved in the screening process (Pham et al., 2014).

[insert **Table 1** about here]

We specifically draw from Moreno, Galeano, and Molina's (2010) work on value generation objects (VGOs) across stakeholders to outline a value system for virtual breeding environments (VBEs) supported by information technologies. We suggest that the DLT business ecosystem has similarities with the VBE model, specifically in how IT brings together independent entities for value creation. Therefore, we are using Moreno, Galeano, and Molina's VGO categorization structure to create our own affinity diagram of VGOs in the DLT ecosystem.

4. Capitals' literature review

The study is focused on identifying a defined set of VGOs or assets that contribute to value creation within the DLT business ecosystem. Capital is viewed as a source of value, a valuable resource of a specific type, which ecosystem entities possess and trade while engaging through DLT (Capital Oxford English Dictionary, 2019). These assets are transformed into tangible and intangible outputs, which are beneficial for other members of the ecosystem. In a successful and enduring ecosystem, all participants contribute and gain mutually, with value co-creation promoting both individual and collective success (Weill & Woerner, 2015). In addition to the VGOs held by participants, DLT itself is acknowledged as a distinct ecosystem entity with inherent value that plays a role in ecosystem sustainability, necessitating a separate examination of its VGOs.

Based on existing research, scholars have expanded the concept of VGOs beyond financial capital to encompass various forms of social and community value. For example, the essential assets driving development in rural communities include built, natural, financial, human, social, political, and cultural assets (Fey et al., 2008). Similarly, the Forum for the Future (2009) introduced a five-capital model—natural, human, social, manufactured, and financial—which has influenced sustainable business research. Ernst & Young's framework also includes intellectual capital as a key element in a company's value creation and risk management within a broader societal context (Ernst & Young, 2016).

Social capital theory views networks of relationships as valuable resources that give members access to information, opportunities, and reputational advantages (Nahapiet & Ghoshal, 1998). In a DLT ecosystem, the strength of these relationships and the role of each participant directly impacts their ability to develop and benefit from social capital. Therefore, social capital in this context is closely linked to the VGOs of DLT members, encompassing both networks and resources mobilized through resource exchange and combination. Applying social capital theory to DLT ecosystems suggests that social capital—deeply rooted in relationships and resource sharing—supports the exchange and combination necessary for DLT sustainability and technical advancement. Resource utilization is crucial for co-creation in business ecosystems, with participants adding value by pooling their resources (Dahl et al., 2016; Valkokari, 2015). In the DLT ecosystem, VGOs consist of tangible resources like digital, financial, data, and fixed assets, along with intangible resources such as intellectual property, brand reputation, and innovation, all categorized under assets, knowledge, and capabilities (Romero et al., 2010). Moreover, process and structural capital, which include capabilities that enhance workflow, are essential for optimizing business processes. Studies on ecosystems and communities have shown that resources not only connect participants but also encourage synergies that drive co-creation, ecosystem growth, and resilience. These attributes support value creation and the sustainable development of ecosystems.

5. Dimensions of Value Generation Objects for DLT business ecosystem

To identify VGOs in a DLT business ecosystem, it is necessary to dive deeper in the concept of "value" and its different aspects. We suggest examining DLT and ecosystem member capitals separately, focusing on tangible and intangible resources. Traditional resource categorization is used for DLT ecosystem member capitals, with additional dimensions related to DLT ecosystems. Our analysis of technology VGOs includes capital dimensions relevant to technology development and the special needs of DLT adoption.

5.1 The traditional approach of economic and social capitals

Drawing on economic and sociological perspectives (Bock & George, 2018), literature examines value within business ecosystems, focusing specifically on social and economic dimensions. Although multiple value types exist, such as familial or moral, this analysis centers on those that drive interactions in ecosystems. Economic value, typically related to a product's price or utility—such as reliability, efficiency, and flexibility—is outside our scope here, as these characteristics pertain to individual products rather than shared ecosystem resources.

Economically, our emphasis is on capital and business value as core production resources. From a sociological view, social capital involves shared beliefs and norms (Macedo et al., 2006), with social value arising from interactions within economic communities. This concept is especially relevant to DLT ecosystems, where interdependent companies create collaborative networks focused on mutual value creation.

5.2 The need for business value conceptualization

The business value approach in DLT ecosystems emphasizes customer perspectives, assessing product benefits, costs, and functionality compared to alternatives, alongside the provider's productivity and expertise (Anderson, Narus, Rossum, 2006). Functional value connects to DLT's role in enhancing product features and the "capitals" that participants bring to the ecosystem. Economic value, foundational in business theory, frames DLT ecosystems as economic communities where consensus mechanisms align business objectives with broader social and economic principles (Moore, 2006; Abu, 2022). For ecosystem sustainability, economic contributions from each member are essential, as interdependent actors rely on each other's efficiency (Iansiti, Levien, 2004). Thus, economic value remains crucial to DLT ecosystem VGOs.

Economic capital includes both the financial value of goods or services and the funds used by organizations to achieve objectives. Business value expands this by including market share, intellectual property, productivity, and efficiency, reflecting the impact of both internal and external relationships on ecosystem value creation. In a DLT ecosystem, total value is generated through partner interactions and evolves based on each participant's role and status within the network.

6. Value Generation Object elements for DLT business ecosystem

To identify Value Generation Objects (VGOs) in Distributed Ledger Technology (DLT) ecosystems, we consider both the combined resources of ecosystem participants and DLT-specific capitals. Table 2 categorizes these VGOs by dimension, distinguishing between tangible and intangible capitals based on both existing literature and the unique attributes of DLT ecosystems.

6.1 DLT business ecosystem member capitals

VGOs for ecosystem members represent the economic, social, and business assets each participant brings, while VGOs specific to DLT reflect the technology's unique contributions to the ecosystem.

6.1.1 DLT business ecosystem member economic capitals

Economic capital consists of the financial assets and resources needed to achieve organizational goals and sustain ecosystem viability (Perez, 2002). This capital supports participants individually and is essential for the development of DLT technology. Since each DLT service or product relies on a network of actors sharing tangible and intangible resources (Zhou et al., 2022), stable financial support is crucial. Insufficient financial resources can disrupt operations, increase transaction validation costs, and slow down DLT processes. Some ecosystem members, such as transaction validators, have roles requiring significant initial investments to facilitate transactions, which ultimately strengthens the ecosystem. Although validation activities are crucial to DLT operations, the required financial resources are classified as member capitals rather than DLT capitals. Unlike conventional platforms, financial capital in DLT ecosystems also extends to the development of decentralized applications (dApps) and validator funding.

Assets, including economic-value resources like fixed assets or digital data, are key VGOs that foster co-creation by facilitating resource exchange and integration (Deo, 2021). Tangible assets, such as patents and copyrights, complement intangible assets in generating ecosystem value. In DLT ecosystems, data is especially significant: beyond its intrinsic value to the owning entity, shared data can drive new knowledge and increase value creation. However, this reliance on data-sharing raises privacy concerns, as some participants may feel that shared data grants undue control to certain entities. Addressing these concerns is crucial to support wider DLT adoption and emphasizes data's essential role in value generation.

6.1.2 DLT business ecosystem member social capitals

In a Distributed Ledger Technology (DLT) ecosystem, *relational capital* refers to the intangible value derived from stakeholder interactions, which can range from cooperation to competition (Araujo & Easton, 1999). Relationships evolve based on participant roles and digital advancements, which both promote competition and strengthen value chains (Khan et al., 2020). DLT enhances relational capital by fostering transparency, trust, and efficiency, facilitating collaboration—even among previously unconnected actors (Lenkenhoff et al., 2018). Relational capital is vital for generating both business and social benefits, as it encourages partnerships, boosts trust, and lowers entry barriers, promoting a more collaborative network.

An actor's role and influence in the DLT ecosystem directly affect both collaboration potential and co-evolution (Adner & Kapoor, 2010). Organizations leverage their roles to shape interactions, with influential participants creating connections across network gaps, enhancing overall value. However, dominant players that engage in superficial collaborations may inadvertently demotivate smaller participants, limiting the ecosystem's scalability and longevity (Granovetter, 1985; Burt, 1997). For the DLT ecosystem to grow sustainably, it is essential to foster balanced, meaningful collaborations and recognize the contributions of niche participants, who bring innovation and diversity.

Although identity capital—an organization's mission, values, and public perception—is sometimes classified with social capital (Romero et al., 2007), it is not considered a separate Value Generation Object (VGO) in this DLT framework. While identity capital can shape organizational culture and external relationships (Dagnino, 2007; Kabadayi, 2020), the DLT context prioritizes an actor's role, influence, and collaborative approach over external identity, as these factors more directly influence value creation and ecosystem transformation.

6.1.3 DLT business ecosystem member business value

In Distributed Ledger Technology (DLT) ecosystems, business value goes beyond financial metrics, encompassing customer perception, efficiency, and organizational expertise (Fedorovich & Fedorovich, 2019). Structural capital, a part of intellectual capital, includes essential systems, data, and processes that preserve knowledge and meet objectives. Customer value in this context affects loyalty, trust, and buying behaviors (Kordupleski & Simpson, 2003). Particularly in B2B settings, DLT fosters trust by limiting opportunistic actions, though concerns about data transparency may arise among dominant entities.

While related to customer value, functional value in the DLT ecosystem requires separate focus. This aspect pertains to product or service features that satisfy user needs and enhance competitive positioning (Sore et al., 2022). DLT adoption, which can redefine a firm's functional boundaries, has the potential to reshape collaborative functionality within the ecosystem. Overall, the DLT ecosystem fosters business value not only through monetary gains but also by promoting trust, operational functionality, and cooperative productivity.

Knowledge capital stands as a core Value-Generating Outcome (VGO) in the DLT ecosystem, reflecting a continuous pursuit of improvement through shared learning and innovation (Hamel, 2000). Within a DLT network, members gain greater access to data, transparency, and trust, all of which are foundational for value creation. The resulting network visibility and trust lower coordination costs, enabling smoother collaboration and reinforcing knowledge exchange. DLT infrastructure also supports efficient management of information flows, enabling members to uncover new insights and driving innovation in a data-centric economy (Treiblmaier & Beck, 2019). Through dynamic capabilities theory, external conditions like collaborative DLT platforms can directly shape an organization’s learning and knowledge development (Teece et al., 1997).

Within this ecosystem, knowledge capital—encompassing human and innovation capital—is vital, propelling adaptability and competitiveness in a digital economy through elements like data, intellectual property, and patents (Bock & George, 2018). Human capital, a key intangible asset, includes the collective expertise, skills, and problem-solving abilities within an organization, influencing both productivity and operational efficiency. Aspects of human capital, such as systems thinking, impact organizational productivity, while social skills contribute to social VGOs, shaping relationships and fostering success within the ecosystem.

Capability resources denote an organization's potential to enhance operational effectiveness and productivity. Advanced capabilities can improve production throughput, service quality, and development efficiency, playing a significant role in co-creating ecosystem value (Ernst & Young, 2016). The DLT ecosystem benefits from enhanced productivity, process efficiency, and data accessibility, all supported by skills like problem-solving and decision-making.

DLT-based data sharing supports knowledge creation, essential for ecosystem value. By enhancing traceability and transparency, DLT fosters process efficiencies that ultimately strengthen the ecosystem. Structural capital—encompassing tools, processes, and routines—helps retain knowledge and meet organizational goals (Aramburu & Sáenz, 2011). Innovation, while sometimes included in structural capital, is categorized here as knowledge capital due to its role in knowledge generation. DLT-enabled openness and data sharing align with innovation practices like resource exchange and collaboration (Quandt & Castilho, 2017), where “co-opetition” promotes new R&D-driven innovations (Seo et al., 2017). Market position and share impact organizational influence within the ecosystem’s social capital framework.

[insert **Table 2** about here]

6.2 DLT Value Generation Objects

In addition to the resources each participant brings, DLT Value Generating Outputs (VGOs) represent a foundational part of the DLT ecosystem's value architecture. Table 1 offers an in-depth examination of these VGOs across the ecosystem. DLT VGOs capture the resources crucial for technological progress and align with DLT's inherent traits and requirements, which, when fulfilled, substantially aid value generation and support the ecosystem's objectives.

Within this structure, we categorize financial, human, and infrastructure resources related to DLT development as VGOs. The social dimension of DLT VGOs involves ecosystem growth potential, which is driven by incentives for member engagement, the effects of network synergies on value co-creation, and the roles and policies established by primary ecosystem players. These social capitals are essential for DLT's collaborative principles, fostering sustainability and enabling value creation across the ecosystem. As the technology advances, its applicability across various industries increases, helping to overcome initial adoption barriers (Wanda, Doskey, Moreland, 2017). Particularly in its early stages, the maturity of the technology can redefine ecosystem expansion and potential.

6.2.1 DLT development capitals

Key resources in DLT development include infrastructure, financial, and human capital (Romero, Galeano, Molina, 2010). Financial capital involves the funding essential for DLT innovation, closely tied to the human resources required to advance the technology. Implementation strategies may vary from building new systems, using components from tech providers, or modifying existing DLT solutions. Regardless of approach, customization is often necessary to align with ecosystem needs, necessitating considerable financial and human resources, especially with emerging technologies like DLT.

Infrastructure capital consists of the physical and technological assets supporting DLT, such as mobile apps, decentralized applications (dApps), serverless platforms, smart contracts, and consensus protocols. Given DLT's relatively recent introduction, it is rapidly evolving to address challenges related to data sharing and visibility. Requirements for off-chain storage and network integration drive continuous innovation, with privacy and interoperability being key factors for scalability and adoption, establishing infrastructure as vital to value creation and expansion in the ecosystem.

6.2.2 DLT social capitals

DLT's scalability and maturity extend beyond technical boundaries to impact the social capital within the ecosystem, motivating members' participation and shaping governance policies, including those set by major players (Philbeck & Davis, 2018). Maturity, often linked to the time since a technology's market introduction (Adner & Kapoor, 2010), helps broaden adoption and overcome initial hurdles. Within DLT, maturity is central to sustainability, with network effects and strategies by key actors enhancing DLT's unique framework.

As DLT is still developing, pilot projects across sectors are testing its strengths and constraints. In fast-evolving environments where cooperation, competition, and co-opetition intersect, these projects assess DLT's potential for broader applications. Moving from permissioned to open systems will require greater technological maturity to increase shared benefits and encourage engagement. With more participants, network effects boost value, attracting further members and supporting ecosystem resilience. A mature DLT infrastructure would enhance interoperability, ensuring smooth interactions and seamless value transfers across ecosystems (Haoyan et al., 2017).

Security measures, such as a consensus mechanism's resistance to potential threats, improve with greater ecosystem involvement. The DLT architecture—including the type of consensus, data permissions, and governance—strongly influences ecosystem stability. Major players guiding DLT adoption must ensure fair value distribution to maintain engagement from smaller participants. The policies set by these dominant players, especially around data access, are instrumental in shaping the ecosystem's value creation dynamics. This equilibrium between governance and inclusivity ultimately determines the ecosystem's resilience and the collective success of DLT-based collaborations.

7. Conclusion and discussion

Organizations deploying DLT should determine anticipated benefits within an ecosystem framework, as DLT emphasizes a model of shared value creation among participants. Our analysis of the DLT business ecosystem reveals key value sources, referred to as Value Generation Objects (VGOs) or “capitals,” that underpin the ecosystem’s sustainability and growth. We classify ecosystem entities into two categories: ecosystem members, who engage in value exchange, and the technology itself, which enables these interactions. Effective participation requires all members to actively contribute to and derive benefits from this ecosystem, which DLT-facilitated interactions support.

This study aims to outline a structured framework of VGOs that foster value creation within the DLT ecosystem. VGOs represent valuable resources or “capitals” that members hold, exchange, and use to sustain the ecosystem. Value systems incorporate these VGOs across entities (Parolini, 1999). Defining this system in the DLT context involves identifying VGOs and understanding their role in value generation. After recognizing VGOs, the next step involves creating a performance measurement system that assesses VGO effectiveness in meeting ecosystem objectives. This system enables participants to track progress and refine activities to enhance value creation.

Our framework proposes three primary VGOs for ecosystem members in DLT settings: economic capital, social capital, and business value. Economic capital encompasses financial assets vital for activity support and positive economic outcomes, underscoring data ownership and management’s strategic significance. Social capital pertains to social behavior and organizational relationships within the ecosystem, enhancing each member’s ability to generate and exchange value. Business value covers intangible elements beyond economic gain, such as customer perspectives, product or service functionality, inter-firm processes, and shared knowledge.

The framework’s second component addresses VGOs linked to technology development, including both tangible and intangible assets and the social factors influencing DLT adoption. Given DLT’s early development stage, its scalability and adoption are closely tied to overcoming technical hurdles.

Identifying VGOs offers stakeholders a tool for understanding and effectively integrating essential capitals, allowing decision-makers to evaluate their engagement within the DLT ecosystem. Value systems drive performance improvements across DLT ecosystems by aligning strategies with defined VGOs and setting performance metrics to assess and manage strategic impact. Due to limited micro-level quantitative data, this study could not determine whether DLT-specific VGOs have met raised expectations or influenced ecosystem sustainability and growth.

Our framework centers on defining VGOs and their role in creating value in the DLT business ecosystem, helping clarify which capitals attract partners and encourage collaborative innovation. Managers can adopt a “positive-sum game” approach, moving beyond firm-focused strategies to foster shared value. Future studies could create performance metrics to quantify co-created value within DLT ecosystems and verify value systems through case studies across DLT fields. Such studies would allow stakeholders to measure progress, compare results with established goals, and evaluate if DLT ecosystem participation delivers expected value. This research provides insights into the DLT business ecosystem structure from a capital perspective, integrating various VGOs.

References

Abu. H (2022). Expansion of Techno-Capital: New Space of Contestation, October 2022 In book: Reading Sociology Decolonizing Canada, Fourth Edition (pp.327-332), Oxford University Press

Adner R, Kapoor R. (2010). Value creation in innovation ecosystems: how the structure of technological interdependence affects firm performance in new technology generations. Strategy Management Journal. 31:306–333. https://doi.org/10.1002/smj.821

Alkhudary, R., Brusset, X. and Fenies, P. (2020), Blockchain in general management and economics: a systematic literature review, European Business Review, Vol. 32 No. 4, pp. 765-783. https://doi.org/10.1108/EBR-11-2019-0297

Anderson, J., Narus, J., Rossum, W. (2006). Customer Value Propositions in Business Markets. Harvard business review. 84. 90-9, 149.

Angelis J. and Ribeiro da Silva E. (2019). Blockchain adoption: A value driver perspective. Journal of Business Horizons. Vol 62, Issue 3,307-314, https://doi.org/10.1016/j.bushor.2018.12.001

Aramburu, N., Sáenz, J. (2011). Structural capital, innovation capability, and size effect: An empirical study. Journal of Management & Organization, 17(3), 307-325. https://doi.org/doi:10.5172/jmo.2011.17.3.307

Araujo, L., Easton, G. (1999). A Relational Resource Perspective on Social Capital. In: Leenders, R.T.A.J., Gabbay, S.M. (eds) Corporate Social Capital and Liability. Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-5027-3_4

Asrar-ul-Haq M. and Anwar. S. (2016) A systematic review of knowledge management and knowledge sharing: Trends, issues, and challenges, Cogent Business & Management, 3:1, DOI: 10.1080/23311975.2015.1127744

Audretsch, D., Link, A. (2018). Innovation capital. The Journal of Technology Transfer. 43 https://doi.org/10.1007/s10961-018-9700-6

Bauer, I. et al (2019). Exploring Blockchain Value Creation: The Case of the Car Ecosystem, Proceedings of the 52nd Hawaii International Conference on System Sciences.

Becker L. A. and Oxman A.D. (2008). Cochrane handbook for systematic reviews of interventions. Higgins J. P. T., Green S., editors. Hoboken, nj: John Wiley & Sons, Ltd; Overviews of reviews; pp. 607–631

Bocek, T., et al. (2017). Blockchains everywhere - a use-case of blockchains in the pharma supply-chain, 772–777, IFIP/IEEE Symposium on Integrated Network and Service Management (IM) https://doi.org/10.23919/INM.2017.7987376

Bock, A., Gerard, G. (2018). The business model book: Design, build and adapt business ideas that thrive. United Kingdom: Pearson Education Limited.pg6

Bφhme, R., Christin, N., Edelman, B., & Moore, T. (2015). Bitcoin: Economics technology, and governance. Journal of Economic Perspectives, 29(2), 213–238. DOI:10.1257/jep.29.2.213

Broch, C., Lurati, F., Zamparini, A., Mariconda, S. (2018). The Role of Social Capital for Organizational Identification: Implications for Strategic Communication. International Journal of Strategic Communication, 12:1, 46-66 https://doi.org/10.1080/1553118X.2017.1392310

Burt, R. S. (1997). The contingent value of social capital. Admin. Sci. Quart. 42 339-365. Cacciatori, E., M. Jacobides. https://doi.org/10.1016/B978-0-7506-7222-1.50014-3

Byukusenge E. and Munene. J. (2017) Knowledge management and business performance: Does innovation matter?, Cogent Business & Management, 4:1, DOI: 10.1080/23311975.2017.1368434

Câmara, S., Buarque, B., Pinto, G., Ribeiro, T and Soares, J. (2022). Innovation policy and public funding to stimulate innovation in knowledge intensive companies: the influence of human and social capital. Journal of Science and Technology Policy Management. 10.1108/JSTPM-09-2021-0135.

Capital Oxford English Dictionary. (2019) Oxford University Press

Dabbagh, M. Sookhak, M. and Safa, N.S. (2019). The Evolution of Blockchain: A Bibliometric Study," in IEEE Access, vol. 7, pp. 19212-19221, 2019, doi: 10.1109/ACCESS.2019.2895646.

Dagnino, G. (2007). Preface: Coopetition Strategy—Toward a New Kind of Inter-Firm Dynamics?, International Studies of Management & Organization, 37:2, 3-10, https://doi.org/10.2753/IMO0020-8825370200

Chiambaretto, P., Dumez, H. (2016). Toward a Typology of Coopetition: A Multilevel Approach, International Studies of Management & Organization, 46:2-3, 110-129, https://doi.org/10.1080/00208825.2015.1093797

Dahl, J., Kock, S., Lundgren-Henriksson E. (2016). Conceptualizing Coopetition Strategy as Practice: A Multilevel Interpretative Framework, International Studies of Management & Organization, 46:2-3, 94-109, https://doi.org/10.1080/00208825.2015.1093794

Deo, P. (2021). Fixed Asset Management-Revisited. Journal of Accounting and Finance, 21(1), 10-22.

Ernst & Young. (2016). Integrated reporting, Linking strategy, purpose and value Available at https://integratedreporting.org/resource/ey-integrated-reporting-linking-strategy-purpose-and-value/

Fedorovich, V.O., Fedorovich, T.V. (2019). Corporate business value: Asymmetric information in calculating economic value added. Financial Analytics: Science and Experience. 12. 183-203 https://doi.org/10.24891/fa.12.2.183

Fey, S., Bregendahl, C., Cornelia, F. (2008). The Measurement of Community Capitals through Research. Online Journal of Rural Research & Policy. https://doi.org/10.4148/ojrrp.v1i1.29

Firdaus, A., et al. (2019). The rise of "blockchain": bibliometric analysis of blockchain study. Scientometrics 120, 1289–1331 (2019). https://doi.org/10.1007/s11192-019-03170-4

Forum For The Future (2009). The Five Capitals Model a framework for sustainability. Available at https://www.forumforthefuture.org/the-five-capitals

Gad, G.A., et al. (2022), Emerging Trends in Blockchain Technology and Applications: A Review and Outlook, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 9, Pages 6719-6742, https://doi.org/10.1016/j.jksuci.2022.03.007.

Gawer A. and Cusumano M.A. (2014). Industry Platforms and Ecosystem Innovation. Journal of Product Innovation Management.31(3),417-433. doi:10.1111/jpim.12105

Gorkhali, A., Ling Li, L. and Shrestha, A. (2020). Blockchain: a literature review, Journal of Management Analytics, 7:3, 321-343, doi: 10.1080/23270012.2020.1801529

Gnanaweera K.A.K. and Kunori N. (2018) Corporate sustainability reporting: Linkage of corporate disclosure information and performance indicators, Cogent Business & Management, 5:1, DOI: 10.1080/23311975.2018.1423872

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. American Journal of Sociology 91 481-510. https://doi.org/10.2307/j.ctv1f886rp.12

Hamel, G. (2000). Leading the Revolution. Boston, MA: Harvard Business School Press.

Haoyan W., et.al (2017). A distribute ledger for supply chain physical distribution validity. Journal of Information, 8,137. https://doi.org/10.3390/info8040137

Heredia Pérez, J., Yang, X., Bai, O., Flores, A.,Heredia, W. (2019). How Does Competition By Informal Firms Affect The Innovation In Formal Firms?, International Studies of Management & Organization, 49:2, 173-190, https://doi.org/10.1080/00208825.2019.1608402

Herath, R., Senaratne, S. and Gunarathne, N. (2021). Integrated thinking, orchestration of the six capitals and value creation. Meditari Accountancy Research. ahead-of-print. 10.1108/MEDAR-01-2020-0676.

Hoeborn, G., et al (2022) Understanding Business Ecosystems Using a Morphology of Value Systems, Research-Technology Management, 65:5, 44-53, doi:10.1080/08956308.2022.2095841

Iansiti M., Levien. R. (2004). The Keystone Advantage: What New Dynamics of Business Ecosystems Mean for Strategy, Innovation, and Sustainability. Harvard Business School Press, Boston, MA. https://doi.org/10.5465/amp.2006.20591015

Kaartemo, V., Akaka, M and Vargo, S. (2017). A Service-Ecosystem Perspective on Value Creation: Implications for International Business, in book: Value Creation in International Business (pp.131-149) Edition: Volume 2: An SME Perspective, Publisher: Springer International Publishing, Editors: Svetla Marinova, Jorma Larimo, Niina Nummela. doi:10.1007/978-3-319-39369-8_6

Kabuye F., Joachim Kato J., Akugizibwe I. and Bugambiro N. (2019) Internal control systems, working capital management and financial performance of supermarkets, Cogent Business & Management, 6:1, DOI: 10.1080/23311975.2019.1573524

Kasper-Fuejrer, E., Ashkanasy, N. (2003). The Interorganizational Virtual Organization : Defining a Weberian Ideal, International Studies of Management & Organization, 33:4, 34-64, https://doi.org/10.1080/00208825.2003.11043688

Khan, Z., Kyu, Y., Rao-Nicholson, R. (2020). The role of dynamic capabilities in global strategy of emerging economies' multinationals, International Studies of Management & Organization, 50:1, 1-4, https://doi.org/10.1080/00208825.2019.1703375

Konstantinidis, I., et al (2018). Blockchain for Business Applications: A Systematic Literature Review. In: Abramowicz, W., Paschke, A. (eds) Business Information

Systems. BIS 2018. Lecture Notes in Business Information Processing, vol 320. Springer, Cham. https://doi.org/10.1007/978-3-319-93931-5_28

Kordupleski, R. (2003). Mastering Customer Value Management: The Art and Science of Creating Competitive Advantage, Pinnaflex Educational Resources inc

LaFayette, B., Curtis, W., Bedford, D. and Iyer, S. (2019). *Structural Capital – Definitions and Growth. Knowledge Economies and Knowledge Work* (Working Methods for Knowledge Management), Emerald Publishing Limited, Bingley, pp. 115-127. https://doi.org/10.1108/978-1-78973-775-220191007

Lenkenhoff, K., Wilkens, U., Zheng, M., Suesse, T., Kuhlenkötter, B. and Ming, X. (2018). *Key challenges of digital business ecosystem development and how to cope with them*. Procedia CIRP. 73. 167-172 https://doi.org/1010.1016/j.procir.2018.04.082

Li, X and Wu, W (2022) *Recent Advances of Blockchain and Its Applications,* Journal of Social Computing, vol. 3, no. 4, pp. 363-394, doi:10.23919/JSC.2022.0016.

Lumineau, F., Wang, W. and Schilke, O. (2021*). Blockchain governance—A new way of organizing collaborations?* Organization Science, 32(2), 500-52

Marikyan,D et. al. (2022). *Blockchain: A business model innovation analysis*, Digital Business, Volume 2, Issue 2,100033, ISSN 2666-9544, https://doi.org/10.1016/j.digbus.2022.100033.

Macedo, P., Sapateiro, C. and Filipe, J. (2006). *Distinct Approaches to Value System in Collaborative Networks Environments*, in L.M. et al (Ed.). Network-Centric Collaboration and Supporting Frameworks. International Federation for Information Processing (IFIP), Volume 224, pp. 111-120, Boston: Springer Publisher https://doi.org/10.1007/978-0-387-38269-2_12

Moore, J. (2006). *Business ecosystems and the view from the firm*, The antitrust bulletin: Vol. 51, No. I/Spring 2006 p. 53 https://doi.org/10.1177/0003603X0605100103

Morkunas, V.J., Paschen, J. and Boon, E. (2019). *How blockchain technologies impact your business model,* Business Horizons, Volume 62, Issue 3, Pages 295-306, https://doi.org/10.1016/j.bushor.2019.01.009.

Nofer, M. et al. (2017). *Blockchain*. Business & Information Systems Engineering, 59(3), 183–187, https://doi.org/10.1007/s12599-017-0467-3

Page, M.J., et al. (2021), *The PRISMA 2020 statement: an updated guideline for reporting systematic reviews*. Systematic Reviews 10, 89. https://doi.org/10.1186/s13643-021-01626-4

Paananen, A. and Seppänen, M. (2013). *Reviewing customer value literature: Comparing and contrasting customer values perspectives*. Intangible Capital, 9(3), 708-729.

Papanikolaou E, Angelis J and Moustakis V. (2021). *Implicit business model effects of DLT adoption*, Procedia CIRP, Volume 103, Pages 298-304, https://doi.org/10.1016/j.procir.2021.10.048.

Parolini, C. (1999). *The Value Net Tool for Competitive Strategy*, in John Wiley & Sons Ltd

Perez, C. (2002). *Technological revolutions and financial capital*, Emerald publishing, p90-98

Philbeck, T. and Davis, N. (2018). *The fourth industrial revolution*. Journal of International Affairs, 72(1), 17-22.

Porras-Paez, A. and Schmutzler, J. (2019). *Orchestrating an entrepreneurial ecosystem in an emerging country: The lead actor's role from a social capital* *perspective*. Local Economy, 34(8), 767-786. https://doi.org/10.1177/0269094219896269

Quandt, C.and Castilho, M. (2017). *Relationship between collaboration and innovativeness: A case study in an innovative organisation*. International Journal of Innovation and Learning. 21. 257. https://doi.org/10.1504/IJIL.2017.083400

Pham, Mai et al. (2014). *A scoping review of scoping reviews: Advancing the approach and enhancing the consistency*. Research Synthesis Methods. 5

Riasanow, T., Burckhardt, F., Soto Setzke, D., Böhm, M. and Krcmar, H. (2020). *The Generic Blockchain Ecosystem and Its Strategic Implications*. Business & Information Systems Engineering, 62(3), 273-287.

Ragnedda, M. (2018) *Conceptualizing Digital Capital. Telematics and Informatics*, 35 (8). pp. 2366-2375. https://doi.org/10.1016/j.tele.2018.10.006

Risius, M. and Spohrer, K. (2017). *A Blockchain Research Framework*. Business & Information Systems Engineering, pp 385-409, Vol 59, https://doi.org/10.1007/s12599-017-0506-0

Romero, D., Galeano, N. and Molina, A. (2010). *Virtual Organization Breeding Environments Value Systems and its Elements*. Journal of Intelligent Manufacturing. 21. 267-286 https://doi.org/10.1007/s10845-008-0179-0

Romero, D., Galeano, N.and Molina, A. (2007). *A Conceptual Model for Virtual Breeding Environments Value Systems*. International Federation for Information Processing Digital Library; Establishing The Foundation Of Collaborative Networks. 243. 43-52 https://doi.org/10.1007/978-0-387-73798-0_5

Salviotti, G., Rossi, L. and Abbatemarco, N. (2018). *A structured framework to assess the business application landscape of blockchain technologies*. Proceedings of the

51st Hawaii International Conference in System Sciences. 10.24251/HICSS.2018.440.

Seo, H., Chung, Y. and Yoon, H. (2017). *R&D cooperation and unintended innovation performance: Role of appropriability regimes and sectoral characteristics*. Technovation 66– 67, 28–42. https://doi.org/10.1016/j.technovation.2017.03.002

Sore, S., Saunila, M. and Helkkula, A. (2022). *Business-to-Business Value Co-creation: Suppliers' Perspective of Essential Information Systems Capabilities*. Journal of Creating Value. 239496432211218. 10.1177/23949643221121857.

Storbacka, K. Frow, P. Nenonen, S. and Payne, A. (2012). *Designing business models for value co-creation*, Special Issue – Toward a Better Understanding of the Role of Value in Markets and Marketing, pp. 51-78.

Sun,Y et al. (2022). *Blockchain as a cutting-edge technology impacting business: A systematic literature review perspective*, Telecommunications Policy, Volume 46, Issue 10, 102443, ISSN 0308-5961, https://doi.org/10.1016/j.telpol.2022.102443.

Teece, D. J., Pisano, G. and Shuen, A. (1997). *Dynamic capabilities and strategic management*. Strategic Management Journal, 18, 509–33. https://doi.org/10.1142/9789812834478_0002

Theodoraki, C., Messeghem, K. and Rice, M.P. (2018). *A social capital approach to the development of sustainable entrepreneurial ecosystems: an explorative study*. Small Business Economics 51, 153–170. https://doi.org/10.1007/s11187-017-9924-0

Thompson, J. D. (1967). *Organizations in Action*. McGraw Hill, New York.

Treiblmaier H. and Beck R. (2019). *Business transformation through blockchain* Vol I. Switzerland: Palgrave Macmillan, 152-156

Trivedi. S., Aggarwal, R., and Singh, G. (2023). *An Umbrella Review of the Literature on Blockchain and Distributed Ledger Technology and Their Roles in Future Banking*. In Perspectives on Blockchain Technology and Responsible Investing, 29-57. Hershey, PA: *IGI Globa*l, 2023. https://doi.org/10.4018/978-1-6684-8361-9.ch002

Valkokari, K. (2015*). Business, Innovation, and Knowledge Ecosystems: How They Differ and How to Survive and Thrive within Them*. Technology Innovation Management Review. 5. 17-24 https://doi.org/10.22215/timreview/919

Varun Grover V, Chiang, R.H.L., Liang, T and Zhang D. (2018) *Creating Strategic Business Value from Big Data Analytics: A Research Framework*, Journal of Management Information Systems, 35:2, 388-423

Wanda, P., Doskey, S. and Moreland, J. (2017). *Technology Maturity Assessments and Confidence Intervals: Technology maturity assessments*. Systems Engineering. 20. 10. https://doi.org/10.1002/sys.21389

Weill, P. and Woerner, S. (2015). *Thriving in an Increasingly Digital Ecosystem*. MIT Sloan Management Review. 56. 27-34.

Widjaja, W. and Tak*ahashi, M. (2016).* Distributed interface for group affinity-diagram brainstorming. Concurrent Engineering, 24(4), 344-358.

Wu, X. and Zhang, W. (2020). *Business Model Innovations in China: From a Value Network Perspective*., Conference "US-China business cooperation in the 21st century opportunities and challenges for entrepreneurs", Indiana University, USA

Xu, M., Chen, X. and Kou, G. (2019) *A systematic review of blockchain*. Financial Innovation 5, 27 (2019). https://doi.org/10.1186/s40854-019-0147-z

Xu, X., Weber, I., Staples, M., Zhu, L., Bosch, J., Bass, L., Pautasso, C. and Rimba, P. (2017). *A Taxonomy of Blockchain-Based Systems for Architecture Design*.

Proceedings of the 2017 IEEE International Conference on Software Architecture (ICSA), Gothenburg, Sweden, 3–7 April 2017; pp. 243–252 https://doi.org/10.1109/ICSA.2017.33

Yli-Huumo, J et al. (2018) *Where Is Current Research on Blockchain Technology? - A Systematic Review*, PLOS ONE, https://doi.org/10.1371/journal.pone.0163477

Yujie Zheng, Y. and Boh, W. F. (2021). *Value drivers of blockchain technology: A case study of blockchain-enabled online community*, Telematics and Informatics, Volume 58, 2021, 101563, ISSN 0736-5853, https://doi.org/10.1016/j.tele.2021.101563.

Zhao, J. L., Fan, S. and Yan, J. (2016). *Overview of business innovations and research opportunities in blockchain,* Financial Innovation, vol. 2, no. 1,

Zheng X.R and Lu, Y. (2022) *Blockchain technology – recent research and future trend*, Enterprise Information Systems, 16:12, https://doi.org/10.1080/17517575.2021.1939895

Zhou, Q., Zhang, Y., Yang, W., Ren, L. and Chen, P. (2022). *Value co-creation in the multinational technology standard alliance: a case study from emerging economies*, Industrial Management & Data Systems, Vol. 122 No. 9, pp. 2121-2141. https://doi.org/10.1108/IMDS-12-2021-0782

Affinity diagram Capital categories	**Capital elements**
Social & Community capital	Relational value
	Identity
	Community interactions
	Human capital
	Political capital
	Cultural capital
	Shared value created by each entity
	Network of relationships
Financial	Resource exploitation
	Fixed assets
	Digital Assets
System capital	Organizational systems and related procedures
	Business process improvement
Intellectual capital	Human expert
	Innovation & learning
	Infrastructure
	Patents &copyrights
	Capabilities
	Knowledge
Technology Capital	Investment in technology
	Firm-specific technology development

Table 1: Affinity diagram for Capital categorization

Value Generation Objects	**Value Generation Object dimensions**	**Value Generation Object elements**
DLT Ecosystem Member Capitals	**Economic Capitals:** The amount of capital that a company needs to survive any risks that it takes	• **Financial resources.** Financial fund to fuel organization activities
		• **Assets.** Resources with economic value that an organization owns or controls and could produce positive economic value
		• **Digital assets.** Data as asset, that create value for the organization that owns them
	Social Capitals: Intangible assets established through social behavior and organization relationships developed in the economic community	• **Role and position in ecosystem.** Refers to the value of deep versus artificial collaboration between ecosystem members and how can the dominant actors affect it.
		• **Channel partner value and relational capital.** Relations with ecosystem members and the value that stems from each specific type of relation.
	Business Value: Other forms of value beyond economic value. Intangible forms of value in a collaborative synergistic process.	• **Customer perspective capitals** Attitudes and behaviors that affect brand choice, purchase frequency and loyalty enhance the trust between an organization and its customers. It impacts DLT ecosystem expansion
		• **Functional value.** Product or service attributes that provide functional utility to the customer or its partners. It is related to the potential of ecosystem members to adopt DLT.
		• **Intra-firm knowledge.** DLT ecosystem members leverage data access, and transparency to create new knowledge and upscale trust between each other. Value generated from data exploitation. Value an innovation actuated by data exploitation through collaboration.
		• **Structural Capital.** Refers to means, such as processes and information, that help an organization retain knowledge, share innovation and achieve its objectives. It Is related with capability resources, productivity and process value.

Distributed Ledger Technology Capitals	**Distributed Ledger Technology Development Capitals:** Capitals needed for the development of technology	• **Financial Capital.** Refers to financial resources to fund the development of the technology
		• **Human Capital.** Refers to human resources needed for the development of the technology
		• **Infrastructure Capital.** Technologies methods, processes and physical resources that support the development of the technology
	Social Capitals: Development, sharing principles and boundaries of the technology	• **Scale up potential.** Refers to the value of network effects created through the mass adoption of the technology and ecosystem expansion.
		• **Maturity of technology.** The more mature a technology the more it has been adopted and the more challenges it has met on the way to achieve its objectives

Table 2: Components of DLT Business Ecosystem Value Generation Objects

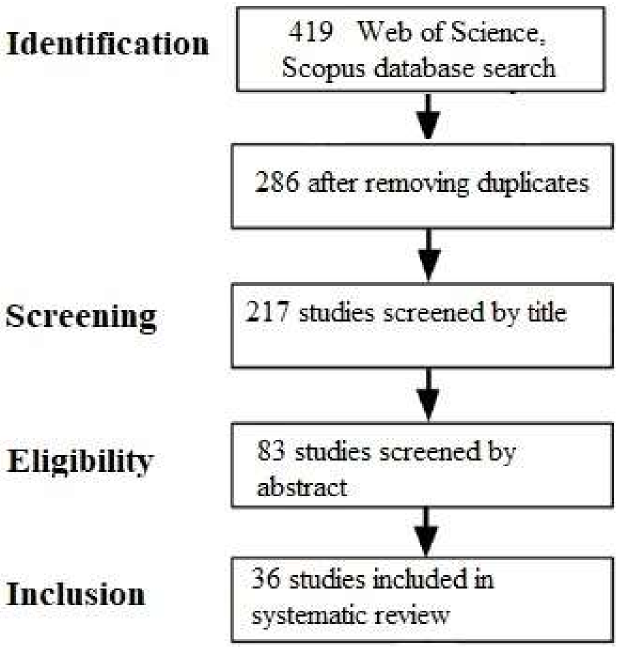

Identification
419 Web of Science, Scopus database search
286 after removing duplicates
Screening
217 studies screened by title
Eligibility
83 studies screened by abstract
Inclusion
36 studies included in systematic review

Figure 1. Capitals and Value system literature review methodology steps

The Impact of AI on the Accounting Profession

Patrick Buckley (University of Limerick)

Research In progress

Abstract

This work-in-progress paper describes a research paper that aims to explore the impact AI at the level of professions. In the modern world, professions are characterized by distinct attributes and practices that set them apart from other occupations. They typically possess a systematic body of theory, professional authority, community sanction, ethical codes, and a unique professional culture. They play a crucial role in creating and maintaining institutions, with different professions focusing on various aspects such as cultural-cognitive frameworks. Professions also tend to have a specific career pathways, norms and codes that are determined by the profession rather then the organisations that professionals work for. At a time when AI is threatening disruption across the entire labor market, this research explores how that disrpution may present at the level of professions.

Keywords: Artificial Intelligence, Labor Market Disruption, Future of Work, Profession

1.0 Introduction

1.1 The Development of Artificial Intelligence

The idea of machines replicating or exceeding human intelligence has been anticipated long before the advent of digital computers. Isaac Asimov introduced the concept of sentient robots and the "Three Laws of Robotics" in 1942 (Asimov, 1950). The origins of formal AI research are often traced back to the Dartmouth Summer Research Project on Artificial Intelligence, held in 1956 (Nilsson, 2009) The field has been notably unpredictable, with cycles of inflated optimism (Crevier, 1993) followed by periods of reduced expectations, known as “AI winters” (Nilsson, 2009).

Currently, AI research is in a phase of sustained growth and enthusiasm. New methods such as genetic algorithms have emerged, while older techniques like neural networks have been revitalized through innovations like deep learning. These theoretical advancements, coupled with increasingly powerful computing platforms and vast data sets generated by the internet, have enabled AI-driven progress in areas such as voice assistants and autonomous vehicles (Badue et al., 2021; Hoy, 2018) . Additionally, AI has made significant strides in fields such as finance, medical decision support, recommender systems, facial recognition, and machine translation (Marr, 2019).

Making predictions of technological development in such a rapidly evolving field is notoriously challenging. One common benchmark in such forecasts is the achievement of artificial general intelligence (AGI), where an AI system attains human-level general intelligence (Baum et al., 2011). Summarizing multiple surveys of AI experts, Bostrom (2016) offers the following median estimates: a 10% chance of AGI by 2022, a 50% chance by 2040, and a 90% chance by 2075.

Despite these optimistic projections, there remains significant uncertainty surrounding the future trajectory and impact of AI (Autor & Dorn, 2013; Mindell & Reynolds, 2022). Some experts predict a scenario where AI systems develop even more advanced AI, potentially triggering a "Cambrian explosion" of intelligence (Muehlhauser & Salamon, 2012). Such forecasts often view AGI as a milestone on the path toward systems that far surpass human cognitive capabilities (Bostrom, 2016). However, this view is not universally held (Mindell & Reynolds, 2022). While acknowledging recent progress, some argue that the path to advanced AI might be more challenging than enthusiasts believe (Penrose & Gardner, 2002). Certain researchers contend that intelligence is fundamentally non-algorithmic, implying that deterministic Turing machines cannot replicate true intelligence (Penrose & Gardner, 2002). Additionally, there is a philosophical debate about whether it is valid to conceive of intelligence as an isolated attribute of a singular entity (Clark, 2005). From this perspective, consciousness and intelligence may instead emerge from a broader cultural feedback loop and cannot be created without a larger social context (Dennett, 2017).

At a more practical level, developing AI models and systems is a complex task, with numerous challenges despite substantial investments from various stakeholders. One key issue is data quality and quantity; most AI systems, especially those based on deep learning, require vast amounts of high-quality, reliable data, which is costly and time-consuming to gather (Halevy et al., 2009)Biased or incorrect data can produce flawed models, posing risks in sensitive applications like finance or criminal justice (Zhang et al., 2018) , and unvetted data can leave AI vulnerable to adversarial attacks (Papernot et al., 2016). Regulatory oversight is another concern, as the increasing influence of AI demands a framework for accountability, especially in critical domains like healthcare and justice (Mindell & Reynolds, 2022). Additionally, the high costs associated with training models like ChatGPT may lead to a concentration of power, limiting broader access and raising questions about data distribution and integrity (Philip Chen & Zhang, 2014). Continuous learning is also challenging, as AI models require frequent updates to remain relevant, yet relying on internet-sourced data introduces validation issues (Kumar et al., 2023). Ethical concerns arise from the opaque nature of many AI models, complicating oversight and accountability in automated decision-making, particularly in high-stakes areas like healthcare and criminal sentencing (Penrose & Gardner, 2002).

In summary the future trajectory of AI development is quite uncertain, with many theoretical and practical challenges. However, given the recent pace of development and the enormous amounts of resources that are being invested in AI development, it would be unwise to assume that AI development will slow in the near future.

1.2 The Impact of Artificial Intelligence

There is also significant uncertainty regarding the societal impact of AI, with predictions ranging from highly optimistic to deeply pessimistic. Optimists, such as Kurzweil (Kurzweil, 2005) , envision AI's positive contributions, including accelerated solutions to global challenges like resource depletion and climate change. In this scenario, AI and robots would handle both physical and cognitive tasks, freeing individuals from labor and offering more choices for personal pursuits, whether in entertainment, creative activities, or traditional economic roles. Conversely, pessimists warn of potential negative outcomes. Bostrom (Bostrom, 2016) suggests that humans, as a cognitively inferior species, may struggle to control more advanced AI, akin to how animals cannot comprehend human motivations. This could lead to scenarios where humanity is sidelined or even faces existential threats (Joy, 2000). Additionally, concerns have been raised about the erosion of human agency as AI systems become practically and philosophically superior (Harari, 2016).

Even if AI systems do not surpass a cognitive threshold that places them beyond human control, skeptics raise significant concerns about the widespread deployment of AI (Arntz et al., 2016). While predictions that AI and robots will handle most or all labor may appear positive, this scenario brings substantial challenges. For instance, the decline in skills like navigation due to reliance on satellite systems illustrates a pattern where tools evolve from supportive "oracles" to dominant "sovereigns," potentially leading to learned helplessness (Bostrom, 2016). Additionally, as AI systems take over economic activities, social and political power may become concentrated in the hands of those who control these systems (Mindell & Reynolds, 2022). This trend could exacerbate existing inequalities, leading to a future where power is concentrated among a small elite, while the majority of people experience diminished agency (Harari, 2018).

1.3 AI and the Labor Market

The uncertainty surrounding long-term, macro-level predictions about AI is also found in shorter-term forecasts, particularly regarding its impact on employment and labor market structures. Technological advancements have historically influenced labor markets, altering the nature and distribution of work without consistently reducing the total volume of available jobs. While there is general agreement that technological change reshapes job roles and reallocates tasks, the broader effect on overall employment levels remains uncertain and varies depending on specific industry dynamics and the adaptability of the workforce.

The prevailing consensus on technological impacts on the labor market is being challenged by the rise of AI and the increasing ubiquity of information technology. The human monopoly on cognitively demanding tasks is being broken (Loebbecke & Picot, 2015). Rifkin (1995) argues that a new economic era is emerging where fewer workers are needed to meet global production demands. Similarly, Ford (2009) notes that as automation and digital transformation continue, labor costs will represent a shrinking portion of company expenses. Empirical studies indicate substantial losses of middle-class jobs and the automation of routine cognitive tasks, signaling potential future disruptions (Levy & Murnane, 2013). Some foresee this trend accelerating without end; Kurzweil (2013) predicts that AI will soon match and surpass human cognitive capabilities, leading to obsolescence in all economic roles. Levy and Murnane (2013) further assert that computers will eventually handle all tasks guided by logical rules or statistical models, including complex processes simplified through structured approaches.

A growing number of academic studies are attempting to quantify these risks. Current research aimed at evaluating the impact of AI automation on occupations was initiated by Frey and Osborne (2017). Their research suggested that 47% of all jobs in the United States might be at risk of automation by 2030. Broadly speaking, the emerging consensus from the latest research is that AI will have a significant destructive effect on at least a portion of the labor market, which will be offset by an unknown degree by transforming some roles and creating new opportunities (Bankins et al., 2024; Cramarenco et al., 2023). At a an operational level, current AI technologies are expensive and difficult for most businesses to adopt (Naudé, 2021). The integration of AI into organizational practices has implications for workers' experiences and job designs, with effects observed at individual, group, and organizational levels (Bankins et al., 2024). Given that this labor market dislocation is happening simultaneously with other trends such as aging populations, rising protectionism and climate change, there is an urgent need for research into this phenomena that can help to both inform and guide policy makers' decision making.

1.4 The Accounting Profession

Professions today are characterized by distinct attributes and practices that set them apart from other occupations. They typically possess a systematic body of theory, professional authority, community sanction, ethical codes, and a unique professional culture (Greenwood, 1957). In the modern world, professions play a crucial role in creating and maintaining institutions, with different professions focusing on various aspects such as cultural-cognitive frameworks, normative prescriptions, or coercive authority (Scott, 2008). Membership of a profession is usually a source of significant status and resources to individuals and is seen as a marker of personal success. For a profession as a whole, it's claimed knowledge base is both a status marker and a large part of the source of value of a profession to wider society (Gardner & Shulman, 2005). The claim that AI systems can commodotize knowledge at scale is particulary disruptive to professions.

2.0 Research Question

In this work-in-progress research paper, we will explore the potential impact AI on how professions are generally organised. We use semi-structured interviews with members of a specific profession, namely the Accounting profession, to explore the potential impact of AI at the level of the profession. We have currently conducted 8 interviews, with approximately 12-15 more interviews to be conducted. Thematic analysis will be conducted to identify key findings, which will be presented at UKAIS 2025.

References

Arntz, M., Gregory, T., & Zierahn, U. (2016). *The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis*. https://doi.org/10.1787/5jlz9h56dvq7-en

Asimov, I. (1950). *I, Robot*. Doubleday.

Autor, D. H., & Dorn, D. (2013). The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. *American Economic Review*, *103*(5), 1553–1597. https://doi.org/10.1257/aer.103.5.1553

Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., Jesus, L., Berriel, R., Paixao, T. M., & Mutz, F. (2021). Self-driving cars: A survey. *Expert Systems with Applications*, *165*, 113816.

Bankins, S., Ocampo, A. C., Marrone, M., Restubog, S. L. D., & Woo, S. E. (2024). A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior*, *45*(2), 159–182. https://doi.org/10.1002/job.2735

Baum, S. D., Goertzel, B., & Goertzel, T. G. (2011). How long until human-level AI? Results from an expert assessment. *Technological Forecasting and Social Change*, *78*(1), 185–195. https://doi.org/10.1016/j.techfore.2010.09.006

Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies* (Reprint edition). Oxford University Press.

Clark, A. (2005). Intrinsic Content, Active Memory and the Extended Mind. *Analysis*, *65*(1), 1–11. JSTOR.

Cramarenco, R. E., Burcă-Voicu, M. I., & Dabija, D. C. (2023). The impact of artificial intelligence (AI) on employees' skills and well-being in global labor markets: A systematic review. *Oeconomia Copernicana*, *14*(3), Article 3. https://doi.org/10.24136/oc.2023.022

Crevier, D. (1993). *AI: the tumultuous history of the search for artificial intelligence*. Basic Books.

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds* (1 edition). W. W. Norton & Company.

Ford, M. R. (2009). *The lights in the tunnel: Automation, accelerating technology and the economy of the future*. Acculant publishing.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Gardner, H., & Shulman, L. S. (2005). The professions in America today: Crucial but fragile. *Daedalus*, *134*(3), 13–18. https://doi.org/10.1162/0011526054622132

Greenwood, E. (1957). Attributes of a Profession. *Social Work*, *2*(3), 45–55. https://doi.org/10.1093/sw/2.3.45

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, *24*(2), 8–12. IEEE Intelligent Systems. https://doi.org/10.1109/MIS.2009.36

Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow*. Random House.

Harari, Y. N. (2018). *Homo Deus: A Brief History of Tomorrow* (Illustrated edition). Harper Perennial.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, *37*(1), 81–88.

Joy, B. (2000). Why the future doesn't need us. *Wired Magazine*, *8*(4), 238–262.

Kumar, M., Mani, U. A., Tripathi, P., Saalim, M., Roy, S., & Roy Sr, S. (2023). Artificial Hallucinations by Google Bard: Think Before You Leap. *Cureus*, *15*(8).

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Penguin.

Kurzweil, R. (2013). *How to create a mind: The secret of human thought revealed*. Penguin.

Levy, F., & Murnane, R. J. (2013). Dancing with robots: Human skills for computerized work. *Washington, DC: Third Way NEXT*.

Loebbecke, C., & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *The Journal of Strategic Information Systems*, *24*(3), 149–157. https://doi.org/10.1016/j.jsis.2015.08.002

Marr, B. (2019). *Artificial intelligence in practice: How 50 successful companies used AI and machine learning to solve problems*. John Wiley & Sons.

Mindell, D. A., & Reynolds, E. (2022). *The work of the future: Building better jobs in an age of intelligent machines*. MIT Press.

Muehlhauser, L., & Salamon, A. (2012). Intelligence Explosion: Evidence and Import. In A. H. Eden, J. H. Moor, J. H. Søraker, & E. Steinhart (Eds.), *Singularity Hypotheses: A Scientific and Philosophical Assessment* (pp. 15–42). Springer. https://doi.org/10.1007/978-3-642-32560-1_2

Naudé, W. (2021). Artificial intelligence: Neither Utopian nor apocalyptic impacts soon. *Economics of Innovation and New Technology*, *30*(1), 1–23. https://doi.org/10.1080/10438599.2020.1839173

Nilsson, N. J. (2009). *The Quest for Artificial Intelligence* (1 edition). Cambridge University Press.

Papernot, N., Mcdaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). *The limitations of deep learning in adversarial settings*. 372–387. Scopus. https://doi.org/10.1109/EuroSP.2016.36

Penrose, R., & Gardner, M. (2002). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (1 edition). Oxford University Press.

Philip Chen, C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347. Scopus. https://doi.org/10.1016/j.ins.2014.01.015

Rifkin, J. (1995). *The end of work: The decline of the global labor force and the dawn of the post-market era.* ERIC.

Scott, W. R. (2008). Lords of the Dance: Professionals as Institutional Agents. *Organization Studies*, *29*(2), 219–238. https://doi.org/10.1177/0170840607088151

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). *Mitigating Unwanted Biases with Adversarial Learning*. 335–340. Scopus. https://doi.org/10.1145/3278721.3278779

Responsible Generative AI in Higher Education: A Brazilian Perspective

Prof. Dr. Renato de Oliveira Moraes
Universidade de São Paulo, Brazil

Prof. Crispin Coombs
Loughborough University, UK

Research In Progress

Abstract

This study explores the integration of Generative AI (GenAI) in Brazilian higher education, focusing on its responsible use and unique challenges in Brazil. Despite global advancements, Latin America lags in establishing GenAI policies, which may deepen educational inequalities. Through a multi-case analysis of public and private institutions, the proposed research will examine GenAI's roles in teaching, learning, and administration, assessing its benefits and risks. Findings will inform policy recommendations tailored to Brazil's educational landscape, aiming to balance accessibility, ethics, and effectiveness in GenAI adoption. It is anticipated that this research will provide insights for policy development to guide GenAI's role in Brazilian education responsibly.

Keywords: Artificial Intelligence, Generative AI, Higher Education, Responsible AI use

1 Introduction

Advances in AI technologies have enabled the application of AI in many fields, including higher education. One of the most significant AI advancements concerns Generative AI (GenAI). In 2020, the launch of ChatGPT generated widespread interest in generative AI, reaching 100 million users within just a few months. For comparison, it took TikTok nine months to achieve the same user milestone (Hu, 2023). ChatGPT, developed by OpenAI, is the most well-known generative AI tool, alongside Copilot from Microsoft and Gemini from Google. These are examples of large language models (LLMs), which are predictive models that require substantial computing power for development. ChatGPT is built on a machine-learning framework, 'Transformers', which was introduced in 2017. It is pre-trained using extensive amounts of internet data, enabling it to generate text in response to user inputs, which is why it's referred to as a 'Generative Pre-trained Transformer' (Webb, 2023). Many educational organisations and learners have been quick to engage with GenAI, leading to international policies to guide how AI technologies should be used. For example, the recommendations from UNESCO and other institutions and individuals emphasise the importance of using AI effectively to advance the Sustainable Development Goals (SDGs) and the Global Education 2030 Agenda (Sabzalieva & Valentini, 2023).

Responsible use of GenAI in higher education involves principles that seek to ensure that this technology benefits society in an ethical and safe manner, and that it does not create or increase the marginalization of historically discriminated groups (UNESCO, 2023). It includes practices that aim to ensure that GenAI respects human rights, promotes inclusion and diversity, and supports sustainable development. This means data privacy protection, transparency and explainability (which can be challenging when using neural networks), avoidance of discrimination and bias, content labelling, education and training of teachers and students, validation and monitoring of security and suitability for educational use, and, lastly, inclusion and equitable access to technologies.

However, access to and utilisation of AI technologies differ in higher education countries. For example, the European Union is attempting to establish itself as a global hub for AI tool development while striving to ensure the tools are trustworthy and human-focused. By contrast, Asian Universities have applied bans on using GenAI for credit-bearing activities (e.g., the University of Hong Kong), and India prohibits using GenAI or other electronic tools during examinations. However, Latin America appears to lag behind many parts of the world in responding to GenAI, with few specific policies or initiatives targeted at GenAI (Hsu and Ching, 2023). Thus, rather than bridging the gap between groups with access to quality education and those without, this situation may perpetuate and even deepen the divide.

In Brazil, the Internet Civil Rights Framework was enacted in 2014, establishing the right to exercise citizenship in digital media, as well as diversity and freedom of expression on the internet, and in 2018 the General Personal Data Protection Law was enacted, which deals with privacy and the use of personal data. These two laws would correspond to the first of the seven steps suggested by UNESCO (2023) for government agencies to regulate GenAI in order to harness its potential in various sectors, including education. A fake news law (tabled in 2020) is still under discussion in the Brazilian

Congress, which deals with combating the spread of false news and regulating and holding the actions of big tech companies accountable. With regard specifically to AI, a project of law has been under discussion in the Senate since 2023 that defines the foundations, principles, prohibitions on use and sanctions related to AI in Brazil, as well as the rights of people affected by the use of these tools. More recently, in 2024, with the sophistication of AI tools, the Superior Electoral Court determined the mandatory requirement of notice when AI is used in electoral propaganda, and prohibited the use of audiovisual content generated or manipulated by AI to create, replace or alter the image or voice of a living, deceased or fictitious person. While these steps demonstrate an increasing awareness among Brazilian regulatory bodies of the need to develop responsible AI use policies, guidance in the education section remains in its infancy.

This lack of policy guidance or regulation leaves the Brazilian Higher Education sector vulnerable to questionable use of GenAI tools. For example, academic institutions in the Laureate educational network in Brazil have used AI to mark essays without students' consent or knowledge (Domenici, 2020). Barros et al. (2023) suggest that this example demonstrates how cost-cutting strategies to increase profits can place additional pressures on academics beyond simply adjusting to new teaching methods. However, Chassignol et al. (2018) offer a more optimistic perspective on the potential applications of artificial intelligence to enhance the teaching and learning experience. They argue that AI should augment the role of educators by creating new learning opportunities rather than replacing teachers or fundamentally altering the educational process. These examples illustrate considerable debate regarding the use of GenAI in education. In this context, and with the evidence that Latin America is lagging in its response to GenAI in education, the objective of this research in progress is to analyse how GenAI is being used in higher education in Brazil and to provide recommendations for the development of new policies and guidance for responsible use.

2 Observations From The Literature

There has been considerable interest in using GenAI in higher education. Studies on this topic can be divided into four themes: the educational benefits, the risks, changes in educational delivery approaches, and educational policy responses.

A range of potential benefits from using GenAI in education have been identified. These benefits include improving students' writing, critiquing and reasoning through assessing ChatGPT outputs or using ChatGPT as a virtual assistant to answer student questions. Further benefits include reducing inequalities by providing additional support for non-native English-speaking students and helping students with mental health issues, through chatbots to support students experiencing degression or anxiety (de la Torre, & Baldeon-Calisto, 2024).

By contrast, several risks have been identified. Concerns regarding the use of GenAI in higher education primarily revolve around issues of cheating and plagiarism. Educational institutions utilise plagiarism detection programs like Turnitin, CopySpider, and Plagscan. There is also a tool developed to identify text generated by ChatGPT called GPTZero (Atlas, 2023; Sullivan et al, 2023). The emphasis on responsible use underscores the significance of addressing plagiarism in a broader context—beyond the traditional definition of copying someone else's work. In this case, the issue is using an automated tool to produce academic assignments rather than direct copying. Atlas (2023) recommends implementing citation practices, increasing transparency, and establishing continuous monitoring to tackle these challenges. Further risks include limitations in the ability to access GenAI technology because of cost, as well as ethical and privacy concerns regarding the sharing of student data (Ferreira Mello et al. 2024). Risks have also been identified regarding compromising academic integrity and some educators requiring training to take advantage of GenAI (de la Torre, & Baldeon-Calisto, 2024). Ferreira Mello et al. (2024). also highlight the importance of educators' needing to be able to explain outcomes in their teaching and that the black-box nature of GenAI does not lend itself to being able to offer detailed explanations.

The interest in using GenAI for education has led to changes in the educational delivery approaches in some contexts. Barros et al. (2023) report that using GenAI has led to changes in role design for educators, with their professional identity shifting from instructors to facilitators of learning. They also found that educators reported that GenAI was being used to provide base content for lectures, presentation slides, curriculum and module design that the educational professional adapted and enhanced.

According to Vygotsky (1988) the student's relationship with the content is mediated by instruments (e.g. book, pencil and blackboard), signs (e.g. words, images and symbols that allow us to remember, imagine, compare or evaluate) and the Other (the partner of the self, with whom social relationships are established, such as the teacher, the colleague or the tutor). AI is a new instrument that transforms language, but it is not the Other. Therefore, AI cannot internalise socially constructed culture, as it does not have higher psychological functions and, consequently, cannot eliminate the role of the teacher, even though it can expand and change the teacher’s actions.

Given these potential benefits, risks and changes to education practice highlighted in the literature, it is striking that less than 10% of schools and universities have institutional policies or guidance for the use of generative AI (WEF, 2024). This has led to researchers proposing possible responses such as an ecological framework for higher education policy for Gen AI (Chan, 2023) or recommending the adjustment of existing regulations to provide a quick and timely response to a rapidly advancing technology (Dotan Techbetter et al., 2024). Furthermore, some policy bodies and academic groups have provided advice, such as WEF’s 7 Principles for AI in Education (WEF, 2024) and the UK Russell Group’s five principles for responsible Gen AI use (Russell Group, 2024). However, Dotan Techbetter et al. (2024) argue that top-down approaches that dictate how Gen AI is used in HE are unlikely to fit with HE values and ways of working. Consequently, they advocate an educational rather than regulatory approach to using GenAI. These studies provide a valuable foundation for understanding the GenAI landscape in education. However, Latin America, particularly Brazil, needs more attention in these studies. Further, the lack of engagement with GenAI in education and the development of policy guidance is surprising given Brazil is one of the highest users of GenAI applications (Freitas, 2024).

In summary, the focus on GenAI in higher education heavily emphasises information systems management while paying insufficient attention to the specific skills of teachers. There is significant discussion about ethical issues related to the use of AI, such as transparency, privacy, fairness, and accountability—topics that are highly relevant to information systems management. However, there is a lack of discourse on higher education didactics, andragogy, or university pedagogy. Additionally, in Brazil, there is no evidence to suggest that teachers are well-versed in these important themes.

This oversight seems to overshadow certain discussions, particularly the connection between the use of generative AI and learning theories. The way students learn, or how teachers perceive student learning, influences teaching methods and, consequently, impacts the integration of AI in their courses. Therefore, this research aims to understand how GenAI is being used in Brazil for higher education, why Brazil is lagging other nations in developing GenAI educational policy, and to provide recommendations for suitable policy responses.

3 Brazilian Higher Education Context

While it is possible to categorise Brazilian Higher Education Institutions (HEIs) into two groups, private and public institutions, doing so can obscure the presence of many high-performing private HEIs, such as numerous religious institutions, as well as public HEIs of lower quality, particularly those located in isolated regions of the country. Enade is an assessment of undergraduate courses administered by Brazil's Ministry of Education and Culture (MEC). We used data from the last 3 Enade evaluations (in 2019, 2017 and 2014) of production engineering courses to build a taxonomy of courses with cluster analysis. The number of groups to form was decided based on the dendrogram generated with hierarchical cluster analysis (Figure 1). This led to the decision to form 3 groups that balance the homogeneity inside groups, the heterogeneity among the groups, and a short number of groups. Using the K-means method, these 3 clusters were generated. The average value of Enade grades among the groups is statistically different (Table 1, p-value < 0.05). Therefore, they were named low, medium and high quality.

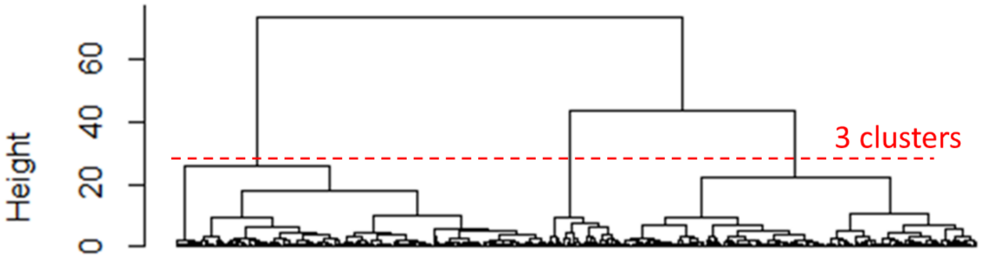


Figure 1 – Dendrogram

Table 1 shows the characteristics of each course group. The first group (low quality) is dominated by private HEIs – 62% of these courses are in this group, and 95% of the group is from these HEIs. The courses from public HEIs are divided into the other two groups (medium and high quality) – 45% in each group. Looking at private courses in the medium-quality group, we find HEIs linked to the church (mostly Catholic), smaller institutions, and units from big educational groups, which are often accused of commodifying education. This suggests that the willingness to offer better quality courses cannot be adequately explained by the size of the HEI and/or the nature of its management and that other essential elements are not captured by the INEP data used in this study.

	Formed Clusters			
Administrative Category	**Low**	**Median**	**High**	**Total Courses**
Private	152 (62%) (95%)	81 (33%) (68%)	12 (5%) (24%)	245 (100%) (74%)
Public	8 (10%) (5%)	38 (45%) (32%)	38 (45%) (76%)	84 (100%) (26%)
Total	160 (49%) (100%)	119 (36%) (100%)	50 (15%) (100%)	329 (100%) (100%)
Year	**Average score at Enade**			
2019	1,71	2,76	3,98	
2017	1,76	2,80	3,87	
2014	1,33	2,36	3,72	

Table 1 – Clusters of undergraduate courses

Figures 2 and 3 illustrate that most courses and vacancies in higher education are in private institutions. The production engineering students are in private HEI, mainly low (62%) and medium (33%) quality courses. Therefore, from a national perspective, improvements in these courses would impact more students.

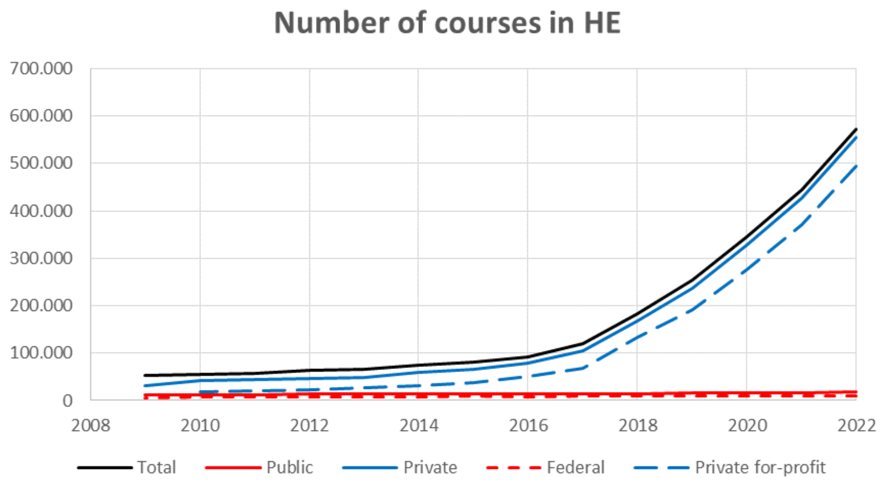


Figure 2 – Number of undergraduate courses. Source: INEP (2022a)

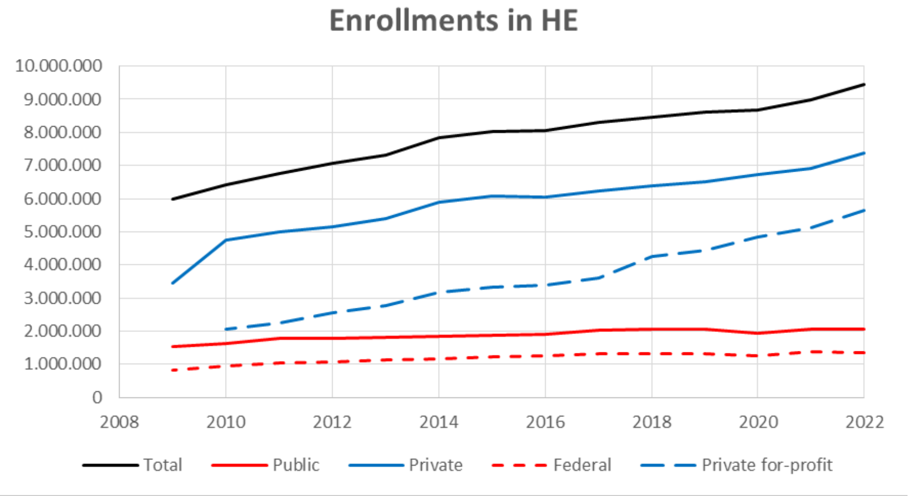


Figure 3 – Number of enrolments. Source: INEP (2022a)

The groups have different objectives. In the lower quality group, the focus is on improving education quality, mainly due to the consistently low performance observed in the MEC evaluation system. This low quality has serious consequences, which can even result in the cancellation of courses or the de-accreditation of HEIs. By contrast, higher-quality HEIs, particularly public ones, face the challenge of increasing the number of vacancies and enrolments in undergraduate programs. Each group aims to utilise the strengths of the other while preserving its own. Public HEIs tend to offer better courses but often have vacancies and serve fewer students. Their main challenge is to recruit a larger number of students, similar to how private HEIs operate. In contrast, private HEIs enrol a larger number of students, but their challenge lies in improving the quality of education, akin to the efforts of public HEIs.

Assuming that Brazilian HEIs can be distinctly categorised into three groups, their desired uses and contributions of GenAI may differ significantly. This leads to the research question: How are these three groups of HEIs using and planning to use GenAI in their undergraduate courses? Since different courses have varying content, practices, and internal cultures, which may act as confounding factors to interpretation, this study will focus on Production Engineering programs.

Individual experiences of different HEIs in Brazil reflect an evolving understanding of the role of GenAI in the country. Analysing Brazilian experiences in the context of regulations from other countries could reveal the degree of alignment or global maturity. However, this approach would overlook the unique aspects of the Brazilian context, particularly the historical divide between public and private HEIs, as well as the significant expansion of courses and available spots in private HEIs, which tend to be owned by educational groups with shares traded on the stock exchange.

4 Method

Multiple case studies will be conducted to analyse the use of GenAI in higher education in Brazil. This research will focus on six case studies classified by quality group (low, medium and high) and type of HEI (public and private). The unit of analysis will be the course. According to Bielschowsky (2020), in 2018, 49% of enrollments in higher education were concentrated in ten educational groups, with five of these being publicly traded on the stock market. This highlights their significance in the higher education landscape. Cluster analysis of production engineering courses reveals notable differences between public and private HEIs in Brazil. Therefore, it is essential to include courses from these groups in the study cases.

Data will be collected mainly through interviews with course coordinators, professors, and students. Representatives of course coordinators may include the course coordinator, course coordination committee members, and core faculty members. In this group, the role of GenAI in the course will be assessed, as well as how GenAI initiatives fit together and are articulated with other actions to achieve the course objectives. As an auxiliary source of information, the course's political-pedagogical project (PPC) will also be analysed. Interviews will be conducted with professors who utilise GenAI resources in their disciplines to understand their role, how these resources align with the content being taught, and how they support the objectives of the discipline according to Bloom's taxonomy. Additionally, updated syllabi and lesson plans (when available) will be analysed. Lastly, interviews will be carried out with students who have previously taken at least one course that incorporated GenAI as an instructional tool.

The analysis of GenAI usage will focus on the following elements:

1. The evolving landscape of AI applications.
2. The characteristics of GenAI projects.

3. Supported processes: teaching, learning, and administration.

4. Intensity of GenAI usage

5. Types of usage:

   - Taxonomy of Bloom: dimensions and levels

   - The role of GenAI

6. The impact of GenAI usage on stakeholders.

5 Anticipated Contributions

This study is expected to offer valuable insights into the current use of GenAI in the Brazilian higher education sector and the benefits and risks associated with its application. The findings will serve as a basis for creating a policy framework for GenAI, aimed at promoting responsible usage in HEIs in Brazil.

6 References

Atlas, S. (2023). *ChatGPT for Higher Education and Professional Development: A Guide to Conversational AI.* https://digitalcommons.uri.edu/cba_facpubs/548

Barros, A., Prasad, A., & Śliwa, M. (2023). Generative artificial intelligence and academia: Implication for research, teaching and service. *Management Learning*, *54*(5), 597–604. https://doi.org/10.1177/13505076231201445

Bielschowsky, C. E. (2020) Tendências de precarização do ensino superior privado no Brasil. Revista Brasileira de Política e Administração da Educação, Goiânia, v. 36, n. 1, p. 241-271. Jan, 2020. https://doi.org/10.21573/vol36n12020.99946.

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, *20*(1), 38. https://doi.org/10.1186/s41239-023-00408-3

Chassignol, M.; Khoroshavin, A.; Klimova, A. & Bilyatdinova, A. (2018) Artificial Intelligence trends in education: a narrative overview. Procedia Computer Science. Vol 136, Pages 16-24. https://doi.org/10.1016/j.procs.2018.08.233.

de la Torre, A., & Baldeon-Calisto, M. (2024). Generative Artificial Intelligence in Latin American Higher Education: A Systematic Literature Review. *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, 1–7. https://doi.org/10.1109/ISDFS60797.2024.10527283

Domenici T (2020) Laureate usa robôs no lugar de professores sem que alunos saibam [Laureaute uses robots instead of lecturers without warning students]. *Agência Publica*, 30 April. Available at: https://apublica.org/2020/04/laureate-usa-robos-no-lugar-de-professores-sem-que-alunos-saibam/

Dotan Techbetter, R., Parker, L. S., Radzilowicz, J. G., & Dotan, R. (2024). *Responsible Adoption of Generative AI in Higher Education: Developing a &quot;Points to Consider&quot; Approach Based on Faculty Perspectives*. https://doi.org/10.1145/3630106.3659023

Ferreira Mello, R., Freitas, E., Pereira, F. D., Cabral, ; Luciano, Tedesco, P., & Ramalho, G. (n.d.). *Education in the age of Generative AI Context and Recent*

Developments. Retrieved November 8, 2024, from https://arxiv.org/pdf/2309.12332

Freitas, F. (2024). *Brazil is the fourth country that most accesses ChatGPT • Artificial Intelligence • Tecnoblog*. https://tecnoblog-net.translate.goog/noticias/brasil-e-o-quarto-pais-que-mais-acessa-o-chatgpt

Hsu, Y. C., & Ching, Y. H. (2023). Generative Artificial Intelligence in Education, Part Two: International Perspectives. *TechTrends*, *67*(6). https://doi.org/10.1007/s11528-023-00913-2

Hu, Krystal (2023). ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*. Retrieved from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

INEP 2022a. *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Censo da Educação Superior*. https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados

INEP 2022b. *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Indicadores de Qualidade da Educação Superior*. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-qualidade-da-educacao-superior

Knowles, M. S.; Holton III, E. F.; Swanson, R. A. (2009) Aprendizagem de resultados: uma abordagem prática para aumentar a efetividade da educação corporativa. Rio de Janeiro: Campus, 2009.

Kolb, D. A. Experiential learning: experience as the source of learning and development. Prentice-Hall Inc., New Jersey, 1984.

Russell Group. (2023). *Russell Group principles on the use of generative AI tools in education*. https://russellgroup.ac.uk/media/6137/rg_ai_principles-final.pdf

Sullivan, M.; Kelly, A. & McLaughlan, P. (2023) ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching*. Vol. 6 No. 1. https://doi.org/10.37074/jalt.2023.6.1.17

Sabzalieva, E. & Valentini, A. (2023). ChatGPT and Artificial Intelligence in higher Education - Quick start guide. UNESCO - United Nations Educational, Scientific and Cultural Organization.

UNESCO (2023). Guidance for generative AI in education and research. Paris, UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000390241

Vygotsky, L. S. (1988) Aprendizagem e desenvolvimento intelectual na idade escolar. In.: Vygotsky, L.S; Luria, A.R.; Leontiev, A. N. Linguagem, desenvolvimento e aprendizagem. 5. ed. São Paulo: Ed. Ícone, 1988.

Webb, M. (2024). *A Generative AI Primer*. JISC. https://nationalcentreforai.jiscinvolve.org/wp/2024/08/14/generative-ai-primer/

World Economic Forum (2024) *7 principles on responsible AI use in education | World Economic Forum*. Retrieved November 8, 2024, from https://www.weforum.org/stories/2024/01/ai-guidance-school-responsible-use-in-education/

Categorising viral videos based on the emotional intensity of football fans and non-football fans - A methodological study

Joseph Asamoah (Manchester Metropolitan University)

Abstract

This study explores how emotional responses and social identity influence the viewing and sharing of viral videos among football fans and non-fans. Grounded in social identity and social sharing of emotions theories, it suggests that identity and emotions drive the sharing behaviour in football fandom. According to social identity theory, fans align strongly with their own team (in-group) while often viewing rival teams and their supporters (out-group) less favourably. The social sharing of emotions theory further suggests that fans are more likely to share videos that elicit intense emotions. The study develops a methodological framework showing that, when exposed to a viral football video, fans demonstrate higher emotional intensity and are more inclined to share than non-fans. The findings highlight that the likelihood of sharing increases when viewers experience peak affective states, emphasizing the role of emotional intensity in the spread of football-related content.

Keywords: football fandom, YouTube, emotions, affective states, emotional intensity, distribution bell curve, virality, social media.

Introduction

Viral videos are videos that gain popularity by being shared and recommended through online word of mouth (France, Vaghefi and Zhao, 2016). Successful viral videos engage users and appeal to their emotions where they may be produced for a range of applications, including consumer marketing political campaigns and the creation of user generated content (UGC) (France, Vaghefi and Zhao, 2016). Some of these viral videos can be in the shape of football related videos which will be the prime focus of the study. The spectacle of football transcends the mere boundaries of the pitch; it invokes a profound communal experience characterised by a cascade of emotions that ripple through stadiums and across screens worldwide. The medium of the viral video has become a central vehicle for such experiences, capturing moments of triumph, defeat, unity, and individual prowess, distilling them into potent narratives that are consumed and shared by millions. In exploring the importance of emotions elicited by these videos, one observes a fascinating dichotomy: the die-hard football fan and the non-affiliated spectator. For fans, football is often inseparable from their identity; the sport elicits a profound affective commitment that is reflected in their response to viral content (Bandyopadhyay, 2024). Emotions here are heightened, tied to a sense of belonging, and often, to the collective memory of the fan community (Zubernis, 2023).

The viral football related video becomes not just a replay of a moment but a reaffirmation of their passion, loyalty, and solidarity with the team and its ethos. This emotional investment can amplify the joy of a victory or the agony of a loss, rendering each shared video a fragment of the larger narrative they live and breathe (Biscaia et al.,2012). Contrastingly, non-fans, who may lack allegiance to any club or knowledge of the game's intricacies, might engage with the same content from a different emotional perspective. Their responses might be coloured not by the fervour of loyalty, but by the universal appeal of human drama and athletic excellence that the sport encapsulates. Football videos, thus, have the power to evoke admiration, amusement, empathy, or awe from an audience uninitiated in the sport's culture (Zubernis, 2023). Football-related videos often evoke strong emotional responses, both from avid football fans and those who aren't as invested in the sport. For football fans, these videos can trigger intense feelings of joy, pride, or disappointment, depending on the content. This emotional connection is often tied to their deep psychological bond with their favourite teams and players (Zubernis, 2023). For instance, a video showing a last-minute goal or a dramatic penalty shootout can lead to a surge of excitement and a sense of achievement by football fans. In football folklore, Ole Gunner Solskajer sealing a treble for Manchester United with a quick-fire double in a champion's league final is a prime example of such an impact (Sidle, 2022). On the other hand, non-football fans might not share the same level of investment in the sport itself, but they can still be moved by emotional elements in these videos (Shakina , Gasparetto and Barajas, 2020).

Interestingly, research has shown videos that evoke strong emotions, whether positive or negative, are more likely to be shared and discussed (Wen et al., 2021;Pivecka , Ratzinger and Florack, 2022). This is true for both football fans and non-fans, although the specific emotions and reasons for sharing might differ between the two groups. Thus, it can be argued that there is a need to categorise the emotional characteristics of viral videos for the following reasons: Football fans are likely to engage more deeply with football-related videos by liking, sharing, and discussing them, while non-fans may watch but engage less. Fans typically have more knowledge about the sport and watch for specific details, while non-fans may watch for entertainment or because it's trending. Fans also have stronger emotional reactions, like excitement or disappointment, whereas non-fans view the content more casually. Football fans tend to watch such videos regularly, while non-fans only occasionally. Fans often align with specific teams or players, influencing their perception, whereas non-fans remain more neutral (Mastromartino, Chou and Zhang, 2018). It is important to state that football fans and non-football fans will both depict variations in affective shifts and emotional intensity which will be elucidated below.

Affective (Emotional) Shifts

An **affective** or **emotional shift** can refer to a change in feeling that occurs over time. It can be a shift from one emotion to another, such as from sadness to happiness (negative to positive), or it can be a shift in the intensity of an emotion, such as from mild anger to intense rage (Mitchell, 2021). For example, affective shifts can occur when one is feeling down and then suddenly something happens that makes the person feel happy (or vice versa). This could be something as simple as a football fan seeing a winning goal being scored or being ruled out for offside. Affective shifts in advertising refer to the deliberate changes in emotional tone throughout an advertisement to influence viewers' overall perception and response (Baumgartner, Sujan and Padjet, 1997). This technique leverages the idea that people's emotional reactions to videos are not static but can vary significantly from moment to moment. Previous research shows that viewers' overall judgments of a video ad are heavily influenced by the peak emotional moment and the final emotional state (Baumgartner, Sujan and Padjet, 1997). Furthermore, ads that build up to a strong emotional peak and end on a high note tend to be more memorable and positively received (Baumgartner, Sujan and Padjet, 1997).

Emotional Intensity

Ma (2024) classifies **emotional intensity** when someone experiences emotions to an unusual level of depth harnessing a constant stream of both positive and negative feelings, sometimes together, sometimes from one to another in a short period. Subsequently, emotional intensity in advertising refers to the strength or power of the emotional response elicited by an advertisement (Otamendi and Martin, 2020). Poels and Dewitte (2019) suggests that the intensity of an emotion is determined by the degree of pleasure or displeasure experienced, as well as the level of activation elicited.

In relation to sports, Annamalai et al. (2021) explained that sport fans are affected by the team's performance as their favourite team's success or failure is felt as a personal success or failure. A team's good or bad performance results in emotional reactions among fans. Furthermore, there has been some research on the emotional differences between football fans and non-football fans. Football fans experience a wide range of intense emotions related to their teams such as happiness, surprise, anger and sadness (Friedrich et al., 2020). These emotions can significantly influence their behaviour such as engagement (Shakina, Gasperetto and Barajas, 2020). This engagement can occur through social media views and the sharing of content (Neurolaunch, 2024; Zubernis, 2023). Another study indicated that while fans experience more intense emotions, particularly during losses, the difference in emotional intensity between fans and non-fans is relatively small (Friedrich et al., 2020).This suggests that while fans are more emotionally invested, non-fans also experience emotional fluctuations when watching sporting events, though to a lesser extent. Furthermore, it is important to indicate that the measurement of emotions can come in many forms including self-report , physiological and psychological measures (Asamoah, 2019; Li et al., 2020; Richardson et al.,2020).

Facial Expression Analysis

The data for the study was obtained from facial expression analysis - which is the process of automatically detecting, collecting and analysing facial muscle movement (Farnsworth et al., 2020). This process provides a robust and efficient way of eliciting and collating the emotional responses of its subject participants. The tool for analysing elicited emotions is the FaceReader. The FaceReader is a commercially available software program which can automatically analyse facial expressions regarding seven emotional states: happiness, sadness, anger, surprise, scared, disgust, and neutral, and allows researchers to analyse participants' facial expressions quantitatively (Yu and Ko, 2017).

FaceReader demonstrates high accuracy in classifying various expressions, with rates reported at 94% for Neutral, 82% for Scared, and other studies reporting performance rates ranging from 80 to 89% (Skiendziel et al., 2019). Although, past studies have suggested that “Disgusted” and “Angry” are two emotions that FaceReader recognises less effectively (Terzis, Moridis and Economides, 2013; Suhr, 2017). Consequently, validation studies when comparing the Facereader with other measures are less clear. Suhr (2017) analysed negative emotions in a video data set with the FaceReader and facial electromyography (fEMG), comparing both measures revealed that they were inconsistent. Additionally, Asamoah (2019) noted that there is a discriminant validity between self-report and the use of Facereader , with the drawbacks of using self-report measures aligned to influences by biases (subjectivity). In as much the justification for using this method can stem from (Lewinsky et al., 2014, p.4) that depicted that the “FaceReader is as good at recognising emotions as humans”.

Hence, based on the literature the following research questions can be explored to set the premise for the methodological framework.

RQ1: How can a framework for categorising groups (football fans and non-football fans) based on emotions elicited through facial expression analysis be developed?

Significance to research field: Researchers can gain insights into how different groups (e.g., football fans vs. non-football fans) experience and express emotions in various contexts. This can further help in understanding the emotional dynamics within and between groups (i.e, Goldenberg, 2024). Additionally, and as explained prior facial expression analysis provides an objective method to measure emotions, reducing the biases associated with self-reports (Lewinsky et al.,2014). Finally, understanding the emotional responses of different groups can help tailor marketing and advertising strategies and improve fan engagement (Biscaia et al.,2012).

RQ2: Do football fans and non-football fans have a different inclination to share a viral video in relation to their emotional intensity?

Significance to research field: Practically, marketers can create tailored content that resonates more deeply with specific groups. For example, emotionally intense content might be more effective for football fans, while different emotional appeals might work better for non-football fans which can also be further aligned with creating optimised ad campaigns (Choi, 2022). Furthermore, knowing which emotional intensities are more likely to be shared can help in crafting content with higher viral potential with the aim of having increased visibility and reach on social media platforms (Santos, 2018; Choi, 2022).

Theoretical Framework

To provide a holistic understanding of this study the social identity theory and the social sharing of emotions theory provided the theoretical lens. Social identity theory is a psychological theory that explains how people develop their sense of self based on their group membership (McLeod, 2023) . According to Henri Tajfel, the founder of social identity theory, people derive their sense of self from the groups they belong to, such as social class, family, and football team. Tajfel proposed that social categorisation is a cognitive process that helps us understand and identify objects and people (Tafjel et al.,1979). In the context of social identity theory, social categorization refers to the process of dividing people into groups based on shared characteristics such as race, gender, and nationality (Islam, 2014) .

Social identity theory posits that people tend to form in-groups and out-groups based on social categorization (Tajfel and Turner, 1979). An in-group is a group that a person identifies with, while an out-group is a group that a person does not identify with. People tend to view members of their in-group more positively than members of out-groups. This phenomenon is known as in-group favouritism (Noel, Wann and Brascombe, 1995). Football fandom within the confines of sports provides a typical example of how social identity theory works in practice. Football fans often identify strongly with their favourite teams and view other teams as out-groups (Hirshon, 2020). Fans derive their sense of self from their team membership and often feel a strong sense of loyalty towards their team.

When it comes to watching football-related videos, both football and non-football fans may have different perspectives on the in-group and out-group. Football fans who watch these videos are likely to view their favourite team as the in-group and other teams as out-groups. They may have a strong sense of loyalty towards their team and may exhibit in-group favouritism. This means that they are more likely to view their team positively and other teams negatively. Non-football fans who watch these videos may not have a strong sense of loyalty towards any team. They may view all teams as out-groups or may not have any strong feelings towards any team.

The way football and non-football fans view the in-group and out-group can influence their behaviour when watching football-related videos. Football fans who identify strongly with their team tend to be more emotionally invested in the videos (Zubernis, 2023). They may experience a range of emotions such as joy, happiness, or disappointment depending on the performance of their team. They may also engage in discussions or debates with other fans about the videos. Non-football fans, on the other hand, may watch these videos for entertainment purposes or to gain knowledge about the sport. They may not have any emotional attachment to the teams or players featured in the videos. Thus, this study will intend to prove that football fans and non-football players depict different emotional intensities when watching a football related viral video as exemplified within the social sharing of emotions theory.

According to the social sharing of emotions theory, individuals have an inherent tendency to share their emotional experiences with others, a behaviour that significantly amplifies and disseminates these emotions within their social circles (Rime, 1998). High-arousal emotions, whether positive (such as awe and amusement) or negative (such as anger and anxiety), play a pivotal role in this process (Nikolanikou and King, 2018; Wen et al., 2021; Susannah et al.,2023). Content that evokes strong emotional responses is more likely to be shared, as individuals seek to manage their emotional states and connect with others who share similar feelings (Nikolanikou and King, 2018; Wen et al., 2021; Susannah et al., 2023). This emotional contagion creates a ripple effect, enhancing the video's reach and impact. Furthermore, the act of sharing emotionally resonant videos fosters social bonding and reinforces group identity. When individuals share content that aligns with their values and beliefs, it strengthens in-group cohesion and promotes further sharing within the group (Hayes, Shan and King, 2017; Hoffman et al.,2020). This dynamic is particularly evident in videos that tell compelling stories or depict relatable situations, as viewers see their own experiences reflected and feel compelled to share (Hoffman et al., 2020).

There has been research that has synthesised concepts related to social identity and emotions. To cite an example, Stets and Turner (2014) identified that emotions are intimately related to the identity verification process whilst Mercer (2014) identified that group level emotions are distinct from individual emotions. However , there has been limited research that has explored a clear methodological framework for depicting the emotional intensity elicited and affective state within the context of football fandom and video stimuli. Shakina, Gaspereto and Barajas (2020) identified how emotions impact attendance at football matches, examining whether football fans prefer to watch highly competitive matches or matches between good teams with star-players whilst Doidge, Kossakowski and Mintert (2020) highlighted emotional engagement of football fans, particularly focusing on the ultras. Hence, there is a basis for larger studies into understanding the interplay between emotions and social identity. As pertaining to this research, it is relevant to recognise that emotions can be dichotomously categorised into positive and negative emotions (King, 2013; An et al., 2017). Anger, fear, sadness and disgust encompass negative emotions, and conversely, happiness and surprise make up positive emotions (An et al., 2017). To get a robust picture the research data focused primarily on happiness and surprise as these were the emotions that were elicited highly and per literature has a higher propensity for dissemination (Frigerri et al.,2014). Furthermore, the data was segmented into football fans and non-football fans.

Methods

Participant recruitment

A total of 60 sampled respondents (32 football fans and 28 non-football fans) undertook the facial expression analysis evaluation which was used as the main basis for the study. The respondents comprised lecturers and students from the University as well as carefully selected respondents selected from a freelance website. The students and lecturers were mainly solicited via email correspondence and through face-to-face requests. Participants from the

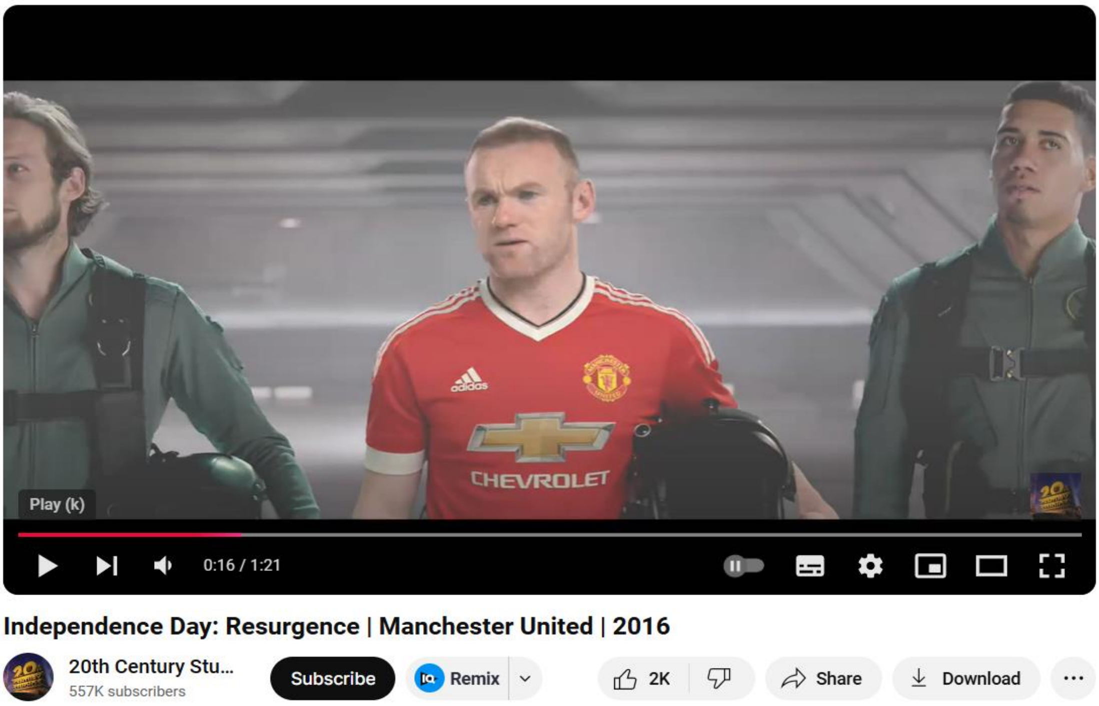
PeoplePerHour website were selected after they sent a proposal. Accepted proposals primarily took into consideration the gender and whether they were football fans or non-football fans to get a balanced perspective from the participants. The sample size of 60 was based on a priori assumption in relation to Berger (2011) which used 40 participants in a similar study undertaken to elicit emotions from participants. The use of stratified simple random sampling was adopted for remote participants on People Per Hour (PPH) who were selected after submitting their bid to undertake the study (Singh, 1996). Subsequently, two of the participants (one football fan and one non-football fan) data were eliminated from the sample study as their emotional responses were not able to be calibrated.

Materials

To undertake the study the data was obtained from facial expression analysis using the FaceReader 6 platform or they could do the test remotely (This also involved a self-recording of themselves which was subsequently uploaded into a Dropbox for further facial expression analysis).

Independence Day resurgence (Test stimulus)

The viral video was evaluated from a publicly accessible YouTube clip which depicted Manchester United players featured in an Independence Day movie trailer which was released alongside the film's launch, it capitalised on the buzz around "Independence Day: Resurgence," with the trailer likely also attracting viewers interested in the movie.



Independence Day: Resurgence | Manchester United | 2016

20th Century Stu...

557K subscribers

Figure 1. **Independence Day viral video stimulus (https://youtu.be/5YyUrlxnFww?si=dzIuPr2BkKem0Mgj).**

Procedure

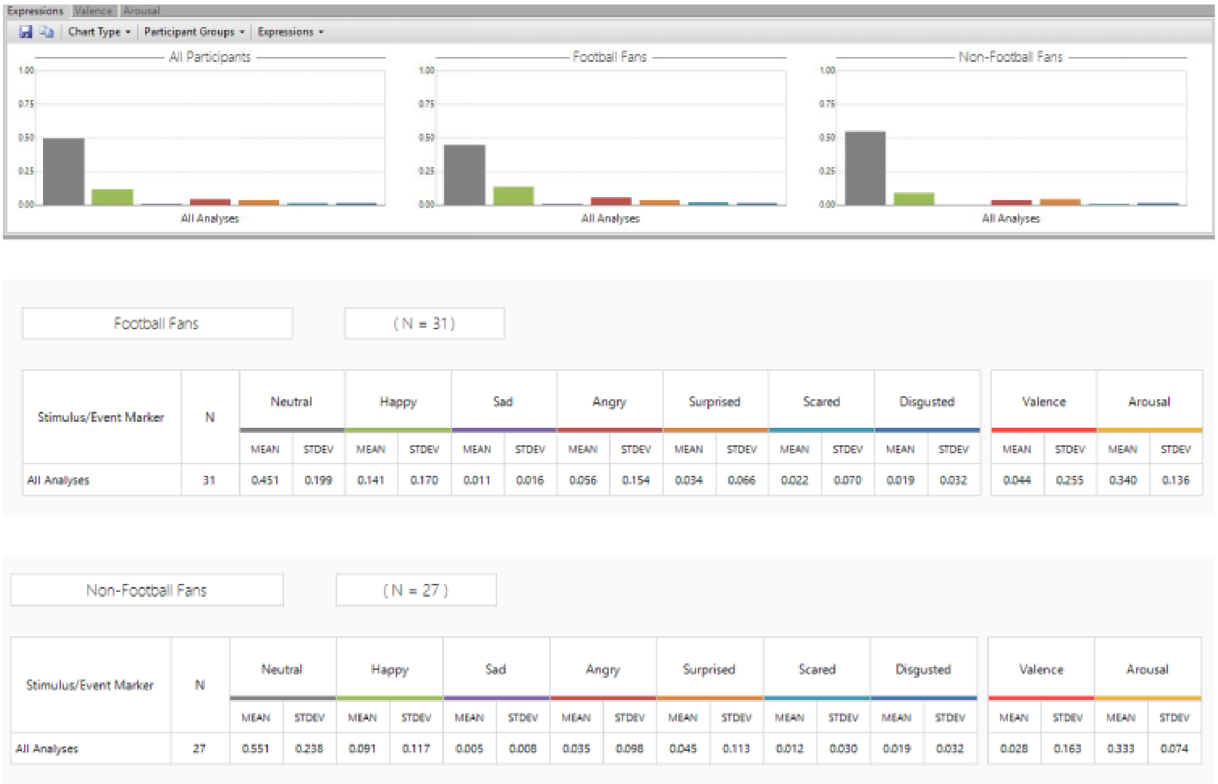
On site participants were required to read a participation information sheet and sign an ethical approval form prior to the start of the study. The participant information sheet depicted the entire process they will go through as well as the scope behind the study. Remote participants had to do likewise and check an online web form that stated that they agree with the modus operandi. Participants who took the study remotely were instructed to record themselves with any suitable recording software and a high-definition webcam, onsite participants had access to a testing suite, a laptop using the Face Reader 6 software and webcam.

The **test stimulus** is less than 1-minute 30s in length. The viral video was chosen due to its widespread circulation at that time – (177500 views in the first 28 days) and hypothesised ability to induce a measurable variation in the mean emotional intensities of the viewing participants.

Results

Emotions summary

The main emotional variables evaluated from the facial expression analysis study were the mean emotion intensities of **happiness** and **surprise** of each of the viewing participants. However, the FaceReader tool was also able to measure the emotional responses for neutral, anger, sadness , disgust and fear. Individual data can be found in the figshare repository. Observed in the figure 2 below are the collective mean emotional responses for both football and non-football fans.



Football Fans | (N = 31)

Stimulus/Event Marker	N	Neutral		Happy		Sad		Angry		Surprised		Scared		Disgusted		Valence		Arousal	
		MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV
All Analyses	31	0.451	0.199	0.141	0.170	0.011	0.016	0.056	0.154	0.034	0.066	0.022	0.070	0.019	0.032	0.044	0.255	0.340	0.136

Non-Football Fans | (N = 27)

Stimulus/Event Marker	N	Neutral		Happy		Sad		Angry		Surprised		Scared		Disgusted		Valence		Arousal	
		MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV	MEAN	STDEV
All Analyses	27	0.551	0.238	0.091	0.117	0.005	0.008	0.035	0.098	0.045	0.113	0.012	0.030	0.019	0.032	0.028	0.163	0.333	0.074

**Figure 2.	Independence Day viral video stimulus emotions data**

An overview of the emotions elicited by football fans “in group” and non-football fans “out group” from the viral video stimulus.

The mean emotional intensity **(for the entire video stimulus)** for happiness in relation to football fans is 0.170 whilst that for non-football fans is 0.117. Indicating that football fans were happier on average when watching the football stimulus when compared to non-football fans. Additionally, happiness was the depicted as the highest mean intensity for both football and non-football fans. Conversely, the emotional element of surprise was elicited less among football fans when compared to non-football fans overall. To gain a holistic overview of the participants affective state a phase-by-phase analysis was undertaken.

Affective shift analysis of football fans and non-football fans

The viral video has a time duration of 1:31 which was divided into three equidistant phases (thresholds):

- Phase 1 (0:00 – 0:30)
- Phase 2 (0:31 – 1:00)
- Phase 3 (1:01+).

The aggregated threshold is represented below:

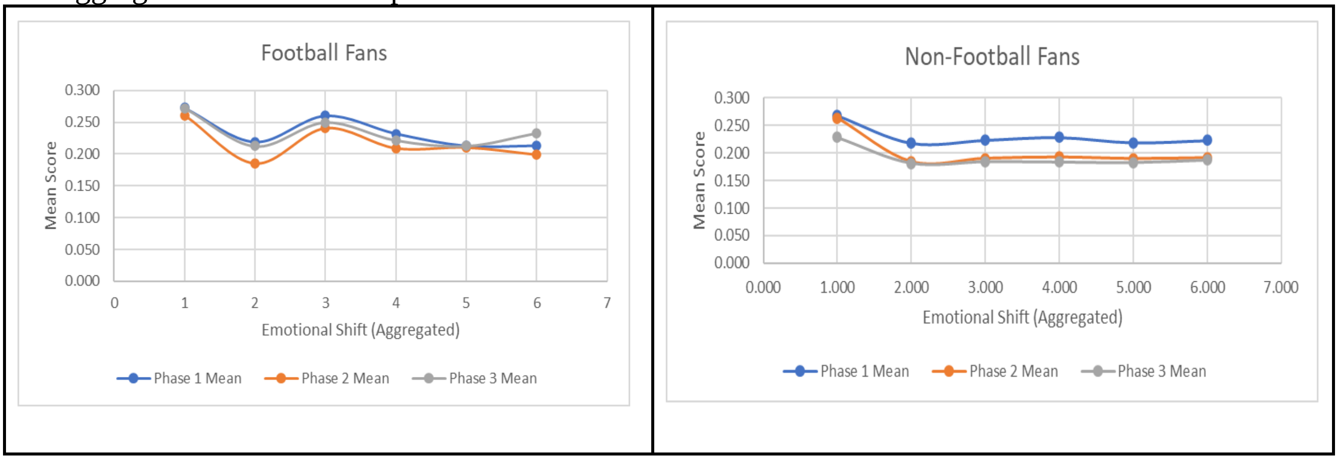
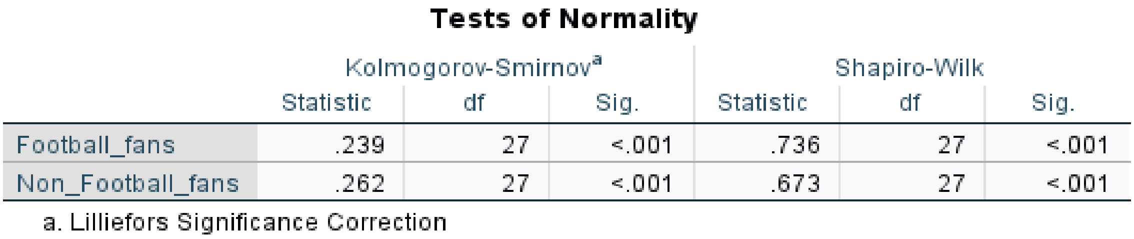


Figure 3. **Affective shifts (from phase 1 to 3) aggregated overview**

There are significant reasons for dividing a video stimulus into three thresholds. A major factor for depicting a video stimulus in three equidistant phases is that it might provide a comprehensive analysis of emotional responses over time. Furthermore, this approach might allow for temporal analysis, highlighting how emotions can evolve throughout the video. Consequently, it will help identify central tendencies and variability, revealing whether emotions are consistent or fluctuate significantly. Comparative insights can be drawn by visually assessing different segments, while outliers and extreme responses are easily detected. The break down of the mean emotional data into 3 equidistant phases can be found within this figshare repository.

Test for normality

Assessing for normality is crucial when working with bell curves (Mcleod, 2023). The Shapiro-Wilk is one of the tests that can evaluate whether a dataset follows a normal distribution (Ghasemi and Zahediasi, 2012; McLead, 2023). To establish a basis the happiness emotional response data was leveraged for both football and non-football fans. The results indicate that using a Shapiro Wilks indicates that the data evaluated was **not normally distributed** for both football and non-football fans using **happiness** as the base.



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Football_fans	.239	27	<.001	.736	27	<.001
Non_Football_fans	.262	27	<.001	.673	27	<.001

a. Lilliefors Significance Correction

Table 1. **Test for normality**

While the data may not strictly follow a normal distribution, using a bell curve can be an approximation or simplification (Sartori, 2006). It can be argued that there are no quantities in nature that are distributed normally, however, there are many quantities in nature with frequency distributions that are very well-approximated by normal distributions (Lyon, 2013). Human emotions elicited are no exceptions to the phenomena. Additionally, the normal distribution is well-known, and its properties are widely understood as it pertains to the central limit theorem which states that the distribution of sample means (from any population) tends to follow a normal distribution as the sample size increases (Sartori, 2006; Montgomery et al., 2014; Varshney, 2021).

Interpreting the affective shift phase with the Bell curve

The premise for this study posits that virality can be stimulated when people elicit positive emotions such as surprise or happiness (Berger and Milkman, 2012; Nelson-Field and Reibe, 2013). The stimulation can be depicted by football fans and non-football fans eliciting emotions using a symmetrical distribution such as the bell-shaped curve. The bell curve is used in various fields to model real world phenomena (Peterson, 2012). A bell-shaped curve represents a graph where the data clusters around the mean, with the highest frequency in the centre, and decreases gradually towards the tails (McLeod, 2023).

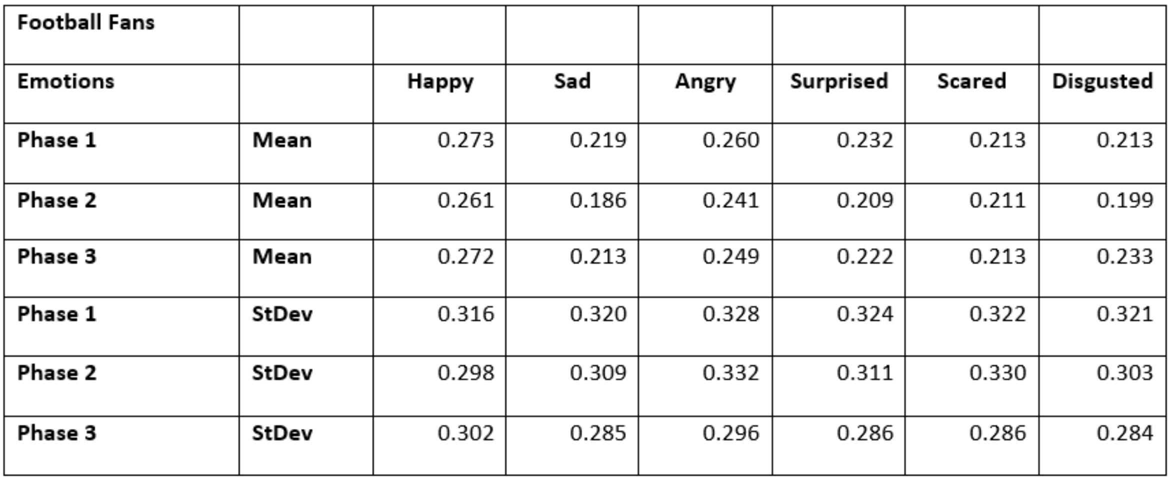
The argument for using a bell curve to model the emotions elicited among football and non-football fans can stem from the following characteristics which would be used as the basis for evaluation:

- Symmetry (If the data is symmetrically distributed, a bell curve can effectively represent the balance of emotions, showing that extreme emotions (both positive and negative) are less common than moderate ones) (Sartori, 2006; Peterson, 2012).
- Central tendency (The bell curve highlights the central tendency of the data, showing where most responses lie. This can help identify the most intense emotional responses) (Sartori, 2006; Peterson, 2012).
- Outlier identification (The tails of the bell curve can help identify outliers or extreme emotional responses, which might be critical for understanding unique or rare reactions and finally (Peterson, 2012).
- Comparative Analysis: It facilitates comparison between different groups or conditions by providing a standardised way to visualise and compare emotional responses (Neurolaunch, 2024). In this context this will be football fans and non-football fans.

- Statistical Inferences: Using a bell curve allows for the application of various statistical tests and inferences, such as confidence intervals and hypothesis testing, which can provide deeper insights into the data (Sartori, 2006).

This research delineates the importance of categorising the characteristics of a viral video based on identifying a person's identity during campaign launch. There are quite a few categorisation models that can be harnessed for classifications such as logistic regression, decision trees or K-Nearest Neighbour (KNN) (Varghese, 2018). However, this study used a distribution bell curve (Gaussian) approach due to its key characteristics such as the use of symmetry, bell shape, mean and standard deviation harnessed from emotions which have been identified prior.

It is worth indicating that standard normal distribution (or bell curve) is primarily used for modelling continuous, numerical data and understanding the distribution of values within a dataset (Riffenburgh, 2012). It's not typically used for categorisation tasks where the goal is to assign discrete labels or categories to data points. However, in some cases, the concept of normal distribution can indirectly relate to categorisation tasks (Brownlee, 2020). A specific example is where a researcher might use a threshold with a normal distribution to create binary categories. For example, if there is data on the heights of individuals and one wants to categorise them as "tall" or "not tall," the researcher could set a threshold based on the mean and standard deviation (Brownlee, 2020; Sun et al.,2022). In this research study, football fans and non-football fans categorisation was adopted using a bell curve distribution. As observed in **figure 4** below the phases which depict the level of emotional intensity can be characterised by using the **bell curve distribution.** The bell curve's familiarity ensures ease of interpretation for stakeholders, and the empirical rule can be applied to understand the intensity and spread of emotions (Sartori, 2006).



Football Fans							
Emotions		**Happy**	**Sad**	**Angry**	**Surprised**	**Scared**	**Disgusted**
Phase 1	**Mean**	0.273	0.219	0.260	0.232	0.213	0.213
Phase 2	**Mean**	0.261	0.186	0.241	0.209	0.211	0.199
Phase 3	**Mean**	0.272	0.213	0.249	0.222	0.213	0.233
Phase 1	**StDev**	0.316	0.320	0.328	0.324	0.322	0.321
Phase 2	**StDev**	0.298	0.309	0.332	0.311	0.330	0.303
Phase 3	**StDev**	0.302	0.285	0.296	0.286	0.286	0.284

Table 2. The data above depicts the emotions elicited during the different phases by football fans ("in group") when watching the viral football video. The mean elicited for "happiness" and "surprise" (Positive emotions) is used for evaluation.

The affective shifts of participants can be evaluated from the distinctive patterns which are characterised by using the **bell curve** distribution as observed below:

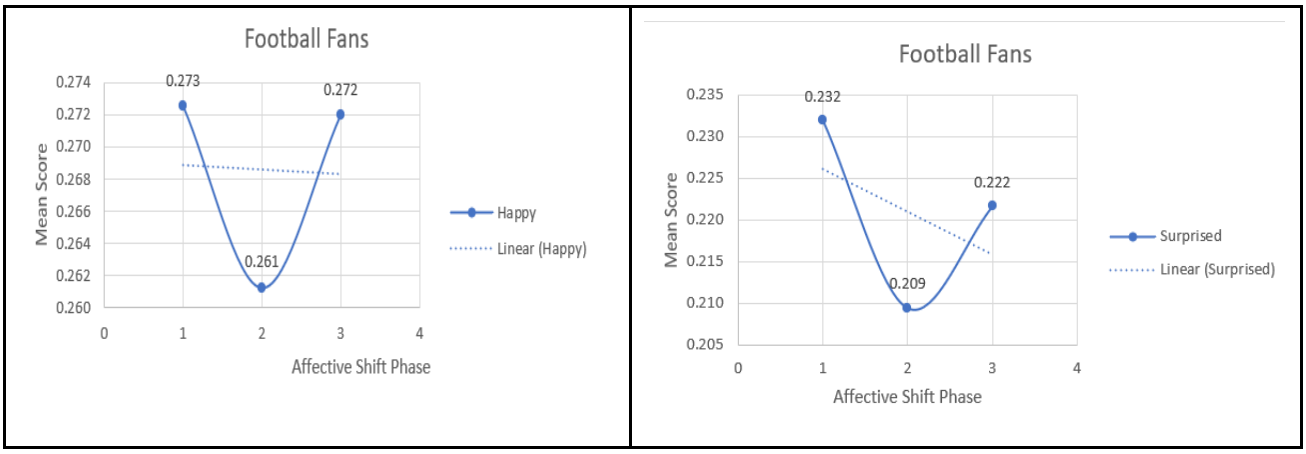


Figure 4 The emotional shift patterns for football fans ("in group") who elicited happiness and surprise when watching the viral football video. The mean elicited for "happiness" and "surprise" are used for evaluation.

Whereas the statistical breakdown of the affective patterns for football fans can be broken down in the following:

Symmetry (Happiness): It can be observed that the football fans who elicited happiness when watching the viral video produced a **perfect downward bell curve**. In this instance the downward bell curve represents the events from the moment the football fan starts feeling an emotion (i.e phase 1) activation , (i.e phase 2) deactivation and reactivation (i.e phase 3) thereby capturing the trajectory of emotional intensity over the duration. This further indicates that the data is symmetrically distributed using a bell curve as it effectively represents the balance of emotions.	**Symmetry (Surprise):** It can be observed that the football fans who elicited surprise when watching the viral video produced a **near downward bell curve** with the third phase not reaching peak intensity.

Central tendency (Happiness): Mean: 0.2687 Median: 0.272 Mode: 0.272 These measures indicate that the central tendency of the dataset is around 0.2687 (mean), with the median and mode both being 0.272. This suggests that the values are quite close to each other, indicating a consistent central tendency.	**Central tendency (Surprise):** Mean: 0.221 Median: 0.222 Mode: 0

Mean: The mean value of 0.2687 suggests the average level of happiness reported across all phases. This gives a general idea of the overall emotional state of the participants. Since the mean is close to the individual values, it indicates that the emotional responses were relatively consistent. **Median:** The median value of 0.272 represents the middle point of the data when ordered. This means that half of the emotional responses were below 0.272 and half were above. The median being close to the mean indicates that the data is symmetrically distributed without extreme outliers. **Mode:** The mode value of 0.272 is the most frequently occurring value in the dataset. This suggests that 0.272 was a common emotional response among the participants, indicating a typical level of happiness experienced during the phases. **Outlier identification:** Q1 (25th percentile): 0.2665 Q3 (75th percentile): 0.2725 IQR (Interquartile range): 0.0060 Lower bound for outliers: 0.2575 Upper bound for outliers: 0.2815 **Outliers:** There are no outliers in the data set, as all values fall within the range of **0.2575 to 0.2815**. The absence of outliers means that there were no extreme emotional responses that could skew the data, reinforcing the consistency of the emotional experiences.	**Mean:** The mean value of 0.221 suggests the average level of happiness reported across all phases. This gives a general idea of the overall emotional state of the participants. Since the mean is close to the individual values, it indicates that the emotional responses were relatively consistent. **Median:** The median value of 0.222 represents the middle point of the data when ordered. This means that half of the emotional responses were below 0.222 and half were above. The median being close to the mean indicates that the data is symmetrically distributed without extreme outliers. **Mode:** The mode cannot be calculated as there are no repeating values in the dataset. **Outlier identification:** Q1 (25th percentile): 0.2155 Q3 (75th percentile): 0.227 IQR (Interquartile range): 0.0115 Lower bound for outliers: 0.19825 Upper bound for outliers: 0.24425 **Outliers:** There are no outliers in your dataset, as all values fall within the range of **0.19825 to 0.24425**. The absence of outliers means that there were no extreme emotional responses that could skew the data, reinforcing the consistency of the emotional experiences.

Table 3. The data above depicts the depicts the symmetry, central tendency and outliers of football fans who elicited happiness and surprise.

Non-Football fans

Conversely, for each phase the following emotional intensities were attributed for non-football fans:

Non - Football Fans							
Emotions		Happy	Sad	Angry	Surprised	Scared	Disgusted
Phase 1	Mean	0.267	0.218	0.223	0.228	0.219	0.223
Phase 2	Mean	0.263	0.184	0.190	0.193	0.190	0.191
Phase 3	Mean	0.228	0.181	0.184	0.183	0.183	0.187
Phase 1	StDev	0.296	0.317	0.314	0.311	0.317	0.314
Phase 2	StDev	0.289	0.292	0.288	0.287	0.290	0.288
Phase 3	StDev	0.264	0.280	0.279	0.279	0.280	0.277

Table 4. The data above depicts the emotions elicited during the different phases by non-football fans ("out group") when watching the viral football video. The mean elicited for "happiness" and "surprise" (Positive emotions) is used for evaluation.

The affective shifts of participants can be evaluated from the distinctive patterns which are characterised by using the **bell curve** distribution as observed in **figure 5** below:

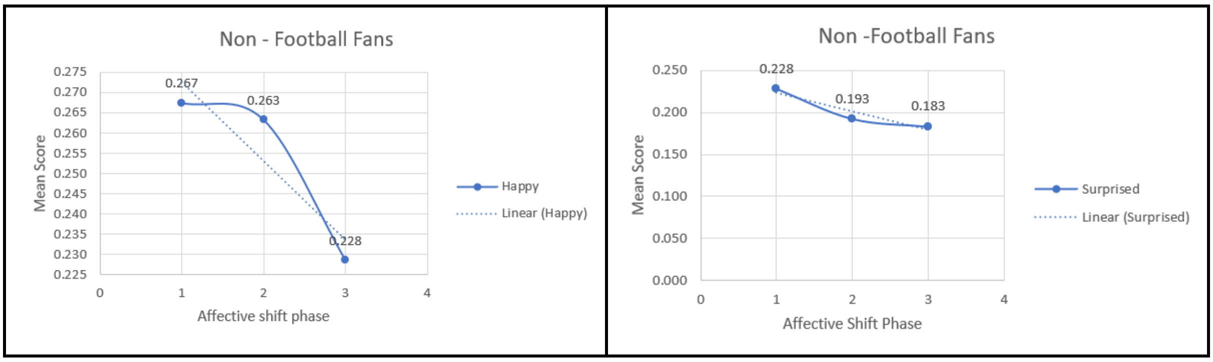


Figure 5 The emotional shift patterns for non-football fans ("out group") who elicited happiness and surprise when watching the viral football video. The mean elicited for "happiness" and "surprise" are used for evaluation.

Whereas the statistical breakdown of the affective patterns for non-football fans can be broken down in the following:

Symmetry(Happiness): These measures suggest that the emotional responses for happiness are relatively close to each other, with a slight decrease in Phase 3. This could indicate a minor drop in happiness during Phase 3 compared to the other phases.	**Symmetry(Surprise):** These measures suggest that the emotional responses for surprise are relatively close to each other, with no extreme values indicating any outliers.

Central Tendency Measures (Happiness):	**Central tendency (Surprise):**

Mean: 0.2527 Median: 0.263 Mode couldn’t be calculated as there are no repeating values in the dataset. **Mean:** The average happiness level across all phases is 0.2527, indicating a general trend of emotional responses. **Median:** The middle value is 0.263, suggesting that half of the responses are below this value and half are above. **Mode:** Since there are no repeating values, there’s no most frequent emotional response. **Outlier Identification:** Q1 (25th percentile): 0.2455 Q3 (75th percentile): 0.265 IQR (Interquartile Range): 0.0195 Lower bound for outliers: 0.2163 Upper bound for outliers: 0.2943. There are no outliers in your dataset, as all values fall within the range of 0.2163 to 0.2943.	Mean (Average): 0.201 Median: 0.193 Mode: No mode (since all values are unique) **Mean (Average):** The mean happiness score of 0.201 suggests a general level of happiness across all phases. This average indicates that, overall, the emotional state is relatively positive but not extremely high. It reflects a consistent, moderate level of happiness. **Median:** The median score of 0.193 represents the middle value when all happiness scores are arranged in order. This means that half of the phases have happiness scores below 0.193 and half above. The median is useful for understanding the central point of emotional experiences, showing that the typical phase has a happiness level around 0.193. **Mode:** There is no mode in this dataset because all values are unique. In the context of emotions, this suggests that each phase has a distinct level of happiness, indicating variability in emotional experiences across different phases. **Outlier Identification:** Q1 (25th percentile): 0.2155 Q3 (75th percentile): 0.227 IQR (Interquartile Range): 0.0115 Lower bound for outliers: 0.19825 Upper bound for outliers: 0.2442 There are no outliers in your dataset, as all values fall within the range of 0.19825 to 0.24425.

Table 5. The data above depicts the depicts the symmetry, central tendency and outliers of non-football fans who elicited happiness and surprise.

Discussion

Viral videos elicit a unique progression of affective emotions. Hence, among football fans the viral video indicated that the highest emotional intensity for happiness was in phase 1(0.273) and 3 (0.272) (highest peak). This is significant as previous research has indicated that peak phases are correlated to “liking” (Baugmarter, Sujan and Padjet, 1997) and that “happiness” as an emotion is highly correlated to sharing of content (Berger and Milkman, 2012; Tellis et al.,2019). Thus, proving theoretically that the more disruptive and intense the emotional experience is, the more frequently the event is shared with social partners (Rime et al., 1992; Rime et al.,1998). More so, this suggests that the video will be more likely to be liked and shared in the first and third phase in relation to happiness elicited when the emotional intensity is at its peak. Similarly, prior studies have shown that an element of surprise is key to diffusion (Nelson-Field, Reibe and Newstead, 2013). This peak of emotional intensity was observed in the first phase and partially in the third phase possibly correlating to the key periods when the goal was scored (i.e 13s) and player celebrations.

Interestingly, for non-football fans the highest emotional intensity for happiness was in phase 1(0.267) which is significantly higher than the lowest emotional intensity encountered by football fans (0.261). In relation to surprise the highest emotional intensity (0.228) is also significantly higher than the lowest emotional intensity when compared to football fans (0.209). The phase also correlates to the key periods when the goal was scored (i.e 13s of play) but not for celebrations which will be in the third phase. Non-football fans being emotionally detached from celebrations further supports the view from (Zubernis, 2023), who argued that highly identified fans feel a strong psychological connection to their team as opposed to non-fans.

Limitations of the methodological framework

It is important to note the limitations when using bell curve distributions. In the context of emotions, a normal distribution could be suitable if the emotions follow a pattern where most observations cluster around a central value (e.g., neutral emotions) with fewer extreme values (e.g., intense happiness or extreme sadness) However, emotions are complex and may not always conform to a perfectly symmetric distribution. For example, extreme emotions (such as intense joy or profound grief) might be more common than a normal distribution predicts leading to skewness, or a lack of symmetry, between what falls above and below the mean (Bloomental, Silbersterin and Munichiello, 2023).

Having considered emotions, an important variable that was not taken into consideration within this study is the effect of mood on emotions when subjects are placed in a condition to view a video stimulus. Inherently, mood is defined as a consumer’s affective state that is relatively global in nature, as opposed to emotions, which tend to have a specific cause (Gardner, 1985). Baggozi, Gopinath and Nyer (1999) posits that the line between emotions and moods is difficult to draw but often by convention involves conceiving of a mood as being longer lasting (from a few hours up to days) and lower in intensity than emotion and yet still another distinction is that emotions are intentional whereas moods are generally nonintentional. Mitchell (2021) indicated that people oscillate between a range of different emotions across a period. Further strengthening the phenomenon of ‘affective shifts’ since people are ‘shifting’ between different affective states as in this context when people were watching the video stimuli.

Additionally, gender was not factored in the framework as an additional variable. Within the context of social identity and emotions the variable could have provided a more extensive overview of the affective states between male football fans and non-football fans and female football fans and female non-football fans. However, the sample of participants who considered themselves as female was rather limited and will have implications to the studies validity and generalisability (Faber and Fonseca, 2014).

Future work

The subsequent phase of the study will be tested for **“robustness”**. Vander et al. (2001) defined robustness as the ability to reproduce the (analytical) method in different laboratories or under different circumstances without the occurrence of unexpected differences in the obtained results. The expectation will be to observe similar affective state patterns for surprise and happiness among football fan and non-football fans when evaluating a different set of video stimuli. Within the study an additional viral video and two non-viral videos stimuli were used for the study but were not included in the initial analysis.

The final phase will be to use a **Cauchy distribution** from the data set. The Cauchy distribution (also known as the Lorentzian distribution) has heavier tails than the normal distribution. It is characterised by its long tails, which allow for extreme values (Alzaatreh et al., 2016). In the context of emotions, the Cauchy distribution could be more appropriate if emotions exhibit significant variability and outliers. For instance, if certain stimuli evoke extreme emotional responses in some individuals, the Cauchy distribution might capture this better. A comparative analysis between Cauchy and bell curve distribution will offer a methodological insight as to which distribution model could be a better fit for understanding the variations of emotions elicited by people watching a viral stimulus.

Practical implications

From a practical perspective, the research is relevant for those interested in the development of viral videos such as those used in sports. Although the results must be confirmed and replicated in real-world settings and industries, it is experimentally evident that developing videos with different affective shifts and emotional intensity can systematically affect engagement such as sharing and views. By identifying and analysing the affective states of videos it will be possible to crop or streamline videos where the affective state is most effective in relation to a video content going viral. This comes particularly useful in the creation of “micro videos” or “shorts/Reel” videos. Micro videos are short, engaging clips designed to capture attention quickly and are used in social media platforms such as YouTube, Instagram and Tik Tok (sahu, 2014).

However, the most significant aspect of the study will lean into the incorporation of machine learning into the analysis of fan responses to viral football content which will further add a potent tool for marketers and producers. By leveraging data sets from both football fans and non-fans, machine learning algorithms can discern patterns and nuances in emotional reactions that might elude human observation(i.e distribution bell curve and Cauchy distribution).These algorithms will be able to analyse vast quantities of emotional data from FaceReader which can help identify which aspects of a video trigger the most intense emotional responses, such as a goal celebration or a display of sportsmanship. It can also differentiate between the subtleties of a fan’s elation or a non-fan’s admiration. By training models with this data, content producers can predict future content performance and tailor their videos to amplify desired emotional responses. Marketers can use these insights to segment their audience more effectively, personalising content to fan categories, enhancing user experience, and increasing engagement.

Conclusive Summary

The study of emotions and fan categorisation is pivotal for marketers and producers within the video content industry. By understanding the intricate emotional tapestries woven by football fans and non-fans alike, content creators can tailor their narratives to resonate more deeply with diverse audiences. For marketers, recognizing the intensity of fans' emotional investments allows for the crafting of targeted campaigns that leverage loyalty and passion, turning every piece of content into a touchpoint for engagement and brand loyalty reinforcement. Simultaneously, acknowledging the broader appeal that football holds for non-fans—often driven by universal themes of human drama and achievement—opens avenues for inclusive content strategies that appeal to a wider demographic, capitalising on the shared human experience.

References

Alzaatreh, A., Lee, C., Famoye, F. et al. The generalized Cauchy family of distributions with applications. J Stat Distrib App 3, 12 (2016). https://doi.org/10.1186/s40488-016-0050-3.

Annamalai, B., Yoshida, M., Varshney, S., Pathak, A.A. and Venugopal, P., 2021. Social media content strategy for sport clubs to drive fan engagement. Journal of retailing and consumer services, 62, p.102648.

An, S., Ji, L.-J., Marks, M. and Zhang, Z., 2017. Two sides of emotion: Exploring positivity and negativity in six basic emotions across cultures. *Frontiers in Psychology*, 8, p.610.

Asamoah, Joseph. "Measuring user emotionality on online videos: A comparison between self-report and facial expression analysis.". UKAIS (2019).

Bagozzi, R. P., Gopinath, M., & Nyer, P. U. (1999). The Role of Emotions in Marketing. Journal of the Academy of Marketing Science, 27(2), 184-206. https://doi.org/10.1177/0092070399272005.

Bandyopadhyay, K. (2024) 'Introduction: perspectives on fans and identities in soccer', *Soccer & Society*, 25(4–6), pp. 385–396. doi: 10.1080/14660970.2024.2342167.

Baumgartner, H., Sujan, M. and Padgett, D., 1997. Patterns of affective reactions to advertisements: The integration of moment-to-moment responses into overall judgments. Journal of Marketing Research, 34(2), pp.219-232.

Benta, K., Kuilenburg, H., Xolocotzin Eligio, U. and den Uy, M. (2009). Evaluation of a System for Real-Time Valence Assessment of Spontaneous Facial Expressions. Cluj-Napoca: International Romanian – French Workshop.

Berger, J. (2011). Arousal Increases Social Transmission of Information. *Psychological Science*, 22(7), pp.891-893.

Berger, J. and Milkman, K.L., 2012. What makes online content viral?. *Journal of marketing research*, *49*(2), pp.192-205.

Biscaia, R., Correia, A., Rosado, A., Maroco, J. and Ross, S., 2012. The effects of emotions on football spectators' satisfaction and behavioural intentions. *European Sport Management Quarterly*, *12*(3), pp.227-242.

Brownlee, J. (2020) *4 types of classification tasks in machine learning*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/types-of-classification-in-machine-learning/ (Accessed: 20 September 2023).

Brownlee, J. (2020) *4 types of classification tasks in machine learning*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/types-of-classification-in-machine-learning/ (Accessed: 20 September 2023).

Bloomenthal, A., Silberstein , S. and Munichiello, K. (2023) *Bell curve definition: Normal distribution meaning example in finance*, *Investopedia*. Available at: https://www.investopedia.com/terms/b/bell-curve.asp (Accessed: 18 April 2024).

Choi, C.(2022). *The Effect of Emotional Intensity, Arousal, and Valence On Online Video Ad Sharing.* (Doctoral dissertation). Retrieved from https://scholarcommons.sc.edu/etd/6656.

Doidge, M., Kossakowski, R. and Mintert, S., 2020. It's only a game?: Centralising emotions in football fandom. In *Ultras* (pp. 43-67). Manchester University Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4), pp.169-200

Faber J, Fonseca LM. How sample size influences research outcomes. Dental Press J Orthod. 2014 Jul-Aug;19(4):27-9. doi: 10.1590/2176-9451.19.4.027-029.ebo. PMID: 25279518; PMCID: PMC4296634.

Farnsworth, B. et al. (2023) Facial expression analysis: The Complete Pocket Guide, iMotions. Available at: https://imotions.com/blog/learning/best-practice/facial-expression-analysis/ (Accessed: 20 September 2023).

France, S.L., Vaghefi, M.S. and Zhao, H., 2016. Characterizing viral videos: Methodology and applications. Electronic Commerce Research and Applications, 19, pp.19-32.

Friedrich M. Götz, Stefan Stieger, Tobias Ebert, Peter J. Rentfrow, David Lewetz; What Drives Our Emotions When We Watch Sporting Events? An ESM Study on the Affective Experience of German Spectators During the 2018 FIFA World Cup. Collabra: Psychology 1 January 2020; 6 (1): 15. doi:

Friggeri, A., Adamic, L., Eckles, D. and Cheng, J., 2014, May. Rumor cascades. In *proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 101-110).

Gardner, M. P. (1985). Mood states and consumer behavior: A critical review. *Journal of Consumer Research*, *12*(3), 281–300. https://doi.org/10.1086/208516

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, *10*(2), 486–489. https://doi.org/10.5812/ijem.3505.

Goldenberg, A., 2024. What makes groups emotional?. *Perspectives on Psychological Science*, *19*(2), pp.489-502.

Hayes, J. L., Shan, Y. and King, K. W. (2017) 'The interconnected role of strength of brand and interpersonal relationships and user comment valence on brand video sharing behaviour', International Journal of Advertising, 37(1), pp. 142–164. doi: 10.1080/02650487.2017.1360576.

Hirshon, N. (2020). Social Identity Theory in Sports Fandom Research. In R. Dunn (Ed.), Multidisciplinary Perspectives on Media Fandom (pp. 172-191). IGI Global. https://doi.org/10.4018/978-1-7998-3323-9.ch010

Hoffman, L. H., Baker, A., Beer, A., Rome, M., Stahmer, A., & Zucker, G. (2020). Going viral: Individual-level predictors of viral behaviours in two types of campaigns. Journal of Information Technology & Politics, 18(2), 117–124. https://doi.org/10.1080/19331681.2020.1814930.

Islam, G. (2014). Social Identity Theory. In: Teo, T. (eds) Encyclopedia of Critical Psychology. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-5583-7_289.

Jiang, L., Miao, Y., Yang, Y., Lan, Z. and Hauptmann, A.G., 2014, April. Viral video style: A closer look at viral videos on youtube. In *Proceedings of International Conference on Multimedia Retrieval* (pp. 193-200)

John P. Lehoczky, Distributions, Statistical: Special and Continuous,
Editor(s): James D. Wright, International Encyclopedia of the Social & Behavioral Sciences (Second Edition), Elsevier, 2015, Pages 575-579, https://doi.org/10.1016/B978-0-08-097086-8.42115-X.
(https://www.sciencedirect.com/science/article/pii/B978008097086842115X)

King, P.S., 2013. Emotions: Positive and negative. In: R.J.R. Levesque, ed., *Encyclopedia of Adolescence*. New York: Springer, pp. 1000-1008.

Kong, Q., Rizoiu, M.A., Wu, S. and Xie, L., 2018, April. Will this video go viral: Explaining and predicting the popularity of youtube videos. In *Companion Proceedings of The Web Conference 2018* (pp. 175-178).

Lewinski, P., Den Uyl, T.M. and Butler, C., 2014. Automated facial coding: validation of basic emotions and FACS AUs in FaceReader. *Journal of neuroscience, psychology, and economics*, *7*(4), p.227.

Li, K., Shen, X., Chen, Z., He, L. and Liu, Z., 2020, August. Effectiveness of Emotion Eliciting of Video Clips: A Self-report Study. In The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (pp. 523-542). Cham: Springer International Publishing.

Lyon, Aidan. (2013). Why are Normal Distributions Normal?. The British Journal for the Philosophy of Science. 65. 621-649. 10.1093/bjps/axs046.

Ma, CX., Song, JC., Zhu, Q. *et al.* EmotionMap: Visual Analysis of Video Emotional Content on a Map. *J. Comput. Sci. Technol.* **35**, 576–591 (2020). https://doi.org/10.1007/s11390-020-0271-2.

Ma, I. (2023) Feeling intensely: The wounds of being 'Too much', Psychology Today. Available at: https://www.psychologytoday.com/us/blog/living-with-emotional-intensity/201805/feeling-intensely-the-wounds-of-being-too-much?msockid=281c2ab8dff261eb35a33edade8860fa (Accessed: 17 October 2024).

Mastromartino, B., Chou, W.H.W. and Zhang, J.J., 2018. The passion that unites us all: the culture and consumption of sports fans. In *Exploring the rise of fandom in contemporary consumer culture* (pp. 52-70). IGI Global.

Mitchell, J. Affective shifts: mood, emotion and well-being. Synthese 199, 11793–11820 (2021). https://doi.org/10.1007/s11229-021-03312-3

Mcleod, S. (2023) *Bell shaped curve: Normal distribution in statistics*, *Simply Psychology*. Available at: https://www.simplypsychology.org/normal-distribution.html (Accessed: 18 April 2024).

Mcleod, S. (2023) *Social Identity theory in Psychology*, *Simply Psychology*. Available at: https://www.simplypsychology.org/social-identity-theory.htm (Accessed: 18 Octoberl 2024).

Montgomery, Douglas C.; Runger, George C. (2014). Applied Statistics and Probability for Engineers (6th ed.). Wiley. p. 241. ISBN 9781118539712.

Nelson-Field, K., Riebe, E. and Newstead, K., 2013. The emotions that drive viral video. *Australasian Marketing Journal*, *21*(4), pp.205-211.

NeuroLaunch (2024) Fan Behavior: Understanding the Psychology and impact of sports enthusiasts. Available at: https://neurolaunch.com/fan-behavior/(Accessed: 18 October 2024).

Nikolinakou, A. and King, K. W. (2018) 'Viral Video Ads: Examining Motivation Triggers to Sharing', *Journal of Current Issues & Research in Advertising*, 39(2), pp. 120–139. doi: 10.1080/10641734.2018.1428247.

Noel, J. G., Wann, D. L., & Branscombe, N. R. (1995). Peripheral ingroup membership status and public negativity toward outgroups. Journal of Personality and Social Psychology, 68, 127–137.

Otamendi, F.J. and Sutil Martín, D.L., 2020. The emotional effectiveness of advertisement. Frontiers in Psychology, 11, p.2088.

Pivecka, N., Ratzinger, R.A. and Florack, A., 2022. Emotions and virality: Social transmission of political messages on Twitter. Frontiers in psychology, 13, p.931921

Peterson, C. (2012) For whom the Bell holds, Psychology Today. Available at: https://www.psychologytoday.com/us/blog/the-good-life/201206/for-whom-the-bell-holds (Accessed: 18 October 2024).

Poels, K. and Dewitte, S. (2019) ‘The Role of Emotions in Advertising: A Call to Action’, Journal of Advertising, 48(1), pp. 81–90. doi: 10.1080/00913367.2019.1579688.

Richardson, D.C., Griffin, N.K., Zaki, L., Stephenson, A., Yan, J., Curry, T., Noble, R., Hogan, J., Skipper, J.I. and Devlin, J.T., 2020. Engagement in video and audio narratives: Contrasting self-report and physiological measures. Scientific Reports, 10(1), p.11298.

Riffenburgh, R.H. Chapter 6 - Finding Probabilities, Editor(s): R.H. Riffenburgh, Statistics in Medicine (Third Edition), Academic Press, 2012, Pages 117-136,ISBN 9780123848642, https://doi.org/10.1016/B978-0-12-384864-2.00006-8. (https://www.sciencedirect.com/science/article/pii/B9780123848642000068).

Rimé B., Philippot P., Boca S., Mesquita B. (1992). Long-lasting cognitive and social consequences of emotion: Social sharing and rumination. European review of social psychology, 3(1), 225–258. https://doi.org/10.1080/14792779243000078.

Rimé, Bernard & Finkenauer, Catrin & Luminet, Olivier & Zech, Emmanuelle & Philippot, Pierre. (1998). Social Sharing of Emotion: New Evidence and New Questions. European Review of Social Psychology - EUR REV SOC PSYCHOL. 9. 145-189. 10.1080/14792779843000072.

Sidle, R. (2022) *Never before seen footage shows new angle of Manchester United winning the Treble, SPORTbible*. Available at: https://www.sportbible.com/football/new-footage-emerges-of-manchester-united-winning-the-treble-20220526 (Accessed: 27 February 2025).

Sahu, N. (2024). Micro Videos: What, Why, Where, and Best Examples. [online] Www.b2w.tv. Available at: https://www.b2w.tv/blog/micro-videos.

Santos, M.J.S.D.A.P., 2018. *Going viral: the influence of emotional content and gender on social transmission* (Doctoral dissertation).

Sartori, R. The Bell Curve in Psychological Research and Practice: Myth or Reality?. Qual Quant 40, 407–418 (2006). https://doi.org/10.1007/s11135-005-6104-0

Singh, R., Mangat, N.S. (1996). Stratified Sampling. In: Elements of Survey Sampling. Kluwer Texts in the Mathematical Sciences, vol 15. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-1404-4_5.

Shakina, E., Gasparetto, T. and Barajas, A., 2020. Football fans’ emotions: Uncertainty against brand perception. Frontiers in psychology, 11, p.659.

Skiendziel, T., Rösch, A. G., and Schultheiss, O. C. (2019). Assessing the convergent validity between the automated emotion recognition software Noldus FaceReader 7 and facial action coding system scoring. *PLoS One* 14:e0223905. doi: 10.1371/journal.pone.0223905

Suhr, Y.T., 2017. *FaceReader, a promising instrument for measuring facial emotion expression? A comparison to facial electromyography and self-reports* (Master's thesis).

Sun, Y., Ming, Y., Zhu, X. and Li, Y., 2022, June. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning* (pp. 20827-20840). PMLR

Susannah B. F. Paletz et al. ,Emotional content and sharing on Facebook: A theory cage match.Sci. Adv.9,eade9231(2023). DOI:10.1126/sciadv.ade9231

Szanto, T. and Krueger, J., 2017. Introduction: Empathy, Shared Emotions, and Social Identity. Topoi, 36(2), pp.169-178.

Stets, J.E. and Turner, J.H., 2014. Emotions in Identity Theory. In: J.E. Stets and J.H. Turner, ed., Handbook of the Sociology of Emotions: Volume II, Cham: Springer International Publishing, pp.291-316.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. Organizational identity: A reader, 56-65.

Tellis, G. J., MacInnis, D. J., Tirunillai, S., & Zhang, Y. (2019). What Drives Virality (Sharing) of Online Digital Content? The Critical Role of Information, Emotion, and Brand Prominence. Journal of Marketing, 83(4), 1-20. https://doi.org/10.1177/0022242919841034.

Terzis, V., Moridis, C.N. and Economides, A.A., 2013. Measuring instant emotions based on facial expressions during computer-based assessment. *Personal and ubiquitous computing*, *17*, pp.43-52.

Tong, L.C., Acikalin, M.Y., Genevsky, A., Shiv, B. and Knutson, B., 2020. Brain activity forecasts video engagement in an internet attention market. *Proceedings of the National Academy of Sciences*, *117*(12), pp.6936-6941

Trzciński, T. and Rokita, P., 2017. Predicting popularity of online videos using support vector regression. *IEEE Transactions on Multimedia*, *19*(11), pp.2561-2570.

Vander Heyden, Y., Nijhuis, A., Smeyers-Verbeke, J., Vandeginste, B.G.M. and Massart, D.L., 2001. Guidance for robustness/ruggedness tests in method validation. Journal of pharmaceutical and biomedical analysis, 24(5-6), pp.723-753.

Varghese, D. (2019) *Comparative study on classic machine learning algorithms*, *Medium*. Available:https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222 (Accessed: 17 April 2024).

Varshney, P. (2021) *The powers of 'normal distribution'*, *LinkedIn*. Available at: https://www.linkedin.com/pulse/powers-normal-distribution-paras-varshney/ (Accessed: 10 May 2024).

Wen, T. J. et al. (2021) 'Empowering Emotion: The Driving Force of Share and Purchase Intentions in Viral Advertising', Journal of Current Issues & Research in Advertising, 43(1), pp. 47–67. doi: 10.1080/10641734.2021.1937408.

Zubernis , L. (2023) Why Football Fans Get So Emotional, Psychology Today. Available at: https://www.psychologytoday.com/gb/blog/the-science-of-fandom/202302/why-football-fans-get-so-emotional (Accessed: 19 September 2023).

THE IMPACT OF SOCIAL MEDIA FATIGUE ON SWITCHING INTENTIONS: INSIGHTS FROM AN INFORMATION-BASED APPROACH

Fulya Acikgoz (University of Sussex) and Abdelsalam Busalim (Technological University Dublin)

Research In progress

Abstract

Social media fatigue has been widely studied, primarily from psychological and behavioral perspectives. However, the role of information quality remains underexplored. Drawing upon the Information Adoption Model, we propose a conceptual model to allow explore how key information-related factors including relevance, timeliness, accuracy, and comprehensiveness, along with source expertise and trustworthiness, contribute to social media fatigue. Additionally, we investigate how social media fatigue influences users' intention to switch from X (formerly Twitter) to alternative platforms. We also outline the research medthod and data collection and discuss the excepted findings.. We expect this study to offers novel insights into the cognitive and behavioral impact of information exposure in social media context.

***Keywords**: Social Media Fatigue, Switching Intention, Information Adoption Model, Source Credibility*

1.0 Introduction

Social media fatigue refers to the tendency of users to disengage from platforms when overwhelmed by the excessive volume of content, interactions, and the time required to maintain online connections (Zheng & Ling, 2021) It reflects constraints on users' cognitive processing abilities, making it increasingly difficult to manage the information flow and communication demands presented on social media platforms (Baj-Rogowska, 2023; Lee et al., 2016). This phenomenon is characterized by weariness, exhaustion, apathy, and reduced enthusiasm, resulting from the continuous cognitive and emotional engagement required to keep up with social media interactions and contenet (Fan et al., 2021; Jabeen et al., 2023; Ou et al., 2023).

The excessive use of social media has intensified awareness of social media fatigue, as users navigate fluctuating cycles of engagement and disengagement (Kaur & Javaid, 2023). As a result, research on social media fatigue has gained significant momentum, particularly in marketing and information systems literature. Existing research predominantly examines social media fatigue throught psychological and behavioral lenses, focusing on factors such as fear of missing out (FOMO) (Bright & Logan, 2018; Hattingh et al., 2022; Jabeen et al., 2023; Shen et al., 2020; Talwar et al., 2019; Tandon et al., 2021), personality traits (Kim, 2022; Xiao & Mou, 2019), and privacy concerns (Bright et al., 2022; Fan et al., 2021). Additionally, research highlight information overload as a key driver of fatigue (Jiang, 2022; Sharma et al., 2023; Pang, 2021; Zhang et al., 2020). However, there remains a limited research to explored the specific role of information-related characteristics, such as information equivocality (Xie & Tsai, 2021) and information quality (Xiao et al., 2023), in contributing to social media fatigue.

Beyond psychological and behavioral factors, information and source-related factors also play a crucial role in shaping social media fatigue (Li & Sheng, 2024). Fatigue may emerge when users struggle to keep pace with real-time updates, which foster a perveived obligation to stay informed.However, not all information triggers fatigue equally. Zhang et al. (2022) suggest that information relevance may both contribute to and alleviate social media fatigue, yet its precise effects remain underexplored. Similarly, Xiao et al. (2023) emphasise that analysing information characteristics on social media platforms is crucial for understanding their role in social media fatigue and emotional dissonance. This highligts the need to examine specific dimensions of information quality such as relevance, timeliness, accuracy, and comprehensiveness in relation to fatigue.

A critical aspect of social media fatigue is user switching and attrition, as platforms face increased competition for user attention (Zhang et al., 2022). Excessive information exposure often leads to fatigue, disengagement and switching behaviors (Pang & Ruan, 2023). Jeyaraj et al. (2023) highlight that technological discontinuance and switching intention have become increasingly relevant due to the growing diverse of available social media platforms. While previous studies have examined the relationship between social media fatigue and discontinuance (Adhikari & Panda, 2020; Lee et al., 2021; Xie & Tsai, 2021), the link between social media fatigue and platform switching remains largely unexplored.

To address this gap, we investigate how information-related factors (i.e., relevance, timeliness, comprehensiveness, and accuracy) and source-related factors (source expertise and trustworthiness) on social media fatigue, using the Information Adoption Model as a theoretical lens. Despite these dimensions are well established in the context of information adoption, their role as antecedents of social media fatigue has not been systematically examined. This study extends the information adoption model by examining their influence on user disengagement and switching intentions, offering novel insights into the mechanisms driving social media fatigue. Additionally, we explore the relationship between social media fatigue and users' likelihood of switching from X (formerly Twitter) to alternative platforms. The findinsg of this study will offer practical implications for both researchers and platform developers in understanding and mitigating fatigue-induced user attrition.

2.0 Theoretical Framework

2.1. Information Adoption Model

The Information adoption model (Sussman & Siegal, 2003) provides a foundation for understanding how individuals process and adopt information within computer-mediated communication environments. Information adoption model posits that information quality and source credibility are critical determinants of information adoption, typically influencing engagement and decision-making in online environments (Cheung et al., 2008). However, while information adoption model has been widely used to study positive outcomes such as trust, purchase intention, and knowledge-sharing behaviors (e.g., Acikgoz et al., 2023; Jiang et al., 2021), its application to negative consequences like fatigue and disengagement remains underexplored.

A growing body of literature suggests that information-related factors are as influential as psychological (e.g., FOMO, anxiety) and social (e.g., privacy concerns, peer influence) factors in shaping social media fatigue (Zhang et al., 2022). However, existing studies have primarily focused on macro-level factors such as information overload (Sharma et al., 2023) or information equivocality (Xie & Tsai, 2021), without systematically analysing specific information attributes that may contribute to fatigue. This study addresses this gap by investigating four key dimensions of information quality—relevance, timeliness, accuracy, and comprehensiveness—and their impact on social media fatigue and switching behavior.

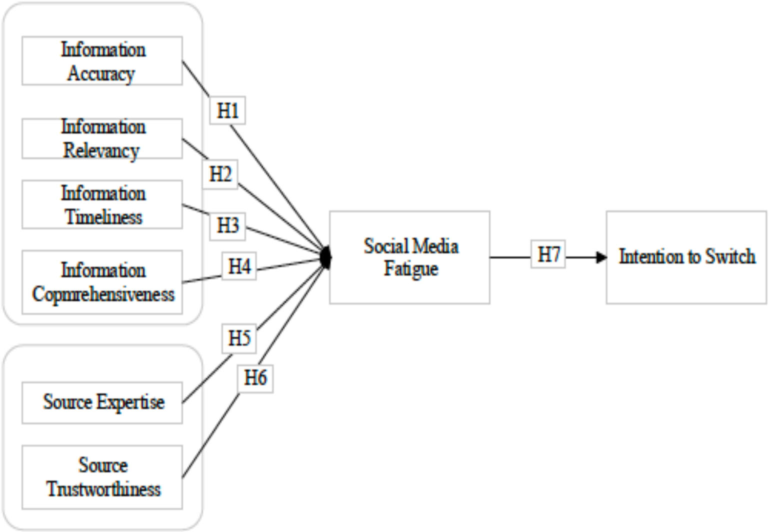


Fig.1. Conceptual Model

Information adoption model identifies argument quality as a major determinant of information adoption, typically measured through relevance, timeliness, accuracy, and comprehensiveness (Cheung et al., 2008; Zhang et al., 2022). However, these dimensions may also play a role in driving disengagement and fatigue when users encounter excessive, low-quality, or cognitively demanding information. Excessive irrelevant content increases cognitive load, leading to user fatigue as they struggle to filter useful information. (Zhang et al., 2022). Timeliness: delayed or outdated information can increase frustration and disrupt cognitive processing, leading to disengagement (Li et al., 2022). Accuracy: exposure to misinformation or contradictory facts leads to skepticism and emotional exhaustion, reducing trust in social media platforms (Pan et al., 2024). Comprehensiveness: incomplete or fragmented information requires additional effort to verify and interpret, increasing cognitive strain and contributing to fatigue (Jiang, 2022). Source credibility also plays a significant role in social media fatigue. When users perceive information sources as lacking expertise, they may feel overwhelmed by uncertain or unreliable content, which may lead to greater cognitive effort and fatigue (Cheung et al., 2008; Elwalda et al., 2021). Similarly, content from untrustworthy sources can cause doubt, skepticism, and increased vigilance, forcing users to fact-check information, which amplifies mental fatigue and disengagement (Jiang et al., 2021; Xiao et al., 2023). The need for constant verification and the emotional burden of processing unreliable information can heighten user fatigue, further driving disengagement and switching behavior. By incorporating both argument quality and source credibility, this study extends the Information adoption model beyond its traditional focus on information adoption, investigating how low-quality information and untrustworthy sources drive social media fatigue and switching behavior.

3.0. Proposed Methodology and Expected Results

We will examine the proposed research model using a survey-based approach, collecting data from X (formerly Twitter) users who are regularly exposed to a wide range of information, including political content and current events. The analysis will be conducted in two stages, leveraging both quantitative and configurational methods to provide a comprehensive understanding of social media fatigue and switching behavior. In the first stage, we will employ Partial Least Squares Structural Equation Modeling (PLS-SEM) to test the relationships between key constructs, including information relevance, timeliness, accuracy, comprehensiveness, source expertise, source trustworthiness, social media fatigue, and switching intention. PLS-SEM is chosen due to its ability to handle complex models with latent variables and its suitability for exploratory research in information systems. In the second stage, we will conduct a Fuzzy Qualitative Comparative Analysis (fsQCA) to identify specific configurations of information- and source-related factors that contribute to social media fatigue and switching intention. Unlike traditional regression-based methods, fsQCA allows for the identification of multiple causal pathways leading to similar outcomes, providing deeper insights into the combinations of conditions that drive users' decisions to disengage or switch platforms. By integrating PLS-SEM and fsQCA, this study aims to offer both a structural and configurational perspective on how information and source-related factors shape social media fatigue and switching behavior, providing new insights for both academics and platform designers.

4.0 Conclusion

The dynamic nature of social media and the rise of social media fatigue represent a valuable opportunities for research, particularly in understanding user behaviour on information-rich platforms such as X. This research contributes to the literature by proposing a model that explores the influence of information and source-related factors on social media fatigue and switching intentions. By focusing on constructs such as information relevance, timeliness, accuracy, comprehensiveness, source expertise, and trustworthiness, our research highlights the nuanced ways in which information quality and source credibility affect user fatigue. To our knowledge, this model represents one of the first attempts to explore how these information and source-based factors drive social media fatigue and switching intentions, providing valuable insights for both academic research and practical applications in the field of social media platform management.

References

Acikgoz, F., Busalim, A., Gaskin, J., & Asadi, S. (2023). An Integrated Model for Information Adoption&Trust in Mobile Social Commerce. *Journal of Computer Information Systems,* 1-23.

Adhikari, K., & Panda, R. K. (2020). Examining the role of social networking fatigue toward discontinuance intention: the multigroup effects of gender and age. *Journal of Internet Commerce*, *19*(2), 125-152.

Baj-Rogowska, A. (2023). Antecedents and outcomes of social media fatigue. *Information Technology & People, 36*(8), 226-254.

Bright, L. F., Kleiser, S. B., & Grau, S. L. (2015). Too much Facebook? An exploratory examination of social media fatigue. *Computers in Human Behavior, 44,* 148-155.

Erkan, I., & Evans, C. (2016). The influence of eWOM in social media on consumers' purchase intentions: An extended approach to information adoption. *Computers in Human Behavior, 61*, 47-55.

Fan, X., Jiang, X., Deng, N., Dong, X., & Lin, Y. (2021). Does role conflict influence discontinuous usage intentions? Privacy concerns, social media fatigue and self-esteem. *Information Technology & People, 34*(3), 1152-1174.

Hattingh, M., Dhir, A., Ractham, P., Ferraris, A., & Yahiaoui, D. (2022). Factors mediating social media-induced fear of missing out (FoMO) and social media fatigue: A comparative study among Instagram and Snapchat users. *Technological Forecasting and Social Change, 185,* 122099.

Jabeen, F., Tandon, A., Sithipolvanichgul, J., Srivastava, S., & Dhir, A. (2023). Social media-induced fear of missing out (FoMO) and social media fatigue: The role of narcissism, comparison and disclosure. *Journal of Business Research, 159*, 113693.

Jeyaraj, A., Dwivedi, Y. K., & Venkatesh, V. (2023). Intention in information systems adoption and use: Current state and research directions. *International Journal of Information Management, 73,* 102680.

Jiang, S. (2022). The roles of worry, social media information overload, and social media fatigue in hindering health fact-checking. *Social Media, Society, 8*(3), 20563051221113070.

Kaur, A., & Javaid, M. (2023). Download. Scroll. Post. Repeat. New social media apps are exhausting us. Retrieved from: https://www.washingtonpost.com/technology/2023/07/07/social-media-platforms-threads-twitter-fatigue

Lee, A. R., Son, S. M., & Kim, K. K. (2016). Information and communication technology overload and social networking service fatigue: A stress perspective. *Computers in Human Behavior, 55*, 51-61.

Li, Y., & Sheng, D. (2024). Determinants of public emergency information dissemination on social networks: A meta-analysis. *Computers in Human Behavior, 152,* 108055.

Li, K., Zhou, C., Luo, X. R., Benitez, J., & Liao, Q. (2022). Impact of information timeliness and richness on public engagement on social media during COVID-19 pandemic: An empirical investigation based on NLP and machine learning. *Decision Support Systems, 162,* 113752.

Ou, M., Zheng, H., Kim, H. K., & Chen, X. (2023). A meta-analysis of social media fatigue: Drivers and a major consequence. *Computers in Human Behavior, 140*, 107597.

Pan, T., Sun, Y., Guo, X., & Zhang, M. (2024). Unraveling the impact of infodemic stress on information and health behaviors: a double effect perspective. *Internet Research.* DOI: 10.1108/INTR-12-2023-1137.

Pang, H., & Ruan, Y. (2023). Determining influences of information irrelevance, information overload and communication overload on WeChat discontinuance intention: *The moderating role of exhaustion. Journal of Retailing and Consumer Services, 72*, 103289.

Shen, Y., Zhang, S., & Xin, T. (2020). Extrinsic academic motivation and social media fatigue: Fear of missing out and problematic social media use as mediators. *Current Psychology,* 1-7.

Talwar, S., Dhir, A., Kaur, P., Zafar, N., & Alrasheedy, M. (2019). Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior. *Journal of Retailing and Consumer Services, 51*, 72-82.

Tandon, A., Dhir, A., Talwar, S., Kaur, P., & Mäntymäki, M. (2021). Dark consequences of social media-induced fear of missing out (FoMO): Social media stalking, comparisons, and fatigue. *Technological Forecasting and Social Change, 171*, 120931.

Xie, X. Z., & Tsai, N. C. (2021). The effects of negative information-related incidents on social media discontinuance intention: Evidence from SEM and fsQCA. *Telematics and Informatics, 56,* 101503.

Zhang, Y., He, W., & Peng, L. (2022). How perceived pressure affects users' social media fatigue behavior: a case on WeChat. Journal of Computer Information Systems, 62(2), 337-348.

Zheng, H., & Ling, R. (2021). Drivers of social media fatigue: A systematic review. *Telematics and Informatics, 64,* 101696.

From action research to activist research: digital surveillance in Africa

Anand Sheombar & Oliver Kayas
HU University of Applied Sciences Utrecht, Netherlands
Liverpool Business School - Liverpool John Moores University, United Kingdom

Research In progress – developmental paper

Abstract

This paper examines the proliferation of state-enabled digital surveillance technologies in Africa, where governments exploit security narratives to justify rights-infringing monitoring. Addressing gaps in African-focused research, it identifies stakeholders—exporting/importing states, firms, citizens, media, and academia—and applies Gaventa's powercube framework to analyse power dynamics across spaces (arenas), forms (visibility), and levels (engagement). The study proposes four pressure points for resistance: reputational (naming/shaming), economic (sanctions, boycotts), political (policy advocacy), and technological (privacy tools), derived from tensions between surveillance proponents and civil rights actors. Integrating Hamdali et al.'s theory-practice impact framework bridges surveillance studies research with actionable strategies, emphasising multi-stakeholder collaboration. Case examples, such as litigation in Kenya and Nigeria, illustrate pathways for civil society to counter authoritarian control and neocolonial technology exports. The research advocates context-sensitive interventions, urging policymakers to balance security with rights and tech actors to develop ethical alternatives, ultimately fostering democratic accountability in Africa's digital landscape.

Keywords: engaged scholarship, impact, digital surveillance, activism, Africa, activist research.

1.0 Surveillance of African citizens

African citizens are increasingly subjected to surveillance, profiling, and targeting online, often in ways that infringe upon their rights. African governments frequently invoke pandemic or terrorism-related security risks to justify the expansion of their surveillance capabilities, significantly increasing their acquisition of monitoring apparatus and technologies with billions of dollars invested (Roberts et al., 2023). Surveillance has become a prominent strategy employed by African governments to restrict civic space (Roberts et al., 2023). Although numerous studies have examined illegal state surveillance in the United States, China, and Europe (Feldstein, 2019; Feldstein, 2021), there is a lack of research on the supply and utilisation of surveillance technologies in Africa, nor their impact on privacy or human rights violations on that continent (Dratwa, 2014; Jili, 2020; Roberts & Mohamed Ali, 2021; Roberts et al., 2021). State surveillance refers to any observing, listening, monitoring, or recording by a state or its agents to track citizens' movements, activities, conversations, communications or correspondence, including the recording of metadata, through digital technology (Roberts et al., 2021). The aim of this paper is to *move from observation to impactful actions to counter* states exporting and supplying surveillance technology to African countries where it is abused for human rights violations.

To study the supply of surveillance technologies to Africa, this research draws on the concepts of surveillance stakeholder groups, interests, and resistance literature to identify and analyse the actors involved (Table 1). Three main groups are identified: (1) state agencies and law enforcement authorities (LEA), (2) surveillance sector companies, and (3) citizens (Schuster et al., 2017). Wright et al. (2015) argue that consultancies that advise governments on implementing surveillance need to be included as relevant actors, as well as those who aim to regulate, oversee, or critique surveillance, such as policymakers and civil society organisations. Martin et al. (2009) propose a multi-actor framework to examine the dynamics between the surveyors, the surveilled, and others involved. They distinguish between governments and law enforcement agencies implementing surveillance, surveillance technology companies, international governmental and non-governmental agencies, and (non-human) surveillance technologies. This paper extends the debate by adding two additional actors who play a vital role in the study of surveillance technologies: the news media and academia.

There are economic benefits for surveillance sector firms exporting to willing governmental clients in Africa, sometimes stimulated by intensive marketing campaigns or proving 'battle-tested' capabilities (Cook, 2020; Feldstein, 2019; Hicks, 2022; van der Lugt, 2021). Northern-based supplier-side governments see the export of surveillance technologies as instrumental for their geopolitical interests in particular countries and for increasing the economic growth of their surveillance industry (Greitens, 2020). On the recipient side, van der Lugt (2021) argues that African states can decide what technologies to purchase and from whom to acquire them. Sometimes, attractive loans by the exporting states incentivise the acquisition of surveillance technologies. Another motive for acquiring such technologies could be, as Hillman (2021) stated, 'a vanity project' capturing votes during elections for ruling politicians introducing smart city projects. A motive that further undermines civil rights is the exertion of authoritarian control over citizens by using digital surveillance technologies (Feldstein, 2019; van der Lugt, 2021).

Stakeholder group	**Interests**
Governments of exporting states	Supplier-side: geopolitical interests via 'spyware diplomacy'.
Government (demand side country): state agencies and law enforcement authorities	Demand side: vanity projects for electoral gain, or authoritarian control over citizens
Private sector: i.e. Northern-based surveillance technologies firms	Economic benefits. Spin-off market for 'battle-tested' surveillance technologies.
Stakeholder group	**Resistance pathways against state surveillance**
Citizens	Pushing back against the erosion of digital rights. On individual level technical countermeasures.
Civil society i.e., an extension of the group interested in countering digital state surveillance.	Pushing back against the erosion of citizens' digital rights in a collective manner. Knowledge sharing of technical countermeasures and requesting policy changes for improving the protection of citizens' (digital) rights.
Media	Revealing erosion of democratic and digital rights (international and in African demand-side countries) and demanding accountability
Academia	Revealing erosion of democratic and digital rights vs complicity in the creation of these surveillance technologies (in the exporting country).

Table 1. Stakeholders involved in proliferation and countering digital surveillance.

1.1 Powercube as Analysis & Mitigation Tool for Citizens' Rights-violating Surveillance Technologies Proliferation

This research draws on the powercube concept that looks into the power dynamics between parties involved in three dimensions: (1) *spaces*, the arenas of power; (2) *forms*, the degree of visibility of power; and (3) *levels*, the places of engagement (Gaventa, 2019). The powercube is useful for advocacy campaigns for change, like here, where the aim is to counter the (illegal) mass surveillance of citizens. The powercube asks questions about what forms of power are used, what spaces of power are used, or what levels of power are used. The result of the powercube analysis of our African surveillance study is extensively discussed in another paper, (Sheombar & Klovig Skelton, 2025), and presented in Table 2.

Pressure Points for motives	**Target surveillance and resistance actors**	**Approach(es) and examples for actions**
Reputational: • surveillance for political gain • surveillance as legitimacy for state security	• Governments (both importing and exporting/supply -side) • surveillance technologies suppliers	• Government should engage in **dialogue with citizens** to build trust and confidence in digital surveillance systems to ensure transparency and accountability. For example, the creation of the Digital Rights Group in Malawi is evidence that resistance is also emergent. Public participation in the development of surveillance policies will amplify the democratic process. • Calling for the establishment of an i**ndependent watchdog** would strengthen accountability. • **Awareness building**: Civil society, especially human rights organisations: launch **awareness and education campaig**ns about the issue of digital surveillance. • **Naming and shaming i**n international forums • **Media strategies** (traditional and social media) by civil society, especially human rights defence organisations and (independent) journalists • Citizens' **documentation of repression**
Economic: • surveillance as a tool for development (e.g. smart or safe cities) • surveillance as a business opportunity	• Governments • Surveillance technologies suppliers	• Governments and civil society: find a **balance between national security and the protection of citizens' right**s • **Economic pressure campaigns** • Pressure campaigns against companies • **Protect civic space**: Human rights organisations: create archival records of surveillance cases and push for litigation against suppliers of surveillance technologies that are likely to be abused. • Corporate boycotts
Political: • surveillance as legitimacy for state security • surveillance for political gain • surveillance as diplomacy • Surveillance as neocolonialism	• Government (potentially both supplier and demand side country)	• **Advocacy for Policy Change:** Governments and policymakers need to establish clear and transparent legal frameworks that govern the use of digital surveillance for legitimate reasons • Civil society organisations can help **mapping** surveillance technologies and their **supply chains:** in Nigeria, Action Group on Free Civic Space (AGFCS) uses the diversity and expertise of their network to initiate joint action research projects. • Electoral challenges to incumbents • **Mobilise for direct action: sanctions** • **Monitor to hold actors accountable**: academics and researchers in the country must research digital and communication surveillance and its implications for citizens. • Civil society: capacity building in knowledge of surveillance, and **push for legal framework** guided by international human rights standards and principles, including the rights to privacy and freedom of expression. For example, in Nigeria, groups like SERAP use public interest litigation as a strategy for demanding surveillance contract transparency. Another example, is the Kenyan civil society took the government to court for illegal surveillance practices through biometric ID and won. • Call for government **restrictions** (e.g., export controls)
Technological: • surveillance as a business opportunity	• Surveillance technologies suppliers • Every actor being surveilled	• Civil Society and Tech for good actors: Develop **technical solutions t**hat protect citizens' privacy and security, such as anti-spyware tools, thus becoming a sousveillance actor. • Collaborating with companies or **digitally capable actors** from civil society on technical counter and protection solutions and anti-repression tools,

Table 2. Summary of powercube analysis of the African surveillance study Source: Sheombar and Klovig Skelton (2025)

The analysis reveals a pattern of interests from all stakeholders involved in the tension between the interests of proponents of surveillance technologies and citizen rights organisations countering their unbridled proliferation, as shown in Figure 1. They are grouped into four categories: economic, reputational, political, and technological. These categories are drawn from Feldstein (2021) but have been expanded to include an additional category: technological actions to counter harmful state-enabled surveillance. In each of these categories, tensions are identified between the interests of citizens and adversary actors. They provide approaches to counter digital surveillance.

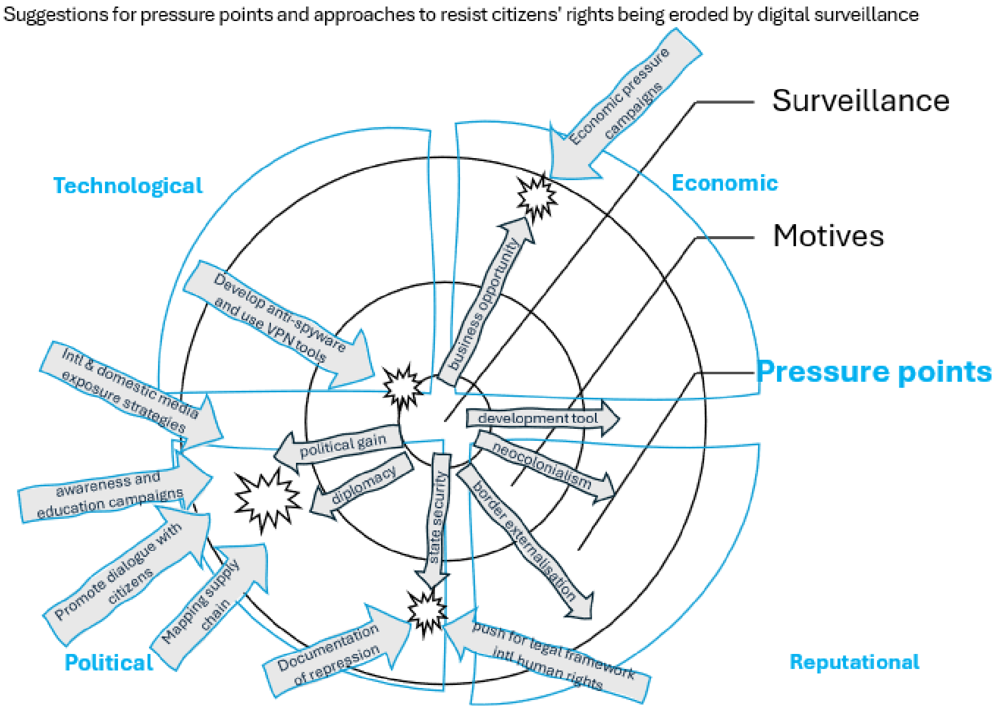

Suggestions for pressure points and approaches to resist citizens' rights being eroded by digital surveillance
Surveillance
Motives
Pressure points
Technological
Economic
Political
Reputational
Economic pressure campaigns
business opportunity
development tool
neocolonialism
border externalisation
state security
political gain
diplomacy
Develop anti-spyware and use VPN tools
Intl & domestic media exposure strategies
awareness and education campaigns
Promote dialogue with citizens
Mapping supply chain
Documentation of repression
push for legal framework intl human rights

Figure 1. Four categories of digital surveillance interests and counter approaches. Source: authors own.

Reputational pressure points are methods intended to disrupt the purportedly good reputation that some actors try to to retain. This type of pressure could be used in situations where state monitoring is driven by a desire to legitimise state security, acquire political advantage, or participate in diplomatic actions.

The economic pressure point could be strategically used in situations where surveillance is pursued as a business opportunity or as a development tool, such as the building of smart or safe cities. This approach would seek to financially impact all players who benefit from the export, purchase, and deployment of surveillance technologies that violate citizens' rights.

The political pressure point can be used strategically in situations where monitoring is driven by political gain, diplomatic goals, or neocolonial ambitions. This action could target governmental actors in both supplying and receiving/demanding states.

Regarding the technological pressure point, using Gaventa's powercube analysis, an intervention for dealing with invisible forms of power, such as surveillance firms' technological capabilities, and supporting 'claimed free of intrusive surveillance' could consist of raising awareness of the dangers of rights-violating surveillance and creating an alternative (safer) digital environment to protect citizens. To counter intrusive surveillance technologies, regardless of the motivations for state surveillance, civil society could develop technical solutions that protect citizens' privacy and security in collaboration with non-adversary actors such as some technology-for-good companies, digitally capable human rights organisations, and universities.

Conclusion

Hitherto, this research has identified the actors involved in the supply and use of state surveillance technologies in Africa. It has also started to use the powercube to analyse how to mitigate against surveillance technologies that violate citizens' rights.

By applying the powercube, we suggest various pathways for resistance that can be grouped into reputational, economic, political and technological actions to target actors involved in the transfer of surveillance technologies to change the power, agency or rights dynamics to favour the rights of citizens. Research has a role to play in this resistance. Drawing on the framework developed Hamdali et al. (2024) to create an impact by impacting practice, their framework provides a powerful tool to conceptualise impactful theory and reflect on the relationship between the production and usage of knowledge to have a practice-based impact (Figure 2).

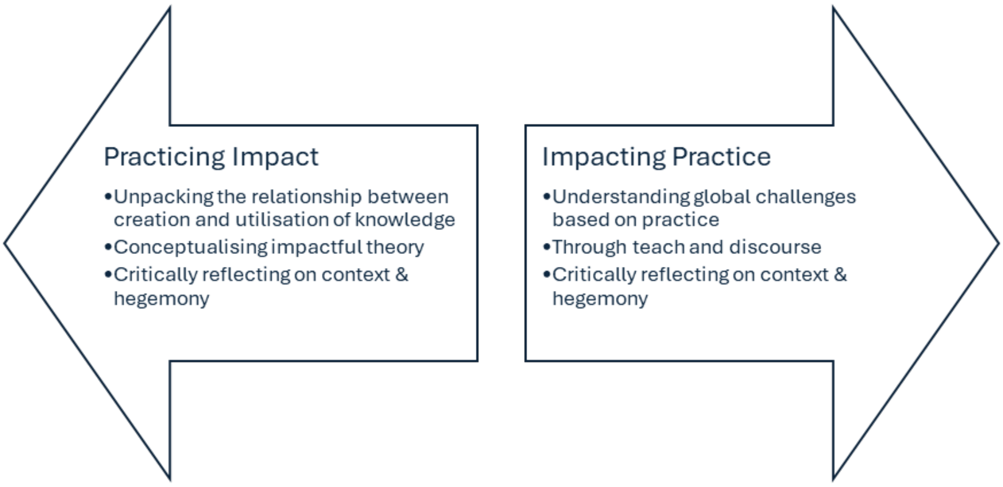

Practicing Impact
•Unpacking the relationship between creation and utilisation of knowledge
•Conceptualising impactful theory
•Critically reflecting on context & hegemony
Impacting Practice
•Understanding global challenges based on practice
•Through teach and discourse
•Critically reflecting on context & hegemony

Figure 2. A framework developed by the authors based on Hamdali et al. (2024) for conceptualising impactful theorising vs practice impact.

By integrating Hamdali et al.'s theory-practice impact framework, this research contributes to bridging surveillance studies research with actionable strategies - while critically reflecting on context and power dynamics -, emphasising multi-stakeholder collaboration, and the positioning of the researcher's impact, either leaning toward practising impact or toward impacting practice. The research advocates context-sensitive interventions, urging policymakers to balance security with rights and surveillance technology actors to develop ethical alternatives, ultimately fostering democratic accountability in Africa's digital landscape. This will enable the development of an impactful approach to countering state surveillance by citizens and other involved actors.

References

Cook, J. (2020). *Israeli Spyware Technology, Tested on Palestinians, Now Operating in a City Near You*. American Educational Trust (AET) Inc. https://www.wrmea.org/2020-january-february/israeli-spyware-technology-tested-on-palestinians-now-operating-in-a-city-near-you.html

Dratwa, J. (2014). Ethics of security and surveillance technologies*.)^(Eds.): 'Book Ethics of security and surveillance technologies' (EGE Opinion Report, 2014, edn.).*

Feldstein, S. (2019). *The global expansion of AI surveillance* (Vol. 17). Carnegie Endowment for International Peace Washington, DC. https://carnegieendowment.org/files/AISurveillanceGlobalIndex.pdf

Feldstein, S. (2021). *The Rise of Digital Repression: How Technology is Reshaping Power, Politics, and Resistance*. Oxford University Press. https://books.google.nl/books?id=W3QjEAAAQBAJ

Gaventa, J. (2019). Applying power analysis: using the "Powercube" to explore forms, levels and spaces. *Power, empowerment and social change*, 117-138.

Greitens, S. C. (2020). Dealing with demand for China's global surveillance exports. *Brookings Institution Global China Report*.

Hamdali, Y., Skade, L., Jarzabkowski, P., Nicolini, D., Reinecke, J., Vaara, E., & Zietsma, C. (2024). Practicing impact and impacting practice? Creating impact through practice-based scholarship. *Journal of Management Inquiry*, *33*(3), 230-243.

Hicks, J. (2022). Export of Digital Surveillance Technologies from China to Developing Countries.

Hillman, J. E. (2021). *The digital silk road: China's quest to wire the world and win the future*. Profile Books.

Jili, B. (2020). The spread of surveillance technology in Africa stirs security concerns. *Africa Center for Strategic Studies*, *11*.

Martin, A. K., Brakel, R. v., & Bernhard, D. J. (2009). Understanding resistance to digital surveillance: Towards a multi-disciplinary, multi-actor framework. *Surveillance and Society*, *6*, 213-232.

Roberts, T., Gitahi, J., Allam, P., Oboh, L., Adekunle Oladapo, O., Appiah-Adjei, G.,…Sheombar, A. (2023). *Mapping the supply of surveillance technologies to Africa: case studies from Nigeria, Ghana, Morocco, Malawi, and Zambia*.

Roberts, T., & Mohamed Ali, A. (2021). Opening and Closing Online Civic Space in Africa: An Introduction to the Ten Digital Rights Landscape Reports. In *Digital Rights in Closing Civic Space: Lessons from Ten African Countries*. Brighton: Institute of Development Studies, United Kingdom. https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/15964

Roberts, T., Mohamed Ali, A., Farahat, M., Oloyede, R., & Mutung'u, G. (2021). Surveillance Law in Africa: a review of six countries.

Schuster, S., van den Berg, M., Larrucea, X., Slewe, T., & Ide-Kostic, P. (2017). Mass surveillance and technological policy options: Improving security of private communications. *Computer Standards & Interfaces*, *50*, 76-82. https://doi.org/https://doi.org/10.1016/j.csi.2016.09.011

Sheombar, A., & Klovig Skelton, S. (2025). Who supplies digital surveillance technologies to African governments? *Pathways for resistance*. In T. Roberts & A. Mare (Eds.), *Digital Surveillance in Africa: Power, Agency, and Rights* (pp. 183-210). Zed Books. https://doi.org/http://dx.doi.org/10.5040/9781350422117.ch-8

van der Lugt, S. (2021). Exploring the Political, Economic, and Social Implications of the Digital Silk Road into East Africa: The Case of Ethiopia. In S. Florian (Ed.), *Global Perspectives on China's Belt and Road Initiative* (pp. 315-346). Amsterdam University Press. https://doi.org/doi:10.1515/9789048553952-014

Wright, D., Rodrigues, R., Raab, C., Jones, R., Székely, I., Ball, K.,…Bergersen, S. (2015). Questioning surveillance. *Computer Law & Security Review*, *31*(2), 280-292. https://doi.org/https://doi.org/10.1016/j.clsr.2015.01.006

Does Digitalization Imply Uncertainty of Future Firm Prospect? Evidence From Analyst Forecast Accuracy

Guanming He[a*],
[a]*Durham Business School, Durham University, Durham, the United Kingdom; DH1 3LB*
Email: guanming.he@durham.ac.uk

April Zhichao Li [b*]
[b]*Business School, University of Exeter, Exeter, the United Kingdom; EX4 4PU*
Email: z.li10@exeter.ac.uk

Tiantian Lin [c*]
[c]*School of Economics and Finance, Huaqiao University, China; 362021*
Email: ttlin@hqu.edu.cn

Completed

[*] The authors are listed based on the alphabetical order of the first letter of the authors' surname. We thank the helpful comments and suggestions from Miklos Vasarhelyi, Sabri Boubaker, and conference/workshop participants at the 62nd World Continuous Audit & Reporting Symposium WCARS and the 2nd British Accounting & Finance Association (BAFA) Accounting Information Systems & Emerging Technologies (AISET) Special Interest Group conference, 3rd Contemporary Issues in Financial Markets and Banking Conference, the 2024 Academy of International Business Asia Pacific conference, and Swansea University. All plausible errors, if any, in the paper remain our own responsibility.

Abstract

This paper examines how analyst forecasts for firms are shaped by corporate digitalization. Based on a large sample of Chinese listed firms, we find that analysts covering firms with higher levels of digitalization are likely to make more accurate forecasts. This result is robust to using two-stage least squares regression, difference-in-differences regression, and firm-fixed-effects regression to mitigate potential endogeneity concerns and elicit causal inferences. Our mediation analysis reveals that digital transformation improves firms' operational efficiency, investment efficiency and information quality, thereby increasing analyst forecast accuracy. We also find that the association between corporate digital transformation and analyst forecast accuracy is more pronounced for analysts with greater work capabilities and more information resources, for firms with lower levels of innovation, as well as for high-tech industries. Our study provides new insights into the implications of digitalization for future firm prospect through the lens of analyst forecast accuracy.

Keywords: digitalization, uncertainty of future prospect, analyst forecasts, operational efficiency, investment efficiency, information quality

1. Introduction

In the new digital era, advanced technologies, such as artificial intelligence (AI), blockchain and big data analytics, have undeniably changed the environment where firms operate. Against this backdrop, it seems to be a necessity for firms to use these advanced tools pursuing digitalization to remain competitive in the dynamic market where opportunities and challenges coexist. Indeed, over the past decades, an increasing number of firms worldwide have integrated advanced digital technologies into their operations, management and information systems (McKinsey & Company, 2018). However, applying digital techniques also comes with risks and costs (e.g., Flyvbjerg and Budzier, 2011; Janssen et al., 2017), making the implications of corporate digitalization for future firm prospects uncertain and a concern for investors. Since analyst forecasts help guide investors in capital allocation decisions, understanding the impact of corporate digitalization on forecast accuracy is crucial for market participants and policymakers.

China provides an ideal setting for our study, given its nationwide promotion of corporate digitalization, the big scale of the digital economy, and the availability of comprehensive data. The Chinese government has long been actively stimulating corporate digital transformation via an array of initiatives, such as the "Internet Plus" strategy launched in 2015 and the construction of "Big Data comprehensive pilot zones" starting in 2016.[1] These national programs aim to promote corporate application of digital technologies in various industries, including agriculture, manufacturing, healthcare, etc., to foster innovation and sustainable economic growth. Such national promotion on corporate digitalization not only attracts investments around the globe, but also provides insights generalizable to other capital markets. As the second largest economy in the world and a global leader in corporate digitalization, China has witnessed a 42.8% GDP contribution by digital economy in 2023.[2] Therefore, a study on digitalization of Chinese listed firms should have relevant implications for other developing countries. In addition, the rich data in China, not least data on digitalization, also allow for a powerful, in-depth empirical analysis of corporate digitalization trends across a wide range of firms of different business scales.

[1] Detailed information of the "Internet Plus" strategy and establishment of Big Data pilot zones can be found via the Chinese government official websites – https://www.gov.cn/zhengce/content/2015-07/04/content_10002.htm; https://www.gov.cn/xinwen/2016-11/07/content_5129639.htm.

[2] Data on the GDP contribution by digital economy in China are obtained from the "2024 Research Report on Digital Inclusive Development in China" published by the China Academy of Information and Communications Technology (CAICT). The report in Chinese is available via the link – http://www.caict.ac.cn/kxyj/qwfb/bps/202408/P020240830315324580655.pdf.

Corporate digitalization, or digital transformation, refers to the process of integrating digital technologies into business operational processes and models. It involves utilizing technologies such as AI, big data analytics, cloud computing, and blockchain to streamline workflows and automate essential business activities, among others. Existing research has documented the economic benefits of corporate digitalization, such as cost reduction and enhanced long-term relationships with customers (e.g., Agarwal et al., 2010; Paiola and Gebauer, 2020). However, considering various risks and uncertainties present in digitalization, it is still questionable whether firms adopting digital techniques could achieve the desirable outcomes. Hence, the implication of digital transformation for firms' future performance remains uncertain to capital market participants, including financial analysts. Given that analyst forecasts play a pivotal role in guiding capital providers' expectations about future firm prospects, understanding the impact of corporate digitalization on analyst forecast accuracy is crucial for capital market participants as well as policymakers.

We posit that corporate digitalization impacts the accuracy of analyst forecasts as it has a real effect on firms' business activities and information management. It could improve the firms' operational efficiency, investment efficiency and information quality. Firstly, advanced digital technologies enable firms to better understand customer behavior, target marketing efforts, and provide customized products as well as services, leading to enhanced customer loyalty and more stable sales. Similarly, cost control becomes more efficient under digitalization as it reduces internal communication costs and mitigates reliance on low-skilled workers, resulting in more predictable expenses. Together, the resulting stable, favorable sales and costs, reflected in higher operational efficiency, would improve the accuracy of analyst forecasts. Secondly, digitalization boosts investment efficiency by delivering comprehensive data swiftly for informed decision-making and optimizing project selection. The improved decision-making and risk assessments for project selection by firms would provide a more reliable foundation for analyst earnings forecasts. Thirdly, digitalization increases information quality not only by facilitating managers to make timely, accurate forecasts but also by improving internal controls through technologies like blockchain, which enhance data traceability and reduce errors or fraud, further improving analyst forecast accuracy. Nevertheless, implementing digital technologies requires significant investments in infrastructure and human capital. The outcomes of such investments might not fit in with a firm's existing activities or managers' capabilities, causing business uncertainty and disruptions that potentially prevent analysts from making accurate forecasts for the firm. In addition, cybersecurity risks arising from digital transformation, not least damages or leakages of proprietary information, would also add uncertainties to future firm performance. For these reasons, whether and how corporate digitalization would affect analyst forecast accuracy is an open, empirical issue to explore.

Using a sample from Chinese listed companies for the period 2011-2022, we find that analyst forecast accuracy is higher for firms embracing digitalization. This favorable effect is realized through improvements in firms' operational efficiency, investment efficiency and information quality. To strengthen causal inferences, we do a range of analyses, including two-stage least squares (2SLS) regression, difference-in-differences (DID) regression, and alternative measures for both digitization and analyst forecast accuracy. Our finding still holds after all these robustness checks. Our further analyses reveal that the positive impact of digitalization on analyst forecast accuracy is more prominent for analysts with greater capabilities and more information resources, for firms with lower levels of innovation, and for industries with higher levels of technology development.

Our study contributes to the existing literature in three aspects. First, given the pros and cons of corporate digitalization, its implications for future firm prospects remain uncertain, concerning capital market participants. The higher the uncertainty regarding such implications, the higher the likelihood and extent of resource misallocation in the capital market, where less promising firms would have received more capital than they should. While recent studies (e.g., Huang et al., 2020; Yasmin et al., 2020; Zeng et al., 2022; Chen and Srinivasan, 2024; Lem, 2024) have examined the influence of digitalization on firm performance, little research attention has been paid to the plausibly uncertain implications of digitalization for firm future prospect. Our study fills this gap in the literature by providing insights into how digital transformation affects the accuracy of analyst forecasts.

Second, this study adds to the literature on financial analysts. Extant research documents that several business activities, including innovation and ESG engagement, affect analyst earnings forecasts (e.g., Gu and Wang, 2005; Chahine et al., 2021; Canace et al., 2024). Our paper enriches this literature by exploring how core corporate business activities, facilitated by digital technologies, affect analyst forecasts. Specifically, digital technologies improve corporate operational efficiency, investment efficiency and information management. These changes in real activities are critical for analysts in their comprehension of corporate performance and their accurate forecasts of firms' future earnings. Our findings suggest that analysts could capitalize on their work capabilities and available information resources to achieve even higher forecast accuracy for digitalized firms.

Third, this paper sheds light on the nexus between real corporate activities and the predictability of firm performance. It illuminates how digital transformation enhances a firm's ability to stabilize and improve performance, thus reducing unpredictability in financial results, which contributes to analysts' accurate forecasts.

The remainder of this paper is organized as follows: Section 2 reviews the relevant literature and develops our hypothesis. Section 3 describes our data sources, sample selection process, and measurements of variables. Section 4 specifies the research design and discusses the empirical results. Section 5 explores the potential mechanisms underlying our baseline results. Section 6 conducts moderation analyses. Section 7 concludes this research.

2. Literature review and hypothesis development

Analysts are important stock market participants who serve as information intermediaries between firms and investors. They gather and process various corporate information to generate earnings forecasts for firms, aiding investors in evaluating corporate performance and predicting the firms' future prospects (Lang and Lundholm, 1996). Therefore, the accuracy of analyst forecasts is essential in reducing information asymmetry among investors and enhancing stock market efficiency. Firms' digitalization would impact the accuracy of analyst forecasts, in that it affects the predictability and reliability of firms' earnings by reshaping corporate operational efficiency, investment efficiency and information quality, specifically as follows.

First, firm digitalization could improve operational efficiency, not only increasing profitability but also enhancing the consistency and predictability of corporate performance. By using advanced digital technologies, firms could improve their dynamic capabilities (Mikalef et al., 2019),[3] which are vital for survival and continuous growth in a constantly changing environment (Mikalef and Pateli, 2017). Specifically, digital tools such as big data analytics facilitate firms to deepen their understanding of customers' characteristics alongside purchasing behaviors (Chen et al., 2012). These insights enable firms to quickly adapt to the market by executing precise marketing campaigns towards targeted customers (Goldfarb and Tucker, 2019). They also facilitate the provision of customized products or services tailored to the unique needs and preferences of customers (Ryu and Lee, 2018; Blichfeldt and Faullant, 2021). Additionally, the utilization of digital assistance, such as AI-powered service, ensures continuous after-sale support and rapid responses to customers' inquiries, potentially increasing customer satisfaction (Huang and Rust, 2018). Thus, digitalization equips firms with a better

[3] Dynamic capabilities refer to the ability to integrate, build and reconfigure internal and external competencies to address rapidly changing environments (Teece et al., 1997).

ability to resist the negative impact of external uncertainty on future performance and cultivate the loyalty of existing customers, thereby creating relatively stable sales or sales growth for the firms. Besides, with the help of digital techniques as to blockchain, firms could strengthen internal communication and integrate resources across different departments that involve operational processes from raw materials procurement to sales (Iansiti and Lakhani, 2017), ensuring optimal resource utilization, maintaining effective cost controls, and thus enabling more predictable expenses. Furthermore, digital techniques are potential substitutes for certain human tasks, involving mainly low-skilled labor, while complementing the sophisticated workforce (Acemoglu and Restrepo, 2018; Agrawal et al., 2019). This potential shift in labor might induce an increase in fixed costs as a proportion of total costs, leading to more constant costs. Consequently, firms with stable sales and costs could achieve more consistent earnings over time, which lowers the difficulty for analysts in forecasting the firms' future performance. Second, the adoption of digital technologies could improve the investment efficiency of firms (Chen and Jiang, 2024), reducing the uncertainty and risk stemming from corporate investing activities (Jiang et al., 2024). Digitalization empowers firms with the technological advantage of collecting extensive data and information from internal and external environments, facilitating well-informed investment decisions. Existing research (Dorantes et al., 2013) provides evidence that after adopting an advanced digital tool called enterprise system, the predictive ability of firm management is enhanced, which is attributable to the improved access to a comprehensive set of internal information in a timely manner. Such improved capabilities would mitigate managers' overly optimistic expectations and consequently decrease overinvestments (e.g., Chen and Jiang, 2024). From another perspective, digital tools, such as machine learning and big data analytics, also augment firms' ability of analyzing market trends, tracking industry dynamics, and understanding the risk profiles of different investment options. As a result, firms can gain more precise risk assessment and accurately predict market demands to invest in more profitable projects (e.g., product expansion or development). This is particularly important for investment decisions related to innovations which may require several years to complete. Armed with digital techniques, firms are able to make timely adjustments during research and development (R&D) to align with changing market needs, reducing the risk of outdated innovations and staying ahead of competitors. This improves R&D success and reduces costly investment failures (Niebel et al., 2019; Wu et al., 2020). Meanwhile, the precise and real-time monitoring on resource utilization, as enabled by digital techniques, could further ensure that firms' resources are allocated on a timely basis to projects with high value-creation potential. This improved investment efficiency helps the firms reduce potential investment failures, stabilize investment flows, and make investment outcomes more predictable in accordance with market demands. Hence, it provides analysts with a more reliable basis for earnings forecasts and reduces their forecast errors. Besides, firms with high investment efficiency are more likely to disclose investment-related information, such as project progress and expected returns (Elberry and Hussainey, 2020), also aiding analysts in making more accurate forecasts about the firm's future prospects.

Third, digitalization increases the information quality of a firm by fostering a robust internal control (Cheng et al., 2024) as well as improving data accessibility and analysis (Huang et al., 2018; Lem, 2024). Advanced techniques, exemplified by blockchain, big data analytics and AI, contribute to a data-driven management system in the firm, promoting high-quality reporting mechanisms and real-time monitoring. For example, blockchain technology increases the traceability of data across different departments, making it easier to detect any fraud, errors or internal control defects on time (Yermack, 2017). A recent study by Ashraf (2024) shows that digitally advanced firms are indeed subject to few material weaknesses in internal controls and have a high quality of financial reporting. Apart from strengthening internal controls, digital technologies empower firms to access and analyze extensive data efficiently, thereby enabling managers to make more accurate, timely forecasts of the firms' future performance (Dorantes et al., 2013; Lem, 2024). These high-quality management forecasts could, in turn, serve as valuable inputs for analysts' forecasts and enrich corporate information environments. Since analysts rely heavily on accurate and timely information to forecast earnings, improvements in the corporate information environment should increase the accuracy of their earnings forecast.

Although digitalization offers a good deal of benefits to a firm, its implementation requires substantial investments in human capital and physical infrastructures, which may interfere with expected outcomes. Professionals with the necessary knowledge must be in place in order to employ digital techniques smoothly and effectively. Nonetheless, integrating these digital tools into real business activities involves a learning curve that demands considerable time and effort. As a result, the anticipated benefits may be realized slowly with little impact, particularly in the initial stages. Furthermore, for digital technologies to be effective, they need to closely align with a firm's core business activities (Correani et al., 2020). However, such alignment entails risk and uncertainty. If a digital technology fails to integrate seamlessly with the firm's operational realities, it may disrupt existing business processes and workflows, reducing the clarity of financial outcomes. Put differently, it would bring more uncertainties rather than the anticipated stable performance in operation, investment and information management. Consequently, analyst forecast error could increase.

In addition, digital transformation exposes firms to cybersecurity risks. A cyberattack compromising critical internal information could lead to operational disruptions. If the firm's proprietary information is leaked to competitors, its competitive advantage would be undermined. The information leakage, if concerning the firm's business stakeholders, would damage its reputation, erode trust among the stakeholders, and even induce legal repercussions. In all these scenarios, the firm's future performance would become more uncertain, making it challenging for analysts to forecast. Considering the foregoing pros and cons of digitalization to firms' real activities, we propose the following null hypothesis for empirical tests:

H1: Firm digitalization is not associated with analyst forecast errors.

3. Data and sampling

3.1 Variable measurements

3.1.1 Analyst forecast accuracy

Following the existing literature on analyst forecasts (e.g., Mansi et al., 2011; Datta et al., 2011; He and Li, 2024), we measure the accuracy of each analyst forecast ($accuracy_{i,j,t}$) as -1 times the ratio of the absolute difference between actual and forecasted earnings per share (EPS) to the closing stock price on the trading day preceding the analyst's forecast, as specified in Model (1). Consistent with previous research (e.g., Clement and Tse, 2005; Bradley et al., 2017), forecasts issued later than one month before the fiscal year-end are excluded from the construction of $accuracy_{i,j,t}$. In cases in which an analyst issues multiple forecasts for a firm in the fiscal year, only the most recent forecast is retained. For each firm-year observation, the average forecast accuracy ($Accuracy_{i,t}$) for a firm is computed as the mean of forecast accuracy among all analysts following the same firm in a year, as specified in Model (2). Higher values of $Accuracy$ denote greater analyst forecast accuracy for a firm.

$$accuracy_{i,j,t} = -1 * \frac{\left|EPS_{i,j,t}^{Forecast} - EPS_{i,t}^{Actual}\right|}{P_{i,j,t}} \tag{1}$$

$$Accuracy_{i,t} = \sum_{j=1}^{N} accuracy_{i,j,t}/N \tag{2}$$

where $EPS_{i,j,t}^{Forecast}$ is the forecasted EPS released by analyst j for firm i in year t, and $EPS_{i,t}^{Actual}$ is the actual EPS of firm i in year t. $P_{i,j,t}$ is the closing stock price of firm i on the trading day preceding the analyst j's forecast for firm i in year t. N is the number of analysts following firm i in year t.

3.1.2 Firms' digital transformation

Existing literature measures the level of firms' digital transformation using two primary approaches. Some scholars (e.g., Chen et al., 2022; Guo et al., 2023) count on financial report data, specifically the year-end intangible assets related to digital transformation. These digital-related intangible assets include digital-related patents as well as items with names containing keywords related to digital technologies, such as digital software and networks. Other researchers (e.g., Wu et al., 2021; Wen et al., 2022; Tu and He, 2023; Guo et al., 2023) employ textual analysis to quantify digital-related content in the firm's annual report. To this end, they construct a dictionary of digital-related keywords via textual analysis (e.g., Wu et al., 2021; Guo et al., 2023) and then calculate the frequency of digital-related terms, which appear in the annual reports, as the proxy for digital transformation.

To capture both the qualitative and quantitative dimensions of digital transformation in firms' annual reports, we use common factor analysis to create a composite index ($Digit$) from two variables: (i) the natural logarithm of one plus digitalization-related intangible assets disclosed in the firm's annual report ($Digit1$) and (ii) the natural logarithm of one plus the frequency of digital-related words in the management discussion and analysis (MD&A) section of the firm's annual report ($Digit2$). Panels A and B of Appendix C present the results of our common factor analysis. The factor with a higher eigenvalue accounts for a greater proportion of the variance in the digital transformation variables. Panel A reveals that the first factor has an eigenvalue of 1.1925, significantly higher than the eigenvalue of 0.8075 for the second factor. According to the Kaiser's criterion (Kaiser, 1960), which suggests retaining factors with eigenvalues above 1, we retain only the first factor. Panel B shows that both $Digit1$ and $Digit2$ load positively on the first factor and are positively correlated with it. The communality values, which indicate the proportion of the variance in each digital transformation variable explained by the first factor, are all relatively high. Together, these results indicate that the first factor is the most powerful and distinct measure of a firm's overall digital transformation. Therefore, we use this factor as our composite measure of firms' digital transformation ($Digit$).

3.1.3 Control variables

Drawing on the previous literature on analyst forecasts (e.g., Lehavy et al., 2011; He et al., 2024), we control for a battery of firm-level variables, including firm size ($LnSize$), financial leverage ($Leverage$), return on assets (ROA), sales growth ($Sale\_growth$), earnings surprise ($Earnings\_surprise$), stock return volatility ($Return\_volatility$), cash ratio ($Cash\_ratio$), directors' shareholdings ($Director\_shares$), state ownership (SOE), institutional investors' stock ownership ($Insti$), audit quality (Big_4), firm age ($LnAge$) and analyst coverage ($LnAnalyst\_coverage$). As with Li et al. (2023) and Liu et al. (2024), we also control for analyst characteristics, including analysts' educational background ($Education\_analyst$), their all-star status ($Star\_analyst$)[4], work experience ($LnExperience\_analyst$), earnings forecast horizon ($LnHorizon$) and the amount of their firm coverage ($LnFollowing\_analyst$). We take the average of these analyst-level variables if multiple analysts are covering the same firm in a year. Definitions for all the foregoing variables are provided in Appendix A.

3.2 Data sources and sample selection

Our empirical tests are based on Chinese firms listed in mainland China. Data used for the tests comes primarily from two sources. First, the firms' digitalization-related information in their annual reports is obtained from the Chinese Research Data Service (CNRDS) platform. Our sample period begins in 2011, the year when digitalization data were first made available by CNRDS, and ends in 2022, the most recent year for which digitalization data are updated. Second, data utilized to construct the foregoing firm-level and analyst-level variables are taken from the China Stock Market and Accounting Research (CSMAR) database.

Appendix B presents our sample selection procedure. The initial sample consists of 39,111 firm-year observations for firms listed on the Shenzhen or Shanghai Stock Exchange from 2011 to 2022. We exclude 1,658 observations for firms categorized as "Special Treatment (ST)", "Special Treatment with a Risk of Delisting (*ST)" or "Particular Transfer (PT)", as these firms suffer from pending delisting. We remove 1,066 observations from firms in the finance industry. Further, we drop 14,906 observations with missing values in the variable of analyst forecast accuracy. Finally, we eliminate 646 (1,693) observations with missing values in the variables as to analyst (firm) characteristics. This screening process results in a final sample of 19,142 firm-year observations across 3,534 firms, of which 1,493 (2,041) firms are listed on the Shanghai (Shenzhen) Stock Exchange.

3.3 Univariate statistics

To minimize the effect of outliers on our results, we winsorize all the continuous variables at the 1st and 99th percentiles. Panel A of Table 1 presents the descriptive statistics of variables used in the baseline regression analysis. The dependent variable, $Accuracy$, has a mean

[4] The list of star analysts is sourced from the *New Fortune* for each year during our sample period, except for 2018 in which the list is missing from the website and thus sourced alternatively from the *China Securities Analyst Golden Bull Awards* and the *Crystal Ball Awards*.

(median) value of -0.0166 (-0.0098), which aligns with the findings reported by Hou et al. (2022). The level of digital transformation ($Digit$) averages -0.0022 in our sample firms. Panel B reports the Spearman correlation matrix among the variables. The correlation between analyst forecast accuracy ($Accuracy$) and firms' digital transformation ($Digit$) is significantly positive. The absolute values of all correlation coefficients are below 75%, suggesting that multicollinearity is not a substantial concern in our analysis.

[Insert Table 1 about here]

4. Research design and results

4.1 Baseline regression

To test the impact of firms' digital transformation on analyst forecast accuracy, we run the following ordinary least squares (OLS) regression:

$$Accuracy_{i,t}=\alpha+\beta_1 Digit_{i,t}+\beta Controls_{i,t}+IndustryDummies+YearDummies+\varepsilon_{i,t} \quad (3)$$

where $Accuracy$ and $Digit$ are defined as in Section 3.1 for firm i in year t. Industry dummies and year dummies are also included in the regression. Standard errors of coefficients are clustered by firm. If firms' digital transformation is positively (negatively) associated with analyst forecast accuracy, β_1 should be positive (negative) and statistically significant at a conventional level.

[Insert Table 2 about here]

Table 2 reports the regression results. All explanatory variables have variance inflation factors below 5, confirming that multicollinearity is not a concern in our regression estimation.[5] In Column (1), the coefficient for $Digit$ is positive and statistically significant at the 1% level. A one-standard-deviation increase in $Digit$ raises $Accuracy$ by 0.0015, which accounts for 8.99% of the sample mean of $Accuracy$ and is thus economically significant. We replace industry-fixed effects with firm-fixed effects for Model (3), and report the results in Column (2), which appear qualitatively the same as those in Column (1). These findings indicate that analysts covering firms with higher levels of digital transformation are likely to make more accurate forecasts.

4.2 Robustness tests — controlling for endogeneity

[5] The mean of variance inflation factors (VIF) for our baseline OLS model is 1.72, with $Size$ showing the highest VIF at 3.03. In all other regressions estimated in our study, VIF values are below 5. Detailed VIF results are available upon request.

Provided that some variables omitted in our regression are correlated with both firms' digital transformation and analysts' forecast accuracy, our results would be biased. To mitigate this endogeneity concern, we employ two identification strategies: (i) a two-stage least squares (2SLS) regression that includes two exogenous instrumental variables and (ii) a difference-in-differences (DID) regression that involves an exogenous event for a quasi-natural experiment.

4.2.1 Two-stage instrumental-variables regression analysis

For the 2SLS regression analysis, the ideal instruments used in the first-stage regression estimation should capture variances in firms' digital transformation that are exogenous to analyst forecast accuracy. The first instrument we choose is the level of internet development in a city for a year ($Internet$). We do a common factor analysis to integrate four variables, (i) the number of internet users per one hundred people, (ii) the proportion of employees in the computer services and software industry to the total employees, (iii) the amount of per capita telecommunications, and (iv) the number of mobile phone users per one hundred people, into a composite index on the internet development in a city for a year ($Internet$).[6] Firms' digital transformation typically hinges on the development of computer network infrastructure (Vogelsang et al., 2018; Tian et al., 2022), but the latter is unlikely to influence analyst forecast accuracy directly.[7] The second instrumental variable is the average level of digital transformation ($Digit\_same\_city$) among all other firms located in the same city in a year. We posit that while analyst forecast accuracy for a firm is not directly influenced by the average level of digital transformation of other firms in the same city, the firm's own digital transformation might benefit from shared digital resources within the city.

[Insert Table 3 about here]

Column (1) of Table 3 shows the results of first-stage regression, in which $Digit$ is the dependent variable. The instrumental variables, $Internet$ and $Digit\_same\_city$, have positive coefficients that are statistically significant at the 1% level. This indicates that a firm's

[6] Data on the internet development in cities are collected from the China City Statistical Yearbook.

[7] Analysts often do not work in the same regions as the firms they cover. Even when they do, their technological infrastructure and analytical tools are typically integrated into employer-provided, enterprise-grade networks and function independently of regional internet conditions. This ensures consistent performance across locations and minimizes the influence of a firm's local internet environment on their work. Additionally, many analytical tools, such as Python, MATLAB, and Stata, function locally without requiring internet access. Analysts also have access to global data centers and cloud computing resources like Bloomberg, FactSet, and Thomson Reuters Eikon, which are robust, globally distributed systems built for scalability and high-level analytical tasks regardless of geographic location. These attributes ensure that analysts' forecasts remain unaffected by the level of internet development in a firm's headquarters city. Therefore, internet availability satisfies the exclusion restriction criterion for a valid instrumental variable.

digital transformation is positively related to both the internet development in the city where the firm is headquartered and the average level of digital transformation of other firms in the same city. The F-statistic for the instruments is 14.33 and significant at the 1% level, confirming the relevance of $Internet$ and $Digit\_samecity$ as instruments for the 2SLS regression analysis. The p-value for the overidentifying restriction test exceeds 10%, lending further support to the validity of instruments. Column (2) reports the results of the second-stage regression, in which the fitted values of $Digit$ (namely, $Pred\_Digit$) estimated from the first-stage regression are used as the main explanatory variable of interest. The coefficient for $Pred\_Digit$ is positive and statistically significant at the 1% level, substantiating that our baseline results are robust to controlling for endogeneity.

4.2.2 Difference-in-differences regression analysis

Our second identification strategy pertains to a quasi-natural experiment that generates exogenous variation in firms' digital transformation. We leverage the establishment of China's big-data comprehensive pilot zones in 2016 as an exogenous shock to firms' digital transformation and examine its impact on analyst forecast accuracy. In February 2016, the Chinese government established the first big data analytics pilot zone in Guizhou Province. In October of the same year, the initiative expanded to cover seven more regions: Beijing–Tianjin–Hebei, Guangdong, Henan, Shanghai, Inner Mongolia, Chongqing, and Shenyang. This policy aims to integrate and apply regional big data resources and technologies to enhance local firms' digital transformation (Wei et al., 2023). To test how the establishment of China's big data pilot zones impacts analyst forecast accuracy, we construct the following difference-in-differences (DID) OLS regression model:

$$Accuracy_{i,t} = \alpha + \beta_1 Treatment_{i,t} + \beta_2 Treatment_{i,t} * Post_{i,t} + \beta Controls_{i,t} + IndustryDummies + YearDummies + \varepsilon_{i,t} \quad (4)$$

where the treatment indicator variable, $Treatment$, equals 1 (0) for the treatment (control) firms, identified as those headquartered in the eight big-data comprehensive pilot zones (those outside of these regions). The time indicator variable, $Post$, equals 1 if a firm is in a fiscal year during the three-year post-policy period (i.e., 2017-2019), and 0 if in the three-year pre-policy period (i.e., 2013-2015). The variable of interest to our test is the interaction term, $Treatment * Post$, of which the coefficient captures the effect of establishments of China's big-data comprehensive pilot zones on the accuracy of analyst forecasts for firms covered by the policy relative to those that are not. All the control variables are the same as those in Model

(3). *Post* exhibits multicollinearity with year dummies and is therefore excluded from the regression estimation.

[Insert Table 4 and Figure 1 about here]

Table 4 reports the results from DID model specifications that are based on four different samples. Columns (1) and (2) present the results of DID regression run based on the original sample. We test the parallel trends assumption for our DID estimator by including the interaction terms between year dummies and the treatment indicator variable, $Treatment$, in an OLS regression with the same controls as in Model (4). The coefficients for the pre-event interaction terms, $Treatment * Year2013$, $Treatment * Year2014$, and $Treatment * Year2015$, are not statistically significant, thus supporting the parallel trends assumption. Consistent with these results, Panel A of Figure 1 reveals a similar trend of analyst forecast accuracy between the treatment and control groups in the pre-policy sample period. By contrast, the coefficients for the post-event interaction terms, $Treatment * Year2017$, $Treatment * Year2018$, and $Treatment * Year2019$, are positive and statistically significant, indicating that the establishments of China's big-data comprehensive pilot zones increase analyst forecast accuracy in each post-policy year. The increasing magnitude of these coefficients suggests that the effect of the policy on analyst forecast accuracy materializes over time. Column (2) reports the DID regression results. The coefficient of the interaction term, $Treatment * Post$, is positive and statistically significant. This indicates that, following the establishment of China's big data analytics pilot zones in 2016, the accuracy of analyst forecasts for firms in the policy-covered areas increases, relative to that for firms located outside of these areas.

To reduce sample selection bias attributed to potential differences in firm characteristics between the treatment group and control group, we adopt three matching approaches – propensity-score matching (PSM), entropy balancing (EB), and coarsened-exact matching (CEM), respectively, to match each treatment firm, without replacement, with a control firm. The matching is done based on eight covariates concerning the main generic dimensions of firm characteristics that are likely associated with the degree of a firm's digitalization (e.g., Sun et al., 2018; Achim et al., 2022; Audretsch and Belitski, 2024): firm size (*LnSize*), firm performance (*ROA*), growth prospect (*Sale_growth*), corporate governance (proxied by directors' shareholdings, *Director_shares*), earnings news (measured by earnings surprise, *Earnings_surprise*), firm risk (represented by stock return volatility, *Return_volatility*), financial health (captured by the cash ratio, *Cash_ratio*), and research and development ($R\&D\ expenditure$).

Under the PSM approach, each observation from the treatment group is matched with a control observation that has the propensity score closest to that of the treatment observation. Propensity scores are estimated based on the logit regression of $Treatment$ on the foregoing seven matching covariates alongside industry dummies and year dummies (He and Tian, 2013; Chen et al., 2015). There are two assumptions for PSM: (i) covariate balance (i.e., similarity of matching covariates between treatment and control groups) and (ii) common support (i.e., overlap in the distribution of propensity scores). We test these two assumptions in Panels A and B of Appendix D, respectively. The results from running the logit regression of the treatment indicator variable, $Treatment$, on the eight covariates for the unmatched (matched) sample are reported in Column (1) (Column (2)) under Panel A. The statistical insignificance of the coefficients of all the covariates for the post-matched sample substantiates that the covariate balance is achieved after the PSM. Panel B shows that the distributions of propensity scores for the treatment and control groups are similar post-matching, thus meeting the common support assumption of PSM.

The matching method as to entropy balancing reweights each observation in the control sample to balance the mean, variance, and skewness of all the covariates between treatment and control groups (Blanco et al., 2023; Madsen and McMullin, 2020). Unlike PSM, entropy balancing (EB) adjusts the sample weights directly, obviating the need to check covariate balance and common support (Pohl et al., 2022). There is no need to do so either under coarsened-exact matching (CEM), as the coarsening bounds for covariates are chosen *ex-ante* in an automatic manner (Iacus et al., 2012; He et al., 2020; Chen et al., 2024).

Following the PSM, EB and CEM, we first re-do the tests of parallel trends assumption based on the matched sample and report the results in Columns (3), (5), and (7) of Table 4. The results all support the parallel trends assumption, so do the graphs portrayed in Panels B, C, and D of Figure 1. We then re-run the DID regression based on the propensity-score matched sample, entropy balanced sample and coarsened-exact matched sample, and report the results in Columns (4), (6), and (8), respectively. In these columns, the variables of interest, $Treatment * Post$, all have positive and statistically significant coefficients, similar to our initial regression results based on the unmatched sample.

4.3 Robustness checks — alternative variable measurements

We re-run Model (3) using alternative measures of firms' digital transformation and analysts' forecast accuracy. For digital transformation, we employ five alternative measures: $Digit1$ is the natural logarithm of one plus digitalization-related intangible assets disclosed in a firm's annual report; $Digit2$ is the natural logarithm of one plus the frequency of digitalization-related words in the management discussion and analysis (MD&A) section of a firm's annual report; $Digit3$ ($Digit4$; $Digit5$) is the relative level of digital transformation for a firm in a year, measured as the difference between a firm's digital transformation and the average digital transformation for firms in the same industry (city; province) in a year, divided by the standard deviation of digital transformation of these firms. For analyst forecast accuracy, we use $Accuracy1$ as an alternative measure. $Accuracy1$ represents the average forecast accuracy of all analysts following the same firm in a year; yet, the forecast accuracy of individual analyst is calculated as -1 times the ratio of the absolute difference between the actual and forecasted EPS to the actual EPS, rather than using the closing stock price as in the $Accuracy$ measure. In cases in which an analyst releases multiple forecasts for the same firm within a year, the last forecast issued at least one month before the end of the fiscal year is utilized for the variable construction. Panels A and Column (1) of Panels B in Table 5 report the results based on the alternative measures of firms' digital transformation and analyst forecast accuracy, respectively. The coefficients of the digital transformation variables are all positive and statistically significant at the 1% level, confirming again that analysts have higher forecast accuracy for firms with higher levels of digital transformation.

Additionally, we test whether analyst forecasts for more digitalized firms exhibit less optimistic or pessimistic bias. To this end, we construct the variable ($Optimism$ ($Pessimism$)) to capture the average forecast optimism (pessimism) among all analysts following the same firm in a year. For each analyst, if his/her forecasted EPS is higher (lower) than the actual EPS, the forecast optimism (pessimism) is computed as the ratio of the absolute difference between the actual and forecasted earnings per share to the closing stock price on the trading day preceding the analyst's forecast; otherwise, it is assigned a value of zero. If an analyst issues multiple forecasts for the same firm within the year, the most recent forecast released at least one month before the fiscal year-end is used to construct the $Optimism$ and $Pessimism$ measures for regression estimation. Columns (2) and (3) under Panel B of Table 5 present the regression results. The negative and statistically significant coefficients of *Digit* suggest that the improved analyst forecast accuracy, driven by enhanced firm digitalization, is manifested in the reduction of both optimistic and pessimistic forecast biases.

[Insert Table 5 about here]

5. Mechanism tests

Our findings so far support the hypothesis that firms' digital transformation renders analyst forecasts more accurate. We next explore the potential economic mechanisms underlying this hypothesis. We identify three mechanisms, operational efficiency, investment efficiency and information quality, and test these mechanisms in Sections 5.1, 5.2 and 5.3, respectively.

5.1 The mediating effect of operational efficiency

As discussed in Section 2, operational efficiency is arguably an underlying economic mechanism through which digital transformation enhances analyst forecast accuracy. To measure firms' operational efficiency ($Operational\_effi$), we do a data-envelopment analysis (DEA) to construct an efficient production frontier, based on an optimization program, to maximize the production output-to-input ratio (Cheng et al., 2018). Production output is proxied by sales revenues, while production inputs include the number of employees, costs of goods sold, and capital expenditures (Lam et al., 2016). For firms within the same industry and year, those on the efficient frontier are assigned a value of one, while less efficient firms behind the frontier receive efficiency scores between 0 and 1.

[Insert Table 6 about here]

To test the mechanism, we perform a two-step mediation analysis. In the first step, we regress firms' operational efficiency ($Operational\_effi$) on digital transformation ($Digit$). As reported in Column (1) under Panel A of Table 6, the coefficient of $Digit$ is 0.0015 and statistically significant at the 1% level, indicating that digital transformation promotes firms' operational efficiency. In the second step, we estimate Model (3) with the main variable of interest replaced by the predicted values of operational efficiency that are estimated from the first-stage regression ($Pre\_Operational\_effi$). As reported in Column (1) of Panel B, $Pre\_Operational\_effi$ takes on a positive, statistically significant coefficient, indicating that analyst forecast accuracy increases with firms' operational efficiency. Collectively, our evidence suggests that operational efficiency is an underlying economic mechanism through which firms' digital transformation enhances analyst forecast accuracy.

5.2 The mediating effect of investment efficiency

Investment efficiency could be another mechanism that explains the positive effect of firms' digital transformation on analyst forecast accuracy. To test this supposition, we first construct a measure of firms' investment efficiency ($Investment\_effi$). It equals -1 times the absolute values of the residuals of a regression model developed by Richardson (2006), where Tobin's

Q is used as the proxy for the growth prospect of a firm in a year.[8] We then conduct a two-step mediation test. In the first step, we regress firms' investment efficiency ($Investment\_effi$) on digital transformation ($Digit$). Column (2) of Panel A in Table 6 reports the results. The coefficient of $Digit$ amounts to 0.0010 and is statistically significant at the 1% level, implying that digital transformation is conducive to increasing firms' investment efficiency. In the second step, we regress analyst forecast accuracy ($Accuracy$) on the predicted values of firms' investment efficiency that are estimated from the first-stage regression ($Pre\_Investment\_effi$). The results, presented in Column (2) of Panel B, show positive and statistically significant coefficients for $Pre\_Investment\_effi$, suggesting that the accuracy of analyst forecasts is higher for firms with higher investment efficiency. Taken together, our findings suggest that firms' digital transformation improves investment efficiency and thereby increases analyst forecast accuracy.

5.3 The mediating effect of information quality

Information quality might also mediate the positive association between firms' digital transformation and analyst forecast accuracy. To test this conjecture, we construct two variables to measure a firm's information quality. First is the quality of accruals reported by a firm ($Accruals\_quality$), which equals -1 times the absolute values of abnormal accruals estimated based on the modified Jones model (Dechow et al., 1995) for the firm in a year. The second variable is management forecast accuracy ($MF\_accuracy$), measured as the negative of the natural logarithm of the absolute difference between management earnings forecast and actual earnings for a fiscal year, scaled by the firm's stock price at the beginning of the fiscal year. Using these variables as mediators, we conduct a two-stage mediation regression analysis. The first-stage regression concerns the impact of firms' digital transformation on their information quality. The results are reported in Column (3) under Panel A of Table 6. It reveals statistically significant positive coefficients for $Digit$, indicating that firms with high levels of digital transformation are likely to have higher information quality. Column (3) of Panel B presents the results of the second-stage regression, which tests the relationship of analyst forecast accuracy with the predicted values of firms' information quality ($Pre\_Accruals\_quality$ and $Pre\_MF\_accuracy$); the latter is derived from the first-stage regression. The coefficients of $Pre\_Accruals\_quality$ and $Pre\_MF\_accuracy$ are significantly positive, suggesting that

[8] We get qualitatively the same results after using the market-to-book ratio or sales growth as alternative measures of the firm's growth prospect.

analyst forecasts are more accurate for firms with higher information quality. Combined, the evidence aligns with our conjecture that improved information quality is an underlying channel through which firms' digital transformation increases analyst forecast accuracy.

6. Moderation analyses

To further illuminate the impact of digital transformation on analysts, we conduct a series of moderation analyses regarding how the effect of digital transformation on analyst forecast accuracy varies on analyst characteristics, corporate innovation and industrial technology development.

6.1 The moderating effects of analyst characteristics

Analyst characteristics may impact forecast performance substantively, as personal traits are typically bound up with their ability to acquire and process value-relevant information, and with the levels of their professional expertise (e.g., Kim et al., 2011). We focus on two categories of analyst characteristics, (i) internal features referring to analysts' work capabilities and (ii) external features related to analysts' available resources for forecasting, and test how these characteristics moderate the relationship between firms' digital transformation ($Digit$) and analyst forecast accuracy ($Accuracy$). Analysts with superior work capabilities and privileged access to resources are better equipped to interpret complex information and thereby infer the implications of corporate digitalization for firms' future prospects. As a result, the accuracy of their forecasts for the digitalized firms is expected to be higher.

We operationalize analysts' work capabilities by using their work experience ($LnExperience\_analyst$) and past forecast performance ($Past\_accuracy$) and measure them in the following. $LnExperience\_analyst$ is the natural logarithm of the average work experience among all analysts following the same firm in a year, and the work experience of each analyst is measured as the number of years elapsed from the first forecast in his/her career to the latest forecast. $Past\_accuracy$ is the average forecast accuracy over the prior 5 years among all analysts following the same firm in a year. Brown (2001) and Kim et al. (2011) offer evidence that the accuracy of analyst forecasts in a year is positively correlated with that of their forecasts in the previous years.

We measure the analyst's available resources by using the size of her/his research portfolio ($LnPortfolio\_analyst$), the size of her/his affiliated brokerage firm ($Broker\_size$), and her/his social networks with other analysts ($Network\_analyst$). $LnPortfolio\_analyst$ is calculated as the natural logarithm of the average size (measured by total assets) of firms in the portfolios covered by analysts following the same firm in a year. Analysts holding a large research portfolio tend to be skillful and experienced in acquiring and processing information for their forecasts (e.g., He et al., 2024). $Broker\_size$ is computed as the average assets of brokerage houses employing analysts who follow the same firm in a year. Analysts affiliated with larger brokerage houses often have access to more information and more abundant resources, thus facilitating their better forecasting for firms. $Network\_analyst$ is a composite measure of the social network centrality of an analyst for a year, which is derived by using a common factor analysis of three social network centrality variables: degree centrality (the number of direct connections of an analyst to other analysts), closeness centrality (an inverse measure of the average of the minimum network distances between an analyst and each of the other analysts), and eigenvector centrality (the eigenvector of the largest eigenvalue of the non-negative adjacency matrix of the analyst's social network). It is suggested that analysts who have strong social networks with other analysts are better positioned to access more valuable information, thereby enhancing their forecast performance (Cao and Liang, 2024).

[Insert Table 7 about here]

We divide our sample into two subsamples based on the median values of each analyst characteristic, and re-estimate Model (3) for each subsample. The results are presented in Table 7. The coefficients for $Digit$ in Columns (1), (3), (5), (7), and (9) referring to analysts with greater work capabilities and more information resources, are consistently higher in magnitude than those in Columns (2), (4), (6), (8), and (10). Statistical tests confirm significant differences in the coefficients of $Digit$ for each pair of subsamples. These results corroborate that the positive impact of firms' digital transformation on analyst forecast accuracy is more pronounced among analysts with greater work capabilities — characterized by longer work experience and more accurate past forecasts — and among those with more information resources — featured by a larger size of their research portfolio, a larger size of their affiliated brokerage house, and their stronger social networks.

6.2 The moderating effect of corporate innovation

Innovation and digital transformation both involve new technologies and could be interrelated in promoting the development of enterprises (Peng and Tao, 2022). In this regard, we explore whether corporate innovation moderates the effect of firms' digital transformation ($Digit$) on analyst forecast accuracy ($Accuracy$). We conjecture that the effect is more prominent if analysts follow less innovative firms for three reasons. First, firms with lower levels of innovation typically have traditional business models and operational processes that are associated with relatively lower operational efficiency. Digital transformation can compensate for their lack of innovation by rapidly enhancing operational efficiency, enabling analysts to leverage information from digitalized processes to predict future firm performance more accurately. Second, firms with lower innovation capabilities are often slower in market response and adaptability. Digital transformation enhances their ability to quickly adjust investment strategies and improve business decision-making in response to changing external environments, thereby making their future performance more predictable for analysts. Third, firms with a high degree of innovation tend to protect proprietary business information associated with innovation, potentially limiting the role of digital transformation in enhancing information quality for firms.

We construct four measures of corporate innovation: (i) the number of invention patents applied by a firm in a year and subsequently granted by the China National Intellectual Property Administration ($Invention$); (ii) the number of product-modelling patents and product-design patents applied by a firm in a year and subsequently granted by the China National Intellectual Property Administration ($Modelling\,\&\,design$);[9] (iii) research and development (R&D) expenditures incurred by a firm in a year, scaled by sales revenue ($R\&D\ expenditure$); and (iv) the number of R&D employees for a firm in a year ($R\&D\ staff$).

[Insert Table 8 about here]

We partition our sample into two subsamples with high and low corporate innovation based on the median values of $Invention$, $Modelling\,\&\,design$, $R\&D\ expenditure$, and $R\&D\ staff$, respectively, and re-estimate Model (3) for each pair of subsamples. Table 8 reports the regression results. The coefficients of $Digit$ in Columns (1), (3), (5), and (7) for the low-innovation subsamples are smaller in magnitude than those in Columns (2), (4), (6), and (8) for the high-innovation subsamples; the differences in coefficients of $Digit$ are all statistically significant. These findings are consistent with our notion that the positive impact of firms' digital transformation on analyst forecast accuracy is more prominent for less innovative firms.

[9] Invention patents (product-modelling patents and product-design patents) could be classified as the outcomes of radical (incremental) innovation by firms.

6.3 The moderating effect of industry-level technology development

High-technology (hereafter, high-tech) firms typically have well-developed absorptive ability — the ability to identify, assimilate and apply knowledge (e.g., Cohen and Levinthal, 1990), and possess R&D capabilities as fundamental components of their business models (e.g., Henderson and Clark, 1990; Tushman and Anderson, 1986). As such, the foregoing costs and risks associated with undertaking digitalization are lower for high-tech firms than for non-high-tech firms. Hence, we conjecture that the effect of firms' digital transformation on analyst forecast accuracy is more prominent for firms operating in high-tech industries.

To test the supposition, we divide our sample into two subsamples, based on whether the firm operates in high-tech industries ($High\_tech$) and re-estimate Model (3) for each subsample.[10] The regression results, presented in Table 9, show that the coefficient of $Digit$ in Columns (1) for high-tech industries is significantly larger in magnitude than that in Columns (2) for non-high-tech industries. These findings support our conjecture and, more importantly, address a plausible concern that our baseline results are driven primarily by high-tech firms, which account for 35% of our sample. The results for the non-high-tech firms are qualitatively consistent with those obtained from the full sample.

[Insert Table 9 about here]

7. Conclusion

This study provides evidence that firm digitalization enhances the accuracy of analyst forecasts by improving corporate operational efficiency, investment efficiency and information quality. Our findings hold across multiple robustness checks, including utilization of alternative measures for key variables, two-stage least squares regression and difference-in-differences models. We also find that corporate digitalization has a more positive impact on the accuracy of forecasts by analysts with greater work capabilities or more information resources. Moreover, firms with lower innovation levels benefit more from digitalization in terms of analyst forecast accuracy, suggesting that digital transformation may compensate for a lack of innovation-driven insights or efforts. Additionally, firms in high-tech industries could leverage digital

[10] As with prior research (e.g., Liu & Buck, 2007; Shen et al., 2024), we define high-tech firms are those operating in eight high-tech industries, as outlined in the "Classification of High-tech Industries" by the China National Bureau of Statistics. These industries include: chemical material and chemical product manufacturing industry; medical and pharmaceutical manufacturing industry; chemical fiber manufacturing industry; railway, shipbuilding, aircraft, and other transportation equipment manufacturing industry; computer, communication, and other electronic equipment manufacturing industry; instrument manufacturing industry; information transmission, software and information technology services industry; and scientific research and technology services industry.

transformation more effectively to enhance their business operations, resulting in more accurate forecasts by analysts. In a nutshell, our study provides empirical evidence from a representative market, where digitalization is rapidly evolving and influencing corporate practices, and offers insights into the implications of digitalization for future firm prospect via the lens of financial analysts.

We acknowledge that our study has certain limitation, particularly the lack of data to analyze how and to what degree each specific digital tool is utilized by firms for their business activities. Should these data be available, researchers could further explore how various digital technologies might interact with one another in shaping future firm performance.

References

Acemoglu, D., & Restrepo, P., (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review, 108* (6), 1488-1542.

Achim, M. V., Văidean, V. L., Popa, A. I. S., & Safta, L. I. (2022). The impact of corporate governance on the digitalization process: empirical evidence for the Romanian companies. *Digital Finance, 4*(4), 313-340.

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Artificial intelligence: the ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives, 33*, 31-50.

Agarwal, R., Gao, G., DesRoches, C., & Jha, A. (2010). The digital transformation of healthcare: Current status and the road ahead. *Information Systems Research, 21*, 796-809.

Ashraf, M. (2024). Does automation improve financial reporting? Evidence from internal controls. *Review of Accounting Studies*, 1-44.

Audretsch, D. B., & Belitski, M. (2024). Digitalization, resource mobilization and firm growth in emerging industries. *British Journal of Management, 35*(2), 613-628.

Blanco, B., Dhole, S., & Gul, F. A. (2023). Financial statement comparability and accounting fraud. *Journal of Business Finance & Accounting, 50*(7-8), 1166-1205.

Blichfeldt, H., & Faullant, R. (2021). Performance effects of digital technology adoption and product & service innovation–a process-industry perspective. *Technovation, 105*, 102275.

Bradley, D., Gokkaya, S., & Liu, X. (2017). Before an analyst becomes an analyst: Does industry experience matter? *Journal of Finance*, *72*(2), 751-792.

Brown, L. D. (2001). How important is past analyst forecast accuracy? *Financial Analysts Journal*, *57*(6), 44-49.

Canace, T., Li, J., & Ma, T. (2024). Analyst following and R&D investment. *Review of Accounting Studies, 29*(3), 2688-2723.

Cao, S., & Liang, C. (2024). Analyst collaboration networks and earnings forecast performance. *International Review of Financial Analysis, 93*, 103138.

Chahine, S., Daher, M., & Saade, S. (2021). Doing good in periods of high uncertainty: Economic policy uncertainty, corporate social responsibility, and analyst forecast error.

Journal of Financial Stability, 56, 100919.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 1165-1188.

Chen, H. A., Francis, B. B., Shen, Y. V., & Wu, Q. (2024). The impact of hedge fund activism on audit pricing. *The British Accounting Review, 56*(2), 101264.

Chen, T., Harford, J., & Lin, C. (2015). Do analysts matter for governance? Evidence from natural experiments. *Journal of Financial Economics, 115*(2), 383-410.

Chen, Z., & Jiang, K. (2024). Digitalization and corporate investment efficiency: Evidence from China. *Journal of International Financial Markets, Institutions and Money, 91*, 101915.

Chen, W., & Srinivasan, S. (2024). Going digital: Implications for firm value and performance. *Review of Accounting Studies, 29*, 1619-1665.

Chen, N., Sun, D., & Chen, J. (2022). Digital transformation, labour share, and industrial heterogeneity. *Journal of Innovation & Knowledge, 7*(2), 100173.

Cheng, Q., Goh, B. W., & Kim, J. B. (2018). Internal control and operational efficiency. *Contemporary Accounting Research, 35*(2), 1102-1139.

Cheng, W., Li, C., & Zhao, T. (2024). The stages of enterprise digital transformation and its impact on internal control: Evidence from China. *International Review of Financial Analysis, 92*, 103079.

Clement, M. B., & Tse, S. Y. (2005). Financial analyst characteristics and herding behavior in forecasting. *Journal of Finance*, *60*(1), 307-341.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly, 35*(1), 128–152.

Correani, A., De Massis, A., Frattini, F., Petruzzelli, A. M., & Natalicchio, A. (2020). Implementing a digital strategy: Learning from the experience of three digital transformation projects. *California Management Review, 62*(4), 37-56.

Datta, S., Iskandar-Datta, M., & Sharma, V. (2011). Product market pricing power, industry concentration and analysts' earnings forecasts. *Journal of Banking & Finance, 35*(6), 1352-1366.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting earnings management. *The*

Accounting Review, 70, 193-225.

Dorantes, C. A., Li, C., Peters, G. F., & Richardson, V. J. (2013). The effect of enterprise systems implementation on the firm information environment. *Contemporary Accounting Research, 30*(4), 1427-1461.

Elberry, N., & Hussainey, K. (2020). Does corporate investment efficiency affect corporate disclosure practices? *Journal of Applied Accounting Research, 21*(2), 309-327.

Flyvbjerg B., & Budzier A. (2011). Why your IT project may be riskier than you think. *Harvard Business Review*, 89 (9), 601-603

Goldfarb, A., & Tucker, C. (2019). Digital economics. *Journal of Economic Literature, 57*(1), 3–43.

Gu, F., & Wang, W. (2005). Intangible assets, information complexity, and analysts' earnings forecasts. *Journal of Business Finance & Accounting, 32*(9-10), 1673-1702.

Guo, X., Li, M., Wang, Y., & Mardani, A. (2023). Does digital transformation improve the firm's performance? From the perspective of digitalization paradox and managerial myopia. *Journal of Business Research, 163*, 113868.

He, G., & Li, A. Z. (2024). The roles of financial analysts in the stock market, in: Lee, C. F., Lee, A. C., Lee, J. C. (Eds.) *Handbook of Investment Analysis, Portfolio Management, and Financial Derivatives*. World Scientific, pp. 2293-2308.

He, G., Ren, H. M., & Taffler, R. (2020). The impact of corporate tax avoidance on analyst coverage and forecasts. *Review of Quantitative Finance and Accounting, 54*(2), 447-477.

He, G., Sun, Y., & Li, A. Z. (2024). Does analysts' industrial concentration affect the quality of their forecasts? *Financial Markets and Portfolio Management, 38*(1), 37-91.

He, J. J., & Tian, X. (2013). The dark side of analyst coverage: The case of innovation. *Journal of Financial Economics, 109*(3), 856-878.

Henderson, R. M., & Clark, K. B. (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative Science Quarterly, 35*(1), 9–30.

Hou, Q., Li, W., Teng, M., & Hu, M. (2022). Just a short-lived glory? The effect of China's anti-corruption on the accuracy of analyst earnings forecasts. *Journal of Corporate Finance, 76*, 102279.

Hu, Y. (2019). Short-horizon market efficiency, order imbalance, and speculative trading: evidence from the Chinese stock market. *Annals of Operations Research, 281*(1), 253-274.

Huang, F., Li, H., & Wang, T. (2018). Information technology capability, management forecast accuracy, and analyst forecast revisions. *Accounting Horizons, 32*(3), 49-70.

Huang, M. H., Rust, R. T., 2018. Artificial intelligence in service. *Journal of Service Research, 21*(2), 155-172.

Huang, C. K., Wang, T., & Huang, T. Y. (2020). Initial evidence on the impact of big data implementation on firm performance. *Information Systems Frontiers, 22*(2), 475-487.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*(1), 1-24.

Iansiti, M., & Lakhani, K. R. (2017). The truth about blockchain. *Harvard Business Review, 95*(1), 118-127.

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research*, 70, 338-345.

Jiang, K., Zhou, M., & Chen, Z. (2024). Digitalization and firms' systematic risk in China. *International Journal of Finance & Economics*. https://doi.org/10.1002/ijfe.2931.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*(1), 141-151.

Kim, Y., Lobo, G. J., & Song, M. (2011). Analyst characteristics, timing of forecast revisions, and analyst forecasting ability. *Journal of Banking & Finance, 35*(8), 2158-2168.

Lam, H. K., Yeung, A. C., & Cheng, T. E. (2016). The impact of firms' social media initiatives on operational efficiency and innovativeness. *Journal of Operations Management, 47*, 28-43.

Lang, M. H., & Lundholm, R. J. (1996). Corporate disclosure policy and analyst behavior. *The Accounting Review, 71*(4), 467-492.

Lehavy, R., Li, F., & Merkley, K. (2011). The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review, 86*(3), 1087-1115.

Lem, K. W. (2024). Data analytics strategy and internal information quality. *Contemporary*

Accounting Research, 41(2), 1376-1410.

Li, C., Lin, A. P., & Lu, H. (2023). The effect of social skills on analyst performance. *Contemporary Accounting Research, 40*(2), 1418-1447.

Liu, X., & Buck, T. (2007). Innovation performance and channels for international technology spillovers: Evidence from Chinese high-tech industries. *Research Policy*, *36*(3), 355-366.

Liu, D., Lin, T., Chen, C. R., & Feng, W. (2024). Air pollution, analyst information provision, and stock price synchronicity. *Review of Quantitative Finance and Accounting,* 1-49.

Madsen, J. M., & McMullin, J. L. (2020). Economic consequences of risk disclosures: Evidence from crowdfunding. *The Accounting Review, 95*(4), 331-363.

Mansi, S. A., Maxwell, W. F., & Miller, D. P. (2011). Analyst forecast characteristics and the cost of debt. *Review of Accounting Studies, 16*, 116-142.

McKinsey & Company. (2018). Unlocking success in digital transformations. https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/unlocking-success-in-digital-transformations (access 29th October 2024).

Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big data analytics capabilities and innovation: the mediating role of dynamic capabilities and moderating effect of the environment. *British Journal of Management, 30*(2), 272-298.

Mikalef, P., & Pateli, A. (2017). Information technology-enabled dynamic capabilities and their indirect effect on competitive performance: Findings from PLS-SEM and fsQCA. *Journal of Business Research, 70*, 1-16.

Niebel, T., Rasel, F., & Viete, S. (2019). BIG data–BIG gains? Understanding the link between big data analytics and innovation. *Economics of Innovation and New Technology, 28*(3), 296-316.

Paiola, M., & Gebauer, H. (2020). Internet of things technologies, digital servitization and business model innovation in BtoB manufacturing firms. *Industrial Marketing Management, 89*, 245-264.

Peng, Y., & Tao, C. (2022). Can digital transformation promote enterprise performance? — From the perspective of public policy and innovation. *Journal of Innovation & Knowledge, 7*(3), 100198.

Pohl, R. V., Lei, L., & Niedzwiecki, M. (2022). Matching Methods for the Evaluation of

Section 1115 Demonstrations. *Washington, DC: Mathematica*, November.

Richardson, S. (2006). Over-investment of free cash flow. *Review of Accounting Studies, 11,* 159-189.

Ryu, H. S., & Lee, J. N. (2018). Understanding the role of technology in service innovation: Comparison of three theoretical perspectives. *Information & Management, 55*(3), 294-307.

Shen, H., Gao, Y., Cheng, X., & Wang, Q. (2024). The impact of the US export controls on Chinese firms' innovation: Evidence from Chinese high-tech firms. *International Review of Financial Analysis, 95*, 103510.

Sun, S., Cegielski, C. G., Jia, L., & Hall, D. J. (2018). Understanding the factors affecting the organizational adoption of big data. *Journal of Computer Information Systems, 58*(3), 193-203.

Teece D, Pisano G, & Shuen A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal, 18*(7): 509-533.

Tian, G., Li, B., & Cheng, Y. (2022). Does digital transformation matter for corporate risk-taking? *Finance Research Letters, 49*, 103107.

Tu, W., & He, J. (2023). Can digital transformation facilitate firms' M&A: Empirical discovery based on machine learning. *Emerging Markets Finance and Trade*, *59*(1), 113-128.

Tushman, M. L., & Anderson, P. (1986). Technological discontinuities and organizational environments. *Administrative Science Quarterly, 31*(3), 439–465.

Vogelsang, K., Liere-Netheler, K., Packmohr, S., & Hoppe, U. (2018). Success factors for fostering a digital transformation in manufacturing companies. *Journal of Enterprise Transformation, 8*(1-2), 121-142.

Wei, X., Jiang, F., & Yang, L. (2023). Does digital dividend matter in China's green low-carbon development: Environmental impact assessment of the big-data comprehensive pilot zones policy. *Environmental Impact Assessment Review, 101*, 107143.

Wen, H., Zhong, Q., & Lee, C. C. (2022). Digitalization, competition strategy and corporate innovation: Evidence from Chinese manufacturing listed companies. *International Review of Financial Analysis, 82*, 102166.

Wu, L., Hitt, L., & Lou, B. (2020). Data analytics, innovation, and firm productivity.

Management Science, 66(5), 2017-2039.

Wu, F., Hu, H., Lin, H., & Ren, X. (2021). Enterprise digital transformation and capital market performance: Empirical evidence from stock liquidity. *Management World, 37*(07), 130-144.

Yasmin, M., Tatoglu, E., Kilic, H. S., Zaim, S., & Delen, D. (2020). Big data analytics capabilities and firm performance: An integrated MCDM approach. *Journal of Business Research, 114*, 1-15.

Yermack, D. (2017). Corporate governance and blockchains. *Review of Finance, 21*(1), 7-31.

Zeng, H., Ran, H., Zhou, Q., Jin, Y., & Cheng, X. (2022). The financial effect of firm digitalization: Evidence from China. *Technological Forecasting and Social Change, 183*, 121951.

Zhou, K. Z., Su, C., & Bao, Y. (2002). A paradox of price–quality and market efficiency: a comparative study of the US and China markets. *International Journal of Research in Marketing, 19*(4), 349-365.

Appendix A: Variable definitions

Variables	Definitions
$Accuracy$	the average forecast accuracy among all analysts following the same firm in a year. The forecast accuracy of each analyst is measured as -1 times the ratio of the absolute difference between the actual and forecasted earnings per share (EPS) to the closing stock price on the trading day preceding the analyst's forecast. In cases where an analyst releases multiple forecasts for the same firm within a year, the final forecast issued one month before the end of the fiscal year is utilized for the variable construction.
$Accuracy1$	the average forecast accuracy among all analysts following the same firm in a year. The forecast accuracy of each analyst is measured as -1 times the ratio of the absolute difference between the actual and forecasted EPS to the actual EPS. In cases in which an analyst releases multiple forecasts for the same firm within a year, the final forecast issued at least one month before the end of the fiscal year is utilized for the variable construction.
$Optimism$	the average forecast optimism among all analysts following the same firm in a year. For each analyst, if his/her forecasted EPS is larger than the actual EPS, the forecast optimism is measured as the ratio of the absolute difference between the actual and forecasted earnings per share to the closing stock price on the trading day before the forecast was issued, and 0 otherwise. If an analyst issues multiple forecasts for the same firm within the year, the most recent forecast released at least one month before the fiscal year-end is utilized for the variable construction.
$Pessimism$	the average forecast pessimism among all analysts following the same firm in a year. For each analyst, if his/her forecasted EPS is lower than the actual EPS, the forecast pessimism is measured as the ratio of the absolute difference between the actual and forecasted EPS to the closing stock price on the trading day preceding the analyst's forecast, and 0 otherwise. If an analyst releases multiple forecasts for the same firm within a year, the final forecast issued at least one month before the end of the fiscal year is utilized for the variable construction.
$Digit$	the level of digital transformation for a firm in a year, estimated by using common factor analysis to integrate variables - (i) the natural logarithm of one plus digitalization-related intangible assets disclosed in the firm's annual report and (ii) the natural logarithm of one plus the word frequency of digitalization-related information disclosed in the Management Discussion and Analysis (MD&A) section of the firm's annual report, into a composite index.
$Digit1$	the level of digital transformation for a firm in a year, estimated as the natural logarithm of one plus digitalization-related intangible assets disclosed in the firm's annual report.
$Digit2$	the level of digital transformation for a firm in a year, estimated as the natural logarithm of one plus the word frequency of digitalization-related information disclosed in the Management Discussion and Analysis (MD&A) section of the firm's annual report.
$Digit3$	the relative level of digital transformation for a firm in a year, estimated as the difference between a firm's digital transformation and the average digital transformation among firms in the same industry in a year, divided by the standard deviation of digital transformation among firms in the same industry in a year.
$Digit4$	the relative level of digital transformation for a firm in a year, estimated as the difference between a firm's digital transformation and the average digital transformation among firms in the same city in a year, divided by the standard deviation of digital transformation among firms in the same city in a year.
$Digit5$	the relative level of digital transformation for a firm in a year, estimated as the difference between a firm's digital transformation and the average value of digital transformation among firms in the same province in a year, divided by the standard deviation of digital transformation among firms in the same province in a year.
$LnSize$	the natural logarithm of total assets for a firm in the prior year.

$Leverage$ the ratio of total liabilities to total assets for a firm in the prior year.

ROA the ratio of profits before taxes and interests to total assets for a firm in the prior year.

$Sale\_growth$ the difference between sales revenue in the prior year and that in the year before last, divided by the sales revenue in the year before last, for a firm.

$Earnings\_surprise$ the ratio of the difference between a firm's actual EPS in this year and that in the prior year to the firm's closing stock price on the final trading day of the prior year.

$Return\_volatility$ the standard deviation of daily stock returns for a firm in the prior year.

$Cash\_ratio$ the ratio of cash and cash equivalents to total current liabilities for a firm in the prior year.

$Director\_shares$ the ratio of the number of shares held by the directors on the board to the total shares outstanding for a firm in the prior year.

$Insti$ the ratio of the number of shares held by institutional investors to the total shares outstanding for a firm in the prior year.

SOE 1 if the largest ultimate shareholder of a firm in the prior year pertains to a government entity, and 0 otherwise.

Big_4 1 if the financial statement released by a firm in the prior year is audited by a Big-4 auditor (i.e., Deloitte, PricewaterhouseCoopers, KPMG or Ernst & Young), and 0 otherwise.

$LnAge$ the natural logarithm of one plus the number of years, since the firm got listed on the stock market in mainland China, as of the end of the prior year.

$LnAnalyst\_coverage$ the natural logarithm of the number of analysts who release forecasts for a firm one month before the end of a fiscal year.

$Education\_analyst$ the average level of education among all analysts following the same firm in a year. The educational level of each analyst is represented by the values 5, 4, 3, 2, and 1 for doctoral, master's, bachelor's, associate degrees, and lower degrees, respectively.

$Star\_analyst$ the proportion of star analysts among all analysts following the same firm in a year.

$LnExperience\_analyst$ the natural logarithm of average work experience among all analysts following the same firm in a year. Work experience of each analyst is measured as the number of years elapsed from his/her initial forecast to the latest forecast.

$LnFollowing\_analyst$ the natural logarithm of average size of research portfolios of all analysts following the same firm in a year. The size of a portfolio covered by each analyst in a year is measured as the number of firms in the portfolio.

$LnHorizon$ the natural logarithm of the average forecast horizon of all analysts following the same firm in a year. The forecast horizon of each analyst is measured as the number of days between an analyst's forecast issuance date and the fiscal year end date.

$Internet$ the level of internet development for a city in a year, calculated by utilizing common factor analysis to integrate variables – (i) the number of internet users per one hundred people, (ii) the proportion of employees in the computer service and software industry to the total employment, (iii) per capita total telecommunications output, and (iv) the number of mobile phone users per one hundred people, into a composite index.

$Digit\_same\_city$ the average level of digital transformation among all other firms located in the same city for a firm in a year.

$Treatment$ a dummy variable that equals 1 when the firm is located in one of China's big-data comprehensive pilot zones (Guizhou, Beijing–Tianjin–Hebei region, Guangdong, Henan, Shanghai, Inner Mongolia, Chongqing, and Shenyang), and 0 otherwise.

$Post$ a dummy variable that equals 1 (0) for the 3-year period from 2017 to 2019 (2013 to 2015) post (before) the announcement of establishments of China's big-data comprehensive pilot zones in 2016.

$Accruals\_quality$ accruals quality for a firm in a year, measured as -1 times the absolute value of abnormal accruals estimated based on the modified Jones model (Dechow et al., 1995).

$MF\_accuracy$ management forecast accuracy, measured as the negative of the natural logarithm

	of the absolute difference between management earnings forecast and actual earnings for a fiscal year, scaled by the stock price at the beginning of the fiscal year. If a manager makes multiple forecasts for the same year, the final forecast is utilized for the variable construction.
$Investment\_effi$	investment efficiency of a firm in a year, measured as -1 times the absolute value of the residual of a regression model developed by Richardson (2006) and using Tobin's Q as the proxy for growth opportunities.
$Operational\_effi$	operational efficiency for a firm in a year, evaluated by using the Data Envelopment Analysis (DEA) to construct an efficient frontier of production through an optimization program maximizing the output-to-input ratio (Cheng et al., 2018). Production output is quantified by sales revenues, while production inputs include the number of employees, costs of goods sold, and capital expenditures (Lam et al., 2016). For firms within the same industry and year, those on the efficient frontier are assigned a value of one, while less efficient firms behind the frontier receive efficiency scores between 0 and 1.
$Past\_accuracy$	the average forecast accuracy over the prior 5 years among all analysts following the same firm in a year. The average is taken both on the prior 5 years and on all analysts.
$LnPortfolio\_analyst$	the natural logarithm of the average size (measured by total assets) of firms in portfolios covered by all analysts following the same firm in a year. The average is taken both on the portfolio firms' total assets and on all analysts.
$Broker\_size$	the average size of brokerage houses employing analysts who follow the same firm in a year, with the size measured as the brokerage house's total assets.
$Network\_analyst$	a composite measure of the social network centrality of an analyst for a year, which is derived by using a common factor analysis of three social network centrality variables: degree centrality (the number of direct connections of an analyst to other analysts), closeness centrality (an inverse measure of the average of the minimum network distances between an analyst and each of the other analysts), and eigenvector centrality (the eigenvector of the largest eigenvalue of the non-negative adjacency matrix of the analyst's social network).
$Invention$	the number of invention patents applied by a firm in a year and subsequently granted by the China National Intellectual Property Administration.
$Modelling\ \&\ design$	the number of product-modelling and product-design patents applied by a firm in a year and subsequently granted by the China National Intellectual Property Administration.
$R\&D\ expenditure$	research and development (R&D) expenditure incurred by a firm in a year, scaled by sales revenue.
$R\&D\ staff$	the number of R&D employees for a firm in a year.
$High\_tech$	1 if a firm operates in the following high-tech industries: chemical material and chemical product manufacturing industry; medical and pharmaceutical manufacturing industry; chemical fiber manufacturing industry; railway, shipbuilding, aircraft, and other transportation equipment manufacturing industry; computer, communication, and other electronic equipment manufacturing industry; instrument manufacturing industry; information transmission, software and information technology services industry; and scientific research and technology services industry, and 0 otherwise.

Appendix B: Sample selection procedure

	Number of Firm-year observations/ firms	Number of Firm-year observations/ firms
Firm-year observations for companies listed on the Shenzhen or Shanghai Stock Exchange in the years 2011–2022		39,111
Less: observations if the firm is designated as ST, *ST or PT		(1,658)
Less: observations if the firm is in the finance industry		(1,066)
Less: observations with missing values in analyst forecast accuracy		(14,906)
Less: observations with missing values in analyst educational background		(646)
Less: observations with missing values in firm-level control variables		(1,693)
- observations without values of sales growth	(13)	
- observations without values of institutional holdings	(23)	
- observations without values of state ownership	(397)	
- observations without values of audit quality	(26)	
- observations without values of earnings surprise	(551)	
- observations without values of stock return volatility	(121)	
- observations without values of director shares ownership	(562)	
Final firm-year observations		**19,142**
Unique companies		3,534
- Unique companies on the Shenzhen Stock Exchange		2,041
- Unique companies on the Shanghai Stock Exchange		1,493

Appendix C: Common factor analysis

Panel A: Common factor analysis: Eigen-values

$Factor\ 1$	$Factor\ 2$
1.1925	0.8075

Panel B: Common factor analysis: Summary for factor analysis

Variables	Loadings on the first factor	Correlation with the first factor	Communality
$Digit1$	0.7722	0.7735	0.5962
$Digit2$	0.7722	0.7702	0.5962

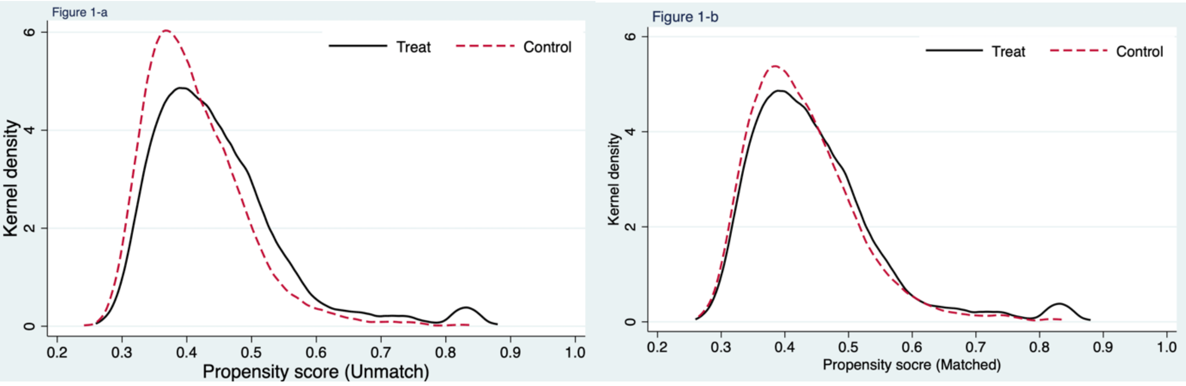
Notes: Panel A of this table reports the eigenvalues for two factors that mimic the correlation matrix of *Digit1* and *Digit2* --- the two individual measures of firms' digital transformation. $Digit1$ is the natural logarithm of one plus digitalization-related intangible assets disclosed in a firm's annual report; $Digit2$ is the natural logarithm of one plus the frequency of digitalization-related words in the management discussion and analysis (MD&A) section of a firm's annual report. *Factor1* and *Factor2* are the common factors obtained by using common factor analysis of the two measures of corporate digitalization. Panel B presents the loadings on the first factor, the correlation of the two digitalization variables with the first factor, and the communality of the two measures of digitalization.

Appendix D: Tests of covariate balance and common support for propensity-score matching

Panel A: Multivariate test of covariate balance for propensity-score matching

Variables	(1)	(2)
	Treatment	
	Un-match sample	Matched sample
LnSize	0.2572***	-0.0302
	(11.45)	(-1.15)
ROA	-0.6289	0.4763
	(-1.35)	(0.91)
Sale_growth	0.0733	-0.0092
	(1.23)	(-0.13)
Director_shares	1.2873***	0.0246
	(9.78)	(0.17)
Earnings_surprise	0.5571	-0.4606
	(0.95)	(-0.70)
Return_volatility	6.2779***	0.7245
	(2.98)	(0.31)
Cash_ratio	0.0598***	-0.0073
	(3.98)	(-0.43)
R&D expenditure	0.0486***	-0.0039
	(8.15)	(-0.56)
INTERCEPT	-7.3505***	0.1072
	(-12.76)	(0.16)
Industry dummies	included	included
Year dummies	included	included
N	8,668	6,566
Pseudo R^2	0.0471	0.0191

Panel B: Kernel density distribution of propensity scores for the treatment and control groups



Notes: This table reports results for the test of covariate balance and common support for propensity-score matching. Panel A shows the results of logit regression of *Treatment* on seven matching covariates: *LnSize*, *ROA*, *Sale_growth*, *Director_shares*, *Earnings_surprise*, *Return_volatility*, *Cash_ratio*, and *R&D expenditure*. Industry dummies (constructed based on the China Securities Regulatory Commission (CSRC) Industry Classification Codes) and year dummies are included in both regressions, but their results are not reported for brevity. The sample period ranges from 2011 to 2022. The treatment indicator variable, *Treatment*, equals 1 if the firm is headquartered in one of the China's big-data analytics pilot zones (i.e., Guizhou, Beijing–Tianjin–Hebei region, Guangdong, Henan, Shanghai, Inner Mongolia, Chongqing, and Shenyang), and 0 otherwise. Definitions of other variables are provided in Appendix A. Column (1) (Column (2)) show the logit regression results for the un-matched (matched) sample before (post) propensity-score matching. Panel B displays the distribution, in the form of kernel density curve, of propensity scores for the treatment group and control group before and after propensity score matching. The horizontal axis represents the propensity scores, and the vertical axis represents the probability density. The left (right) figure shows the distributions of propensity scores before (after) propensity score matching. The solid black (dashed red) curves represent the distribution of propensity scores for the treatment (control) firms.

Panel A: Descriptive statistics

Variables	N	Mean	Std.	P10	P25	P50	P75	P90
$Accuracy$	19,142	-0.0166	0.0220	-0.0346	-0.0181	-0.0098	-0.0055	-0.0031
$Digit$	19,142	-0.0022	0.9953	-1.4255	-0.6932	0.0969	0.7418	1.2248
$Digit1$	19,142	12.1380	6.4031	0	11.9925	14.6983	16.2128	17.5222
$Digit2$	19,142	2.2072	1.3536	0	1.0986	2.1972	3.1781	4.0073
$LnSize$	19,142	22.3632	1.3310	20.8117	21.3820	22.1642	23.1323	24.1885
$Leverage$	19,142	0.4091	0.2011	0.1423	0.2455	0.4027	0.5611	0.6853
ROA	19,142	0.0712	0.0520	0.0205	0.0395	0.0644	0.0964	0.1364
$Sale\_growth$	19,142	0.2276	0.3933	-0.0942	0.0279	0.1535	0.3214	0.5667
$Earnings\_surprise$	19,142	-0.0063	0.0399	-0.0404	-0.0136	0	0.0076	0.0209
$Return\_volatility$	19,142	0.0320	0.0159	0.0192	0.0231	0.0281	0.0350	0.0481
$Cash\_ratio$	19,142	1.0633	1.8766	0.1184	0.2172	0.4374	1.0086	2.3972
$Director\_shares$	19,142	0.1369	0.1898	0	0	0.0096	0.2638	0.4505
$Insti$	19,142	0.4831	0.2568	0.1021	0.2659	0.5135	0.6929	0.8066
SOE	19,142	0.3453	0.4755	0	0	0	1	1
Big_4	19,142	0.0745	0.2627	0	0	0	0	0
$LnAge$	19,142	2.1126	0.7671	1.0986	1.6094	2.1972	2.7726	3.0910
$LnAnalyst\_coverage$	19,142	1.7163	1.0328	0	1.0986	1.7918	2.5649	3.0445
$Education\_analyst$	19,142	4.0216	0.2563	3.8000	4	4	4.0909	4.2500
$Star\_analyst$	19,142	0.2000	0.2247	0	0	0.1667	0.3000	0.5000
$LnExperience\_analyst$	19,142	1.6308	0.3519	1.2040	1.4069	1.6463	1.8524	2.0626
$LnFollowing\_analyst$	19,142	3.4713	0.4630	2.9275	3.1891	3.4723	3.7598	4.0431
$LnHorizon$	19,142	5.0062	0.3819	4.5294	4.8028	5.0195	5.2575	5.5074

Notes: This table reports the descriptive statistics for main variables used in the baseline regression analysis. The sample period ranges from 2011 to 2022. Definitions for variables are provided in Appendix A.

Table 1. Univariate statistics

Panel B: Correlation matrix

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
(1) *Accuracy*	1																			
(2) *Digit*	0.0530***	1																		
(3) *LnSize*	-0.2620***	0.1171***	1																	
(4) *Leverage*	-0.2729***	-0.0010	0.5937***	1																
(5) *ROA*	0.1400***	0.0262***	-0.1309***	-0.3382***	1															
(6) *Sale_growth*	0.0151**	0.0425***	-0.0618***	-0.0008	0.2618***	1														
(7) *Earnings_surprise*	-0.0984***	0.0263***	0.0896***	0.1054***	-0.1729***	-0.0492***	1													
(8) *Return_volatility*	0.1467***	0.0589***	-0.3411***	-0.1691***	-0.0048***	0.0913***	-0.0162***	1												
(9) *Cash_ratio*	0.2568***	0.0387***	-0.4515***	-0.7390***	0.2894***	0.0279***	-0.1086***	0.1521***	1											
(10) *Director_shares*	0.0936***	0.1420***	-0.4380***	-0.3664***	0.1669***	0.1598***	-0.0866***	0.2330***	0.2454***	1										
(11) *Insti*	-0.0281***	-0.0436***	0.4084***	0.2402***	0.0654***	-0.0317***	0.0560***	-0.1755***	-0.1061***	-0.6638***	1									
(12) *SOE*	-0.0688***	-0.1090***	0.4189***	0.3421***	-0.1665***	-0.1335***	0.0704***	-0.2265***	-0.2008***	-0.6346***	0.4203***	1								
(13) *Big_4*	-0.0415***	0.0235***	0.2986***	0.1324***	-0.0004	-0.0463***	0.0455***	-0.1307***	-0.0867***	-0.1981***	0.2530***	0.1645***	1							
(14) *LnAge*	-0.1658***	-0.0086	0.5704***	0.4389***	-0.1961***	-0.1671***	0.1466***	-0.3223***	-0.3672***	-0.5476***	0.2416***	0.4939***	0.1059***	1						
(15) *LnAnalyst_coverage*	-0.0102	0.0853***	0.2242***	-0.0034	0.3657***	0.2211***	0.0703***	-0.0366***	0.0535***	0.0474***	0.2193***	-0.0266***	0.1332***	-0.0395***	1					
(16) *Education_analyst*	0.0344***	0.0301***	-0.0292***	-0.0681***	0.0394***	0.0202***	0.0037	0.0443***	0.0443***	0.0354***	-0.0368***	-0.0361***	-0.0434***	-0.0243***	0.0337***	1				
(17) *Star_analyst*	-0.0300***	0.0016	0.1106***	0.0697***	0.0450***	-0.0077	0.0360***	0.0113	-0.0740***	-0.0289***	0.0753***	0.0204***	0.0213***	0.0693***	0.2087***	0.0020	1			
(18) *LnExperience_analyst*	-0.0740***	0.2190***	0.2055***	0.0722***	0.0144**	-0.0596***	0.0588***	-0.0060	-0.1127***	0.0114	-0.0026	-0.0260***	0.0486***	0.1457***	-0.0301***	0.0055	0.1584***	1		
(19) *LnFollowing_analyst*	0.0594***	0.0695***	-0.0735***	-0.0527***	-0.0196***	0.0359***	0.0056	0.2579***	0.0192***	0.0590***	-0.0761***	-0.0656***	-0.0556***	-0.0382***	-0.0588***	0.0372***	0.1103***	0.2399***	1	
(20) *LnHorizon*	-0.0487***	-0.0654***	-0.0789***	-0.0121*	-0.0601***	-0.0004	-0.2293***	-0.0001	0.0010	-0.0201***	-0.0519***	0.0259***	-0.0457***	-0.0045	-0.1569***	0.0093	-0.0516***	-0.0539***	0.0071	1

Notes: This table reports Spearman correlation for the main variables used in the baseline regression analysis. The sample period ranges from 2011 to 2022. ***, **, and * indicate 1%, 5%, and 10% two-tailed statistical significance, respectively. Definitions for variables are provided in Appendix A.

Table 1. Univariate statistics

Variables	(1)	(2)
	$Accuracy$	
$Digit$	0.0015***	0.0007**
	(6.05)	(2.39)
$LnSize$	-0.0031***	0.0003
	(-9.68)	(0.49)
$Leverage$	-0.0147***	-0.0170***
	(-7.75)	(-6.82)
ROA	0.0752***	0.1156***
	(11.81)	(12.14)
$Sale\_growth$	0.0012***	0.0002
	(2.65)	(0.38)
$Earnings\_surprise$	-0.0376***	-0.0480***
	(-4.34)	(-5.64)
$Return\_volatility$	-0.0224**	-0.0331**
	(-2.13)	(-2.39)
$Cash\_ratio$	-0.0003***	-0.0003***
	(-3.02)	(-2.60)
$Director\_shares$	-0.0001	0.0078**
	(-0.06)	(2.51)
$Insti$	0.0037***	0.0022
	(3.20)	(1.18)
SOE	0.0020***	0.0001
	(3.25)	(0.07)
Big_4	0.0026**	0.0026**
	(2.45)	(2.24)
$LnAge$	-0.0012***	0.0023**
	(-3.24)	(2.12)
$LnAnalyst\_coverage$	0.0009***	0.0009***
	(4.27)	(3.33)
$Education\_analyst$	0.0006	0.0002
	(0.78)	(0.27)
$Star\_analyst$	-0.0022**	-0.0004
	(-2.35)	(-0.48)
$LnExperience\_analyst$	-0.0032***	-0.0011*
	(-5.29)	(-1.72)
$LnFollowing\_analyst$	0.0014***	0.0006
	(3.07)	(1.33)
$LnHorizon$	-0.0005	-0.0011**
	(-1.11)	(-2.13)
$INTERCEPT$	0.0537***	-0.0288*
	(6.50)	(-1.89)
$Industry\ dummies$	included	excluded
$Firm\ dummies$	excluded	included
$Year\ dummies$	included	included
N	19,142	18,594
$Adj.R^2$	0.1552	0.3083

Notes: This table reports the results of the impact of firms' digital transformation on analyst forecast accuracy. The dependent variable, $Accuracy$, is the average forecast accuracy among all analysts following the same firm for a year. The forecast accuracy of each analyst is measured as -1 times the ratio of the absolute difference between the actual and forecasted earnings per share (EPS) to the closing stock price on the trading day preceding the analyst's forecast. In cases where an analyst releases multiple forecasts for the same firm within a year, the final forecast issued one month before the end of the fiscal year is utilized for the variable construction. The main independent variable, $Digit$, is the level of firms' digital transformation, estimated by using common factor analysis to integrate variables - (i) the natural logarithm of one plus digitalization-related intangible assets disclosed in the firm's annual report ($Digit1$) and (ii) the natural logarithm of one plus the word frequency of digitalization-related information disclosed in the Management Discussion and Analysis (MD&A) section of the firm's annual report ($Digit2$), into a composite index. Columns (1) and (2) report the results of ordinary least square (OLS) regression and firm-fixed effect regression, respectively. Definitions of all variables are provided in Appendix A. The sample ranges from 2011 to 2022. Firm dummies, industry dummies (constructed based on the China Securities Regulatory Commission (CSRC) Industry Classification Codes) and year dummies are included where appropriate in the regressions, but their results are not reported for brevity. t-statistics in parentheses are based on robust standard errors clustered by firm. ***, **, and * indicate 1%, 5%, and 10% two-tailed statistical significance, respectively.

Table 2. The impact of firms' digital transformation on analyst forecast accuracy

Variables	(1)	(2)
	$Digit$	$Accuracy$
$Internet$	0.0480***	
	(3.18)	
$Digit\_samecity$	0.1283***	
	(3.64)	
$Pred\_Digit$		0.0148***
		(5.17)
$Control\ variables$	included	included
$Industry\ dummies$	included	included
$Year\ dummies$	included	included
N	11,375	11,375
$Adj.R^2$	0.2684	0.1322
F-stat. for instruments	14.33***	
Overidentification test (p-value)	0.9391	

Note: This table reports the results for the two-stage ordinary least squares (2SLS) regression of firms' digital transformation ($Digit$) on analyst forecast accuracy ($Accuracy$). Column (1) presents the result of the first-stage regression, in which the dependent variable is firms' digital transformation ($Digit$), and the key independent variable are two instrumental variables: first is the level of comprehensive development of internet ($Internet$) for a city for a year, calculated by utilizing common factor analysis to integrate variables – (i) the number of internet users per one hundred people, (ii) the proportion of employees in the computer service and software industry to the total employment, (iii) per capita total telecommunications output, and (iv) the number of mobile phone users per one hundred people, into a composite index; the second instrument is the average level of digital transformation among all other firms located in the same city ($Digit\_samecity$) for a firm for a year. Column (2) shows the result of the second-stage regression of the predicted values of firms' digital transformation (namely, $Pred\_Digit$, which is estimated from the first-stage regression) on analyst forecast accuracy ($Accuracy$). Definitions for all variables are provided in Appendix A. The sample ranges from 2011 to 2022. Industry dummies (constructed based on the China Securities Regulatory Commission (CSRC) Industry Classification Codes) and year dummies are included in all regressions, but their results are not reported for brevity. t-statistics in parentheses are based on robust standard errors clustered by firm. ***, **, and * indicate 1%, 5%, and 10% two-tailed statistical significance, respectively.

Table 3. 2SLS regression analysis

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Basic model		Propensity score matching		Entropy balancing		Coarsened exact matching	
	Test of parallel trends assumption	DID test	Test of parallel trends assumption	DID test	Test of parallel trends assumption	DID test	Test of parallel trends assumption	DID test
	$Accuracy$		$Accuracy$		$Accuracy$		$Accuracy$	
$Treatment * Year2013$	0.0006		0.0002		0.0005		-0.0014	
	(0.55)		(0.14)		(0.46)		(-0.69)	
$Treatment * Year2014$	0.0006		-0.0002		0.0003		0.0012	
	(0.60)		(-0.19)		(0.38)		(0.66)	
$Treatment * Year2015$	0.0014		0.0013		0.0015*		-0.0011	
	(1.44)		(1.33)		(1.65)		(-0.89)	
$Treatment * Year2017$	0.0022**		0.0022**		0.0023**		0.0045**	
	(2.51)		(2.19)		(2.56)		(2.45)	
$Treatment * Year2018$	0.0028***		0.0013		0.0024**		0.0009	
	(3.00)		(1.33)		(2.54)		(0.36)	
$Treatment * Year2019$	0.0041***		0.0038***		0.0041***		0.0026	
	(4.41)		(2.77)		(3.29)		(1.11)	
$Treatment$		0.0009		0.0004		0.0008		-0.0004
		(1.38)		(0.59)		(1.15)		(-0.35)
$Treatment * Post$		0.0021***		0.0020**		0.0021**		0.0032*
		(2.71)		(2.14)		(2.46)		(1.78)
$Control\ variables$	included	included	included	included	included	included	included	included
$Industry\ dummies$	included	included	included	included	included	included	included	included
$Year\ dummies$	included	included	included	included	included	included	included	included
N	9,787	9,787	6,566	6,566	8,668	8,668	997	997
$Adj. R^2$	0.1552	0.1553	0.1737	0.1737	0.1708	0.1709	0.3257	0.3261

Note: This table presents the difference-in-differences (DID) regression results for how the establishment of China's big-data comprehensive pilot zones, the exogenous shock to firms' digital transformation, affect analyst forecast accuracy ($Accuracy$). Columns (1), (3), (5), and (7) report the results from testing the parallel trends assumption based on the original sample as well as the samples formed based on propensity-score matching (PSM), entropy balancing, and coarsened exact matching (CEM), respectively. The treatment indicator variable, $Treatment$, equals 1 if the firm is located in one of the China's big-data comprehensive pilot zones (i.e., Guizhou, Beijing–Tianjin–Hebei region, Guangdong, Henan, Shanghai, Inner Mongolia, Chongqing, and Shenyang), and 0 otherwise. $Year2013$ ($Year2014$, $Year2015$, $Year2017$, $Year2018$, and $Year2019$) equal 1 for the year of 2013 (2014, 2015, 2017, 2018, and 2019), and 0 otherwise. $Treatment$ is omitted due to multicollinearity. Column (2) reports the basic DID regression result. Columns (4), (6), and (8) report the results of the matching-based DID model, which is estimated based on the propensity-score matched sample, entropy-balanced sample, and coarsened exact matched sample, respectively. The matching covariates are $LnSize$, ROA, $Sale\_growth$, $Director\_shares$, $Earnings\_surprise$, $Return\_volatility$, $Cash\_ratio$ and $R\&D\ expenditure$, which we expect not to statistically differ between the treatment and control firms post-matching. $Post$ equals 1 (0) for the 3-year period from 2017 to 2019 (2013 to 2015) post (before) the announcement of establishment of China's big-data comprehensive pilot zones in 2016. The interaction term, $Treatment * Post$, is the variable of interest which captures the impact of establishments of China's big-data comprehensive pilot zones on analyst forecast accuracy for the treated firms ($Treatment = 1$) relative to the control firms ($Treatment = 0$). Definitions for all the variables are provided in Appendix A. The sample ranges from 2013 to 2019 and excludes the year 2016 when the China's big-data comprehensive pilot zones were established. Industry dummies (constructed based on the China Securities Regulatory Commission (CSRC) Industry Classification Codes) and year dummies are included in all regressions, but their results are not reported for brevity. t-statistics in parentheses are based on robust standard errors clustered by firm. ***, **, and * indicate 1%, 5%, and 10% two-tailed statistical significance, respectively.

Table 4. Difference-in-differences regression (DID) analysis

Panel A: Alternative measures of firms' digital transformation

Variables	(1)	(2)	(3)	(4)	(5)
			Accuracy		
$Digit1$	0.0001*** (3.06)				
$Digit2$		0.0012*** (7.02)			
$Digit3$			0.0014*** (6.23)		
$Digit4$				0.0010*** (4.39)	
$Digit5$					0.0013*** (5.68)
Control variables	included	included	included	included	included
Industry dummies	included	included	included	included	included
Year dummies	included	included	included	included	included
N	19,142	19,142	19,139	18,137	19,141
$Adj.R^2$	0.1527	0.1552	0.1555	0.1495	0.1548

Panel B: Alternative measures of analyst forecast accuracy

Variables	(1)	(2)	(3)
	Accuracy1	*Optimism*	*Pessimism*
$Digit$	0.0004*** (5.47)	-0.0015*** (-6.05)	-0.0010*** (-5.30)
Control variables	included	included	included
Industry dummies	included	included	included
Year dummies	included	included	included
N	19,142	19,142	19,142
$Adj.R^2$	0.3396	0.1552	0.3508

Note: Panels A and B of this table report the results from using alternative measures of firms' digital transformation and analyst forecast accuracy, respectively, to test the impact of firms' digital transformation ($Digit$) on analyst forecast accuracy ($Accuracy$). In Panel A, $Digit1$ is the natural logarithm of one plus digitalization-related intangible assets disclosed in a firm's annual report; $Digit2$ is the natural logarithm of one plus the word frequency of digitalization-related information disclosed in the Management Discussion and Analysis (MD&A) section of a firm's annual report; and $Digit3$ ($Digit4$, $Digit5$) is the relative level of digital transformation for a firm in a year, which is estimated as the ratio of, the difference between a firm's digital transformation and the average digital transformation among firms in the same industry (city, province) in a year, to the standard deviation of digital transformation among firms in the same industry (city, province) in a year. In Panel B, $Accuracy1$ is the average forecast accuracy among all analysts following the same firm in a year; the forecast accuracy of each analyst is calculated as -1 times the ratio of the absolute difference between the actual and forecasted earnings per share to actual earnings per share, rather than using the closing stock price as in the $Accuracy$ measure. $Optimism$ ($Pessimism$) is the average forecast optimism (pessimism) among all analysts following the same firm in a year. For each analyst, if his/her forecasted EPS is higher (lower) than the actual EPS, his/her forecast optimism (pessimism) is computed as the ratio of the absolute difference between the actual and forecasted earnings per share to the closing stock price on the trading day preceding the analyst's forecast, and 0 otherwise. In cases where an analyst releases multiple forecasts for the same firm within a year, the final forecast issued at least one month before the end of the fiscal year is utilized for the variable construction. Definitions for all variables are provided in Appendix A. The sample period ranges from 2011 to 2022. Industry dummies (constructed based on the first digit of the China Securities Regulatory Commission (CSRC) Industry Classification Codes) and year dummies are included in all the regressions, but their results are not reported for brevity. t-statistics in parentheses are based on robust standard errors clustered by firm. ***, **, and * indicate 1%, 5%, and 10% two-tailed statistical significance, respectively.

Table 5. Robustness tests using alternative measures of firms' digital transformation and analyst forecast accuracy

Panel A: First-stage results

Variables	(1)	(2)	(3)	(4)
	Operational efficiency	Investment efficiency	Information quality	
	$Operational\_effi$	$Investment\_effi$	$Accruals\_quality$	$MF\_accuracy$
$Digit$	0.0015***	0.0010***	0.0013**	0.0398***
	(3.92)	(4.07)	(2.01)	(2.88)
$Control\ variables$	included	included	included	included
$Industry\ dummies$	included	included	included	included
$Year\ dummies$	included	included	included	included
N	11,119	17,185	19,016	9,943
$Adj.R^2$	0.3160	0.0836	0.0781	0.5271

Panel B: Second-stage results

Variables	(1)	(2)	(3)	(4)
	$Accuracy$			
$Pre\_Operational\_effi$	0.7528***			
	(4.36)			
$Pre\_Investment\_effi$		1.5547***		
		(5.85)		
$Pre\_Accruals\_quality$			1.2132***	
			(6.09)	
$Pre\_MF\_accuracy$				0.0044***
				(7.95)
$Control\ variables$	included	included	included	included
$Industry\ dummies$	included	included	included	included
$Year\ dummies$	included	included	included	included
N	11,119	17,185	19,016	9,943
$Adj.R^2$	0.1422	0.1621	0.1553	0.2603

Note: This table reports the ordinary least square (OLS) regression results of channels through which firms' digital transformation ($Digit$) affects analyst forecast accuracy ($Accuracy$). Panel A shows the results of the mediating effects of three mediators: (i) operational efficiency ($Operational\_effi$) for a firm in a year, evaluated by using the Data Envelopment Analysis (DEA) to construct an efficient frontier of production through an optimization program maximizing the output-to-input ratio (Cheng et al., 2018). Production output is quantified by sales revenues, while production inputs include the number of employees, costs of goods sold, and capital expenditures (Lam et al., 2016). For firms within the same industry and year, those on the efficient frontier are assigned a value of one, while less efficient firms behind the frontier receive efficiency scores between 0 and 1; (ii) investment efficiency ($Investment\_effi$), which is measured as -1 times the absolute values of the residuals of a regression model developed by Richardson (2006) and using Tobin's Q as the proxy for growth opportunities, for a firm in a year; and (iii) information opacity ($Info\_quality$), measured as -1 times the absolute values of abnormal accruals calculated based on the modified Jones model (Dechow et al., 1995) for a firm in a year. Panel B shows the regression results regarding how analysts' forecast accuracy ($Accuracy$) is affected by the predicted values of firms' operational efficiency, investment efficiency, and information opacity (namely, $Pre\_Operational\_effi$, $Pre\_Investment\_effi$, and $Pre\_Info\_quality$, which are estimated from the first-stage regressions). The sample period ranges from 2011 to 2022. Definitions for variables are provided in Appendix A. Industry dummies (constructed based on the China Securities Regulatory Commission (CSRC) Industry Classification Codes) and year dummies are included in all the regressions, but their results are not reported for brevity. t-statistics in parentheses are based on robust standard errors clustered by firm. ***, **, and * indicate 1%, 5%, and 10% two-tailed statistical significance, respectively.

Table 6. Tests of the channels through which firms' digital transformation affects analyst forecast accuracy

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	More work experience ($LnExperience\_analyst$)	Less work experience ($LnExperience\_analyst$)	High accuracy of past forecasts ($Past\_accuracy$)	Low accuracy of past forecasts ($Past\_accuracy$)	Big size of analysts' research portfolio ($LnPortfolio\_analyst$)	Low size of analysts' research portfolio ($LnPortfolio\_analyst$)	Big brokerage house ($Broker\_size$)	Small brokerage house ($Broker\_size$)	Strong social network ($Network\_analyst$)	Weak social network ($Network\_analyst$)
	$Accuracy$									
$Digit$	0.0018***	0.0011***	0.0017***	0.0004***	0.0018***	0.0011***	0.0017***	0.0013***	0.0022***	0.0012***
	(4.77)	(4.19)	(3.88)	(2.63)	(4.86)	(4.04)	(4.82)	(4.51)	(5.60)	(4.41)
$Coefficient\ difference$	(1)-(2) = 0.0007**		(3)-(4) = 0.0013***		(5)-(6) = 0.0007**		(7)-(8) = 0.0004*		(9)-(10) = 0.0010***	
$Control\ variables$	included	included	included	included	included	included	included	included	included	included
$Industry\ dummies$	included	included	included	included	included	included	included	included	included	included
$Year\ dummies$	included	included	included	included	included	included	included	included	included	included
N	9,568	9,574	7,605	7,605	9,571	9,571	9,535	9,535	8,412	8,411
$Adj.R^2$	0.1780	0.1338	0.1545	0.0684	0.1613	0.1357	0.1764	0.1406	0.1699	0.1429

Note: This table reports the results from testing the moderating effects of analyst characteristics on the association between firms' digital transformation ($Digit$) and analyst forecast accuracy ($Accuracy$). Columns (1) and (2), (3) and (4), (5) and (6), (7) and (8), and (9) and (10) show the results for four moderators: (i) the natural logarithm of average work experience among all analysts following the same firm in a year ($LnExperience\_analyst$); work experience of each analyst is measured as the number of years elapsed from his/her initial forecast to the latest forecast; (ii) the average forecast accuracy over the prior 5 years among all analysts following the same firm in a year ($Past\_accuracy$); the average is taken both on the prior 5 years and on all analysts; (iii) the natural logarithm of the average size (measured by total assets) of firms in portfolios covered by all analysts following the same firm in a year ($LnPortfolio\_analyst$); the average is taken both on the portfolio firms' total assets and on all analysts; (iv) the average size of brokerage houses employing analysts who follow the same firm in a year, with the size measured as the brokerage house's total assets ($Broker\_size$); and (v) a composite measure of the social network centrality ($Network\_analyst$) of an analyst for each year, which is derived by using a common factor analysis of three social network centrality variables: degree centrality (the number of direct connections of an analyst to other analysts), closeness centrality (an inverse measure of the average of the minimum network distances between an analyst and each of the other analysts), and eigenvector centrality (the eigenvector of the largest eigenvalue of the non-negative adjacency matrix of the analyst's social network). Definitions for all the variables are provided in Appendix A. The sample period ranges from 2011 to 2022. Industry dummies (constructed based on the China Securities Regulatory Commission (CSRC) Industry Classification Codes) and year dummies are included in all the regressions, but their results are not reported for brevity. t-statistics in parentheses are based on robust standard errors clustered by firm. ***, **, and * indicate 1%, 5%, and 10% two-tailed statistical significance, respectively.

Table 7. The moderating effects of analyst characteristics

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	More invention patents ($Invention$)	Fewer invention patents ($Invention$)	More modelling and design patents ($Modelling$ & $design$)	Fewer modelling and design patents ($Modelling$ & $design$)	More R&D expenditures ($R\&D$ $expenditure$)	Fewer R&D expenditures ($R\&D$ $expenditures$)	More R&D employees ($R\&D$ $staff$)	Fewer R&D employees ($R\&D$ $staff$)
	$Accuracy$							
$Digit$	0.0011***	0.0018***	0.0012***	0.0018***	0.0011***	0.0017***	0.0007**	0.0019***
	(3.27)	(5.58)	(3.08)	(5.69)	(3.51)	(4.09)	(1.95)	(3.98)
$Coefficient\ difference$	(1)-(2) = -0.0007***		(3)-(4) = -0.0006**		(5)-(6) = -0.0006***		(7)-(8) = -0.0008**	
$Control\ variables$	included	included	included	included	included	included	included	included
$Industry\ dummies$	included	included	included	included	included	included	included	included
$Year\ dummies$	included	included	included	included	included	included	included	included
N	9,157	9,967	9,062	10,061	8,476	8,476	5,994	5,996
$Adj.R^2$	0.1699	0.1454	0.1682	0.1463	0.1415	0.1679	0.1592	0.1943

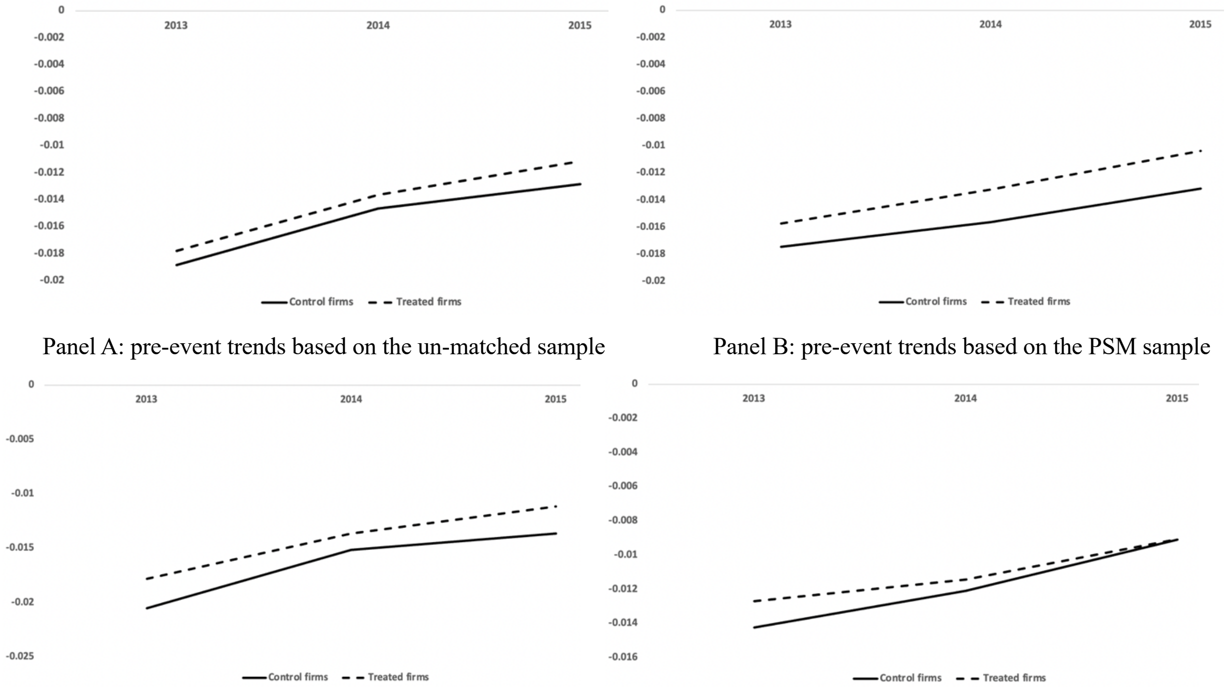
Note: This table reports the results from testing the moderating effect of corporate innovation on the association between firms' digital transformation ($Digit$) on analysts' forecast accuracy ($Accuracy$). Columns (1) and (2), (3) and (4), (5) and (6), and (7) and (8) show the results of four moderators: (i) the number of invention patents applied by a firm in a year and subsequently granted by the China National Intellectual Property Administration ($Invention$); (ii) the number of product-modelling patents and product-design patents applied by a firm in a year and subsequently granted by the China National Intellectual Property Administration for the firm ($Modelling$ & $design$); (iii) the research and development (R&D) expenditures incurred by a firm in a year, scaled by sales revenue ($R\&D\ expenditure$); and (iv) the number of R&D employees for a firm in a year ($R\&D\ staff$), respectively. The sample period ranges from 2011 to 2022. Definitions for variables are provided in Appendix A. Industry dummies (constructed based on the China Securities Regulatory Commission (CSRC) Industry Classification Codes) and year dummies are included in all the regressions, but their results are not reported for brevity. t-statistics in parentheses are based on robust standard errors clustered by firm. ***, **, and * indicate 1%, 5%, and 10% two-tailed statistical significance, respectively.

Table 8. The moderating effects of corporate innovation

Variables	(1)	(2)
	High-tech industries ($High\_tech$)	Non-high-tech industries ($High\_tech$)
	$Accuracy$	
$Digit$	0.0016*** (4.72)	0.0014** (3.86)
$Coefficient\ difference$	(1)-(2) = 0.0002**	
$Control\ variables$	included	included
$Industry\ dummies$	included	included
$Year\ dummies$	included	included
N	6,776	12,366
$Adj.R^2$	0.1458	0.1566

Notes: This table reports the results from testing the moderating effect of industry-level technology development on the association between firms' digital transformation ($Digit$) on analysts' forecast accuracy ($Accuracy$). Columns (1) and (2) show the results based on subsamples of high-tech and non-high-tech industries, respectively. High-tech firms are defined as those operating in the following high-tech industries: chemical material and chemical product manufacturing industry; medical and pharmaceutical manufacturing industry; chemical fiber manufacturing industry; railway, shipbuilding, aircraft, and other transportation equipment manufacturing industry; computer, communication, and other electronic equipment manufacturing industry; instrument manufacturing industry; information transmission, software and information technology services industry; and scientific research and technology services industry (e.g., Liu and Buck, 2007; Shen et al., 2024). Definitions of all the variables are provided in Appendix A. The sample ranges from 2011 to 2022. Firm dummies, industry dummies (constructed based on the China Securities Regulatory Commission (CSRC) Industry Classification Codes) and year dummies are included where appropriate in the regressions, but their results are not reported for brevity. t-statistics in parentheses are based on robust standard errors clustered by firm. ***, **, and * indicate 1%, 5%, and 10% two-tailed statistical significance, respectively.

Table 9. The moderating effect of industry-level technology development



Panel A: pre-event trends based on the un-matched sample

Panel B: pre-event trends based on the PSM sample

Panel C: pre-event trends based on the entropy balanced sample

Panel D: pre-event trends based on the CEM sample

Note: This figure shows the trends in analyst forecast accuracy ($Accuracy$) between the treated and control firms during the three-year period from 2013 to 2015, preceding the establishment of China's big-data comprehensive pilot zones (BDCPZ) in 2016. Panels A, B, C, and D plot the pre-event trends in analyst forecast accuracy, which are derived from the dynamic DID regression based on the un-matched sample, propensity-score matched sample, entropy-balanced sample, and coarsened-exact matched sample, respectively. The dynamic DID regression is adapted from Model (3) by replacing $Treatment * post$ with the interaction terms between $Treatment$ and year dummies.

Figure 1. Pre-event trends in analyst forecast accuracy between the treated and control firms

Does commercial reform embracing digital technologies mitigate stock price crash risk?

Guanming He

Department of Accounting, Business School, Durham University, Durham, UK

Email: guanming.he@durham.ac.uk

Zhichao Li

Department of Finance and Accounting, Business School, University of Exeter, Exeter, UK

Email: z.li10@exeter.ac.uk

Ling Yu

National School of Development, Peking University, Beijing, China

Email: yling@pku.edu.cn

Zhanqiang Zhou

School of Economics, Central University of Finance and Economics, Beijing, China

Email: zhouzhanqiang@cufe.edu.cn

* We appreciate the helpful comments and suggestions from an anonymous reviewer, Xuan Tian, Aishwarya Krishna, Tong Yu, Srijita Chakraborty, Edgar Bellow, Aymen Turki, Juliane Proelss, Jiajun Jiang, Aija Leiponen, Hamid Yahyaei, and seminar participants at the Tsinghua University, Guizhou University, 5th annual conference of the Association of British Chinese Professors (ABCP), 17th International Risk Management conference, 2024 Research Symposium on Finance and Economics, 2024 Academy of International Business Asia Pacific 2024 Conference, 2023 China Finance Review International & China International Risk Forum Joint Conference, 4th International Conference on Digital, Innovation, Financing & Entrepreneurship, 57th World Continuous Audit & Reporting Symposium & the British Accounting and Finance Association (BAFA) Accounting Information Systems and Emerging Technologies conference, 4th International Conference on Economics, 2nd Northeast Digital Marketing Conference, 2023 Clermont Financial Innovation Workshop, 2023 International Conference on Empirical Economics, and 2023 International Conference on "Interdisciplinary Research in Technology & Management". We are liable for any error in this paper.

Does commercial reform embracing digital technologies mitigate stock price crash risk?

Abstract

Over the recent decade or so, the Chinese government implemented a commercial reform that features governmental application of digital technologies to acquire and process firm information. The core objective of commercial reform is to improve information transparency and monitoring on corporate commercial activities. To explore the economic effectiveness of the reform, we examine how it impacts firms' stock price crash risk. We find robust evidence that the commercial reform that digitalizes government regulatory activities mitigates stock price crash risk and achieves so via enhancing information environment and monitoring for firms. This finding is more prominent for firms with higher levels of digitalization and innovation and those with weaker internal governance. Overall, our findings highlight a potential benefit of applying digital technologies to regulatory reform, encouraging governments to adopt digital tools to improve information environments and monitoring for firms, and thereby promoting a more stable and efficient capital market.

Keywords: commercial activities; commercial reform; digitalization; stock price crash risk; innovation; governance

JEL codes: G12; G14; G18

1. Introduction

In the era of digitalization, the Chinese government has adopted digital technologies for commercial reform. It features the governmental utilization of digital technologies to acquire and process firm information for purpose of facilitating real-time monitoring on commercial activities under transparent information environments. The primary goals of the reform are to provide commercial convenience for enterprises, ensure fair and transparent regulation of corporate activities, and promote healthy development of commercial activities within a country. In this study, we investigate the effectiveness of the commercial reform by providing evidence from stock price crash risk.

The digitalization-applied commercial reform involves the utilization of digital technologies by the government to transform and upgrade government activities, with the primary objective of facilitating and regulating corporate commercial activities. The application of digital technologies is an integral part of commercial reform and serves two crucial roles in making the reform plausibly effective. First, it may improve information transparency of firms' commercial activities. By providing convenient digital commercial registration and approval services, the government can efficiently collect an extensive array of commercial information, integrate it into a comprehensible form, promptly analyze this big data, and accurately transmit it among government departments, firms, and the public. Second, digitalization may also enhance the monitoring of firms' commercial activities. Implementing digital and intelligent monitoring in the commercial reform allows the government to improve interdepartmental regulatory cooperation, promote diverse monitoring approaches, and raise firms' awareness of commercial credit. These digital monitoring tools would help standardize firm-relevant commercial conducts, and prevent firms from engaging in suboptimal, illegal, or value-destroying commercial activities.

However, the application of digital technologies in commercial reform may be ineffective in increasing information transparency and enhancing the monitoring of firms' commercial activities if we consider the associated risks and costs. Prior studies document that technological obsolescence (Acemoglu, 2002), privacy concerns (Dinev and Hart, 2006), and cybersecurity risks (Rosati et al., 2022), which are involved in the practices of digitalization, may deter firms from enhancing information transparency and monitoring. In addition, applying digital technologies to commercial reform requires considerable time and entails substantive expenses, learning costs, and uncertainties (Luo, 2022). Therefore, it is unclear whether digitalization-involved commercial reform would improve the information environment and enhance the monitoring of firms' commercial activities.

To address the open question, we investigate the impact of digitalization-applied commercial reform on stock price crash risk. Such risk results from manager opportunism that leads to overvaluation of stocks (e.g., Hutton et al., 2009; Kim et al., 2011), and is closely bound up with both information opacity and inadequate monitoring of corporate activities (e.g., Hutton et al., 2009; Kim et al., 2011). Therefore, by examining the effect of digitalization-involved commercial reform on stock price crash risk, we may shed light on the effectiveness of the government's adoption of digital technologies in commercial reform. If the digitalization-involved commercial reform improves information transparency and monitoring of firms' commercial activities, stock price crash risk is supposed to decrease.

We focus on the digitalization-involved commercial reform in China for two reasons. First, it provides a nice institutional setting for a quasi-natural experiment. Since 2014, the Chinese government has initiated a commercial reform, wherein the Market Supervision Administration (MSA) in each city is established over different years and takes the main responsibility for implementing the commercial reform. For the reform, the municipal MSA actively adopted digital technologies to streamline corporate online applications, acceptances, reviews, license issuances, and publicity for enterprise commercial activities and to process relevant commercial information intelligently for monitoring the activities. This setup provides a reasonable context for employing a stacked difference-in-differences research design to establish causality. Second, the information environment and monitoring of commercial activities are relatively weak in China compared with those of developed countries (e.g., Piotroski and Wong, 2012). Hence, a study on the effectiveness of Chinese commercial reform that embraces digital technologies is potentially generalizable to other countries, especially the developing ones.

We manually collected data on the timing of establishing the MSA in each city to proxy for the timing of enacting the digitalization-involved commercial reform across cities. A difference-in-differences regression model is applied on a stacked propensity-score matched sample to explore whether the digitalization-involved commercial reform mitigates firms' stock price crash risk.[1] We find evidence to suggest that the commercial reform reduces crash risk. The finding is robust to firm-fixed-effects regression analyses, controls of region effects, tests of coefficient stability, placebo tests, and alternative measures of crash risk. Further, we provide evidence that improved information environments and monitoring are the underlying mechanisms through which the attenuating effect of digitalization-involved commercial reform on crash risk realizes. We also find that this mitigating effect is more evident for firms with higher levels of digitalization and innovation and those with weaker internal governance.

Our paper makes two main contributions. First, we extend existing studies on the effect of

[1] A difference-in-differences regression model applied on a stacked sample for staggered events is named stacked difference-in-difference regression design.

digitalization. Prior literature documents the economic consequences of corporate utilization of digital technologies (e.g., Ferreira et al., 2019; Blichfeldt and Faullanti, 2021; Ciampi et al., 2021; Matarazzo et al., 2021; Chen et al., 2022; Xu et al., 2022), and have paid little attention on government application of digital technologies. Our paper is the first to show how governmental adoption of digital technologies in a regulatory reform would achieve the desired regulatory outcomes. By exploring the impact of digitalization-applied commercial reform on crash risk, our research enriches the understanding of the economic consequences of digitalization from a macro perspective. Second, we offer some insights into the policy implementation. By showing that digitalization-involved commercial reform reduces stock price crash risk via effectively improving information transparency and monitoring of firms' commercial activities, we highlight the benefits of applying digital technologies to achieve regulatory objectives, and the benefits of government digitalization to firms and other stock market participants.

The remainder of the paper is organized as follows. Section 2 introduces the institutional background, and proposes the research hypothesis from two aspects – the information channel and the monitoring channel. Section 3 describes the data and methodologies for our empirical analysis. Section 4 discusses empirical results. Section 5 concludes our study.

2. Institutional background and research hypothesis

2.1. The commercial reform in China

In 2013, the Chinese government held several national conferences on reforming the commercial registration system to simplify the registration processes, ease market access, and strengthen the supervision and management of commercial activities.[2] Following these conferences, in 2014, the Chinese government launched a commercial reform nationwide which emphasizes the application of digital technologies. Specifically, local governments in each city are required to provide online services regarding commercial activities for local firms, and use digital technologies to promote data processing as well as data sharing and integration across different departments.

In implementing this digital commercial reform, the Market Supervision Administration (MSA) is established in each city, and responsible for creating various online integrated data platforms, including the National Enterprise Credit Information Publicity Platform (NECIPP), to aggregate a broad spectrum of corporate commercial information and disclose it to the public, not least the media and stock market participants. The information covers financial records, credit ratings, business registration, licensing, regulatory compliance, administrative penalties, commercial transactions, labor relations, shareholder changes, and intellectual property, among other aspects. Data on this diverse information are consolidated and sent to the cloud server, allowing the governments to store and further process them in a big-data platform. Then, leveraging the cloud-based repository, the governments implement a data-sharing system across various departments by using blockchain technology. This ensures trackable data records, data privacy, and seamless data flows among departments. The application of blockchain technology focuses mainly on e-certificates, business registration, and e-invoices. Under the data-sharing system, the same type of credentials and information need to be submitted only once and can be

[2] On 28th February 2013, the Chinese government held the Second Plenary Session of the 18th Communist Party of China (CPC) Central Committee, where it decided to reform the commercial registration system, ease the market access, and strengthen the supervision and management of corporate commercial activities. Later, on 12th November of the same year, the Third Plenary Session of the 18th CPC Central Committee further called for promoting the commercial reform.

used interchangeably across departments.

Meanwhile, big data analytics and cloud computing are employed to analyse and scrutinize the data. On the one hand, governments use these techniques to extract useful information from big databases and gain insights into industry trends, market demands, investment details, patents, bidding, etc. They then share this information with enterprises, assisting them in bolstering their competitive edge. On the other hand, big data analytics enable governments to swiftly pinpoint operational risks, detect potential frauds, issue risk alerts, and initiate appropriate regulatory actions. Furthermore, artificial intelligence (AI) is also incorporated into some government online services. Digital features like AI service expedite the governments' processing of firms' requests by quickly providing guidance and undertaking initial reviews, such as review of business registration, effectively lightening the workload for government employees.[3] All the foregoing information processed by digital technologies will be used for the governmental monitoring on the firms' commercial activities; some processed information, such as the one related to abnormal business operations, will be released publicly, improving the information environments of firms and facilitating public monitoring as well on their commercial activities.

This reform with emphasis on the application of digital technologies integrates government operations, enhances the information system, and elevates management standards for the government. To this end, the Market Supervision Administration (MSA) is established in a staggered way in each city at different years and takes the main responsibility of executing the local commercial reform. Decisions on the timing of the establishment of MSA are autonomously

[3] More information about the application of digital technologies in the government works can be obtained from the "research report on the modernization of national governance in the digital age - experiences, challenges, and responses in using digital technologies for government governance" by the China Academy of Information and Communications Technology (CAICT). The Chinese version of this report can be accessed via the link http://www.caict.ac.cn/kxyj/qwfb/ztbg/202212/P020221207530304282075.pdf.

made by the local government in each city, and are orthogonal to firms' characteristics and events. As firms cannot anticipate the specific timing of establishing the local MSA, they are unlikely to respond to the reform in advance. Therefore, it facilitates us to examine its causal impact on stock price crash risk via a stacked difference-in-differences research design.

2.2. Hypothesis development

Stock price crash risk refers to the possibility of a sudden and significant decline in the stock price (Chen et al., 2001). It is primarily attributed to managers' opportunistic behaviours (e.g., withholding of bad news) leading to investors' overvaluation of stocks (Jin and Myers, 2006). The information asymmetry between investors and managers and the inadequate monitoring of the latter would make it difficult to detect managerial opportunism and potentially hidden corporate bad news, thereby increasing stock price crash risk (e.g., Hutton et al., 2009; Kim et al., 2011; He et al., 2019). Therefore, it is of great importance for regulators to reduce stock price crash risk by enhancing the information environment and monitoring in a commercial reform.

The utilization of digital technologies for commercial reform may enhance the government's ability to collect, process, and share various corporate commercial information, thereby improving the quality and transparency of corporate information as well as external monitoring on firms. Regarding the information acquisition, a variety of digital government services provided during the commercial reform (e.g., online application systems, self-service terminals, and mobile terminals) help firms independently complete commercial registration procedures and swiftly publish commercial information related to their products, services, sales, business expansion, etc., and disclose additional details especially those concerning the creditworthiness of their commercial activities. In such a case, the government can promptly collect a wide range of up-to-date commercial information from different firms, even before its public announcements, and form a big database for comprehensive data analyses on a timely basis.

The application of digital technologies also contributes to effective and efficient information processing. On the one hand, by utilizing advanced big data analytics and cloud computing, the government can classify and group unstructured data from various sources across firms, such as images, news, videos, and audio. This facilitates the government to track and analyze commercial information through the process of a firm's commercial activities, from product design, quality monitoring, marketing, and sales to distribution. Some processed information especially related to abnormal business operations will be published on the government's online service platforms, increasing corporate information transparency. On the other hand, by analyzing the structured data, the government can perform dynamic, real-time, and intelligent monitoring on both the upstream and downstream firms in the supply chain (Gomber et al., 2018; Cong and He, 2019). For instance, using the technique of big data analytics, governments could foresee potential operational risks and generate risk alerts once identified by the digital risk-warning system. Other diversified monitoring through internet technologies, such as e-government platforms in real-time, allows the public to monitor and report in good time any violations of rules related to firms' commercial activities, internal controls, and financial reports. This prompt reporting by the public further facilitates regulators to detect firms' non-compliant activities so that penalties and corrections can be imposed in a timely manner.

Furthermore, using digital technologies such as blockchain significantly improves information sharing across different government sectors. The government can standardize and digitize numerous commercial information, timely transmit valuable commercial information across different departments, and then release it to the public for oversight. Consequently, commercial details concerned by market participants, such as regulatory non-compliance, unethical business practices, poor financial performance, legal complications, and corporate social irresponsibility, would become more transparent. Better information sharing would also eliminate the overlap of regulatory responsibilities among different governmental departments. This strengthens the accountability of each department and fosters better coordination across departments. As a result, the costs of monitoring decrease while the efficiency of monitoring improves.

In essence, the digitalization in commercial reform may help improve both the information environments and monitoring on the firms' commercial activities. The firms' information environments could be ameliorated via media coverage on commercial information processed and released by the government, as the media plays a crucial role in disseminating commercial news to a wide range of stakeholders. The improved information environment would in turn reduce stock price crash risk. For instance, high information quality and transparency enable managers, based on existing commercial information, to conduct more reliable assessments on future commercial investments. This improves firms' investment efficiency and prevent managers from investing in commercial projects that have negative present values (Biddle et al., 2009; Lai et al., 2014). Meanwhile, investors in the transparent environment will have better insight into government policies and firms' commercial investment activities, thus reducing their overvaluation of stocks (Drake et al., 2009; Lee and Lee, 2015). Moreover, information transparency raises the costs for managers to commit malpractice or malfeasance in commercial activities and to hide bad commercial news from investors. As a result, the stock price crash risk will diminish.

The improved monitoring due to digitalization-applied commercial reform further contributes to the reduction in stock price crash risk by mitigating firms' agency conflicts (Fan and Wong, 2005), reducing related-party transactions (Gallery et al., 2008), preventing firms from engaging in suboptimal, illegal, or value-destroying commercial activities, and prompting firms to disclose high-quality commercial information on a timely basis. In addition, digital monitoring in commercial reform can strengthen corporate credit education as well as credit monitoring of firms for their commercial activities. By using diverse digital information disclosure systems, governments can promptly analyze commercial credit information, release it online and issue early warnings when appropriate to relevant parties, thereby guiding and ensuring firms to adhere to laws, regulations, and ethical practices. This is instrumental in fostering the development of a robust commercial credit system and enhancing the standardization and credibility of firms' commercial activities to investors. As the information acquired and processed on a real time basis by the government via digital tools would also be released to the public for oversight, the reform would enhance not only the monitoring by the government but also by the stock market participants.

However, capitalizing on digital technologies in commercial reform does not necessarily increase the transparency of corporate commercial information or the external monitoring of firms' commercial activities. As such, it may not reduce stock price crash risk. This can be attributed to the potential risks and costs that are associated with technological obsolescence, privacy concerns, and cybersecurity risks, among others (Acemoglu, 2002; Dinev and Hart, 2006; Rosati et al., 2022). Technological obsolescence can lead to lower data quality and accuracy, posing challenges for the government to promptly capture the accurate commercial information of firms. Consequently, information opacity will rise (Acemoglu, 2002), impeding the effective monitoring and evaluation of firms' commercial activities and financial performance. Insufficient privacy protection could give rise to mistrust among firms and the public regarding the government's data collection and usage. As such, firms may be reluctant to disclose complete commercial information, hindering the external monitoring of their behaviors. Cybersecurity risks, such as cyber-attacks and data breaches, pose a threat of insecure data, information losses, or information tampering. These vulnerabilities will limit the government from obtaining accurate commercial information and reduce the monitoring effectiveness. Besides, the adoption of digital technologies brings additional expenses and uncertainties. Implementing new technologies properly requires ample time and substantial investments in hardware, software, and staff training. There are also learning costs associated with adopting new technologies and the costs of integrating with the existing government management systems. Considering the foregoing risks and costs associated with applying digital technologies in the commercial reform, it might not be effective in improving the information environment and monitoring on firms' commercial activities and thereby reducing stock price crash risk. Based on the above discussion, we propose the following null hypothesis for empirical tests:

H1: The digitalization-applied commercial reform is unrelated to firms' stock price crash risk.

3. Data and methodologies

3.1. Data sources and sample selection

We focus on listed companies in our study.[4] Data utilized for the empirical tests come mainly from two databases: China Stock Market & Accounting Research (CSMAR) and Chinese Research

[4] There are four reasons for focusing on listed firms for the empirical analysis of the effectiveness of digitalization-applied commercial reform. First, the commercial activities of listed firms involve a myriad of stakeholders and concern public interest, investor protection as well as the stability of capital market, among others. Their commercial information accessible via reputable government websites is trusted and sought highly by the stakeholders. Second, the government's digital platforms form an important channel through which listed firms release value-relevant information to investors. Hence, the commercial reform would affect these firms significantly. Third, listed firms often have greater influence and visibility in the market, so their commercial activities can act as a model for reference by other enterprises. Fourth, from the methodological point of view, a signficantly more comprehensive set of publicly available data from Chinese listed firms, relative to those from non-listed firms, enable us to perform a more rigorous empirical analysis to assure the internal validity of results.

Data Services (CNRDS). Data on the stock trading, financial numbers, and governance structure of firms are taken from CSMAR. Data on media news about a firm are gathered from CNRDS. We hand-collected data on the timing of establishing the Market Supervision Administration in each city by searching the Chinese Industry and Commerce Administration Yearbook and/or the official websites of the municipal governments. Data on firm-level digitalization, which are used later for our moderation analysis, are obtained based on the approach proposed by Chen and Srinivasan (2023). This method employs the Python Crawler technique to search for and collate the digitalization-related keywords in firms' annual reports. Patent data used to construct the moderator variable regarding corporate innovation are collected from the website of the Chinese State Intellectual Property Office.

We focus on the policy implementation period of 2014-2019. Since 2014, the Chinese government across all administrative levels has implemented commercial reform, in which the Market Supervision Administration of each city introduced various digital technologies in a staggered manner. Therefore, we start our policy implementation period from 2014. Considering the confounding impact of COVID-19 on stock price crash risk, we end the policy implementation period in 2019. Meanwhile, we use a six-year period centered on the implementation year of the reform (i.e., a three-year pre-event period and a three-year post-event period) in our difference-in-differences research design. As a result, our treatment group only includes firms headquartered in cities that implemented commercial reform between 2014 and 2017. Therefore, our sample period starts from (ends in) 2011 (2019), three years before (since) 2014 (2017), while covering the period of the enactment of digitalization-involved commercial reform.

Our sample selection starts with the population of Chinese listed firms that have A shares traded on the Shenzhen and Shanghai Stock Exchanges for the period 2011-2019. This initial sample consists of 26,345 firm-year observations, corresponding to 4,016 firms. Following prior studies, we exclude firms that receive Special Treatment (ST or *ST) or Particular Transfer (PT), as these firms are of high delisting risk. We then tease out firms in financial industries because the disclosure requirements and accounting rules for firms in financial industries differ significantly from those in the other industries. Firms cross-listed overseas are also deleted from our analysis, as their stock prices are influenced by foreign stock markets. We further eliminate observations with negative incomes. Finally, we remove firm-year observations that do not have the necessary data to construct the variables of interest for our regression analysis. We end up with 16,237 firm-year observations for 2,577 listed firms. Appendix 1 expounds our sample selection procedure.

3.2. Measures of stock price crash risk

In line with previous research (e.g., Chen et al., 2001; Hutton et al., 2009; Kim et al., 2011; Chen et al., 2016), we measure stock price crash risk by the negative skewness of weekly stock returns (*NCSKEW*) and down-to-up volatility of weekly stock returns (*DUVOL*) over a fiscal year. For *NCSKEW*, we first calculate the firm-specific weekly raw returns by estimating the following equation:

$$r_{i,s} = \delta + \delta_{1,i} r_{m,s-2} + \delta_{2,i} r_{m,s-1} + \delta_{3,i} r_{m,s} + \delta_{4,i} r_{m,s+1} + \delta_{5,i} r_{m,s+2} + \varepsilon_{i,s} \qquad (1)$$

where $r_{i,s}$ is the raw return of stock i in week s; $r_{m,s}$ is the value-weighted market rate of return of all stocks in week s. In particular, the lag terms (i.e., $r_{m,s\text{-}1}$, $r_{m,s\text{-}2}$) and lead terms (i.e., $r_{m,s+1}$, $r_{m,s+2}$) are also included to allow for the nonsynchronous stock trading (Dimson, 1979). $\varepsilon_{i,s}$ is the residual return from Equation (1). The firm-specific weekly return of stock i in week s, $w_{i,s}$, is measured as the natural logarithm of one plus the residual return in Equation (1), that is, $w_{i,s}=ln(1+\varepsilon_{i,s})$ (e.g., Kim et al., 2011).

NCSKEW for a firm *i* in a fiscal year *t* is measured by taking the negative of the third moment of firm-specific weekly returns for each sample firm-year and dividing it by the standard deviation of firm-specific weekly returns raised to the third power:

$$NCSKEW_{i,t} = -\left[n(n-1)^{\frac{3}{2}}\sum w_{i,s}^{3}\right] / \left[(n-1)(n-2)\left(\sum w_{i,s}^{2}\right)^{\frac{3}{2}}\right] \tag{2}$$

where *n* is the number of trading weeks for stock *i* in year *t*.

DUVOL captures asymmetric volatilities between the negative and positive firm-specific weekly returns and is calculated as follows:

$$DUVOL_{i,t} = ln\left[(n_u - 1)\sum_{down} w_{i,s}^{2}\right] / \left[(n_d - 1)\sum_{up} w_{i,s}^{2}\right] \tag{3}$$

where n_u (n_d) is the number of weeks in which the firm-specific weekly returns of stock *i* are higher (lower) than the annual average return. The larger the negative skewness of weekly stock returns (*NCSKEW*) or the down-to-up volatility of weekly stock returns (*DUVOL*), the greater the probability of stock price crashes for the firm.

3.3. Difference-in-differences research design

Given that the municipal MSA is the primary responsible authority for commercial reform in each city, we utilize the timing of establishing municipal MSA to reflect the timing of implementing the commercial reform. MSA is established in different cities at different years, so we adopt a stacked difference-in-differences (DID) approach to evaluate the economic effect of commercial reform on firms' stock price crash risk. The DID research design requires identifying a treatment (control) group, of which firms are (not) subject to the exogenous regulatory event. Accordingly, our treatment group comprises firms headquartered in cities that established MSA from 2014 to 2017. To maintain a clean identification of the control groups for matching with treatment firms for a year *t* (Baker et al., 2022; Roth et al. 2023), we classify firms, headquartered in cities that did not establish MSA during the six-year period from year *t*-3 to *t*+2 nor before year *t*-3, into our control group. For example, if a firm is based in the city where an MSA was established in 2014, the control firms used to match these treatment firms in 2014 are firms with headquarters in cities that did not have an MSA at or before 2016.

The stacked DID regression model is specified as follows:

$$
\begin{aligned}
& NCSKEW_{i,t} \text{ or } DUVOL_{i,t} = \\
& \quad \alpha_0 + \alpha_1 Treat_t \times Post_i + \alpha_2 Treat_t + \alpha_3 size_{i,t} + \alpha_4 soe_{i,t} + \alpha_5 roe_{i,t} + \alpha_6 lev_{i,t} \\
& \quad +\alpha_7 salesgrowth_{i,t} + \alpha_8 cashholdings_{i,t} + \alpha_9 duality_{i,t} + \alpha_{10} boardsize_{i,t} \\
& \quad +\alpha_{11} topshareholdings_{i,t} + \alpha_{12} hhi_{i,t} + \alpha_{13} ceoshare_{i,t} + \alpha_{14} ret_{i,t} + \alpha_{15} sigma_{i,t} \\
& \quad +\alpha_{16} share\_turnover_{i,t} + \alpha_{17} roa\_volatility_{i,t} + year\_dummies \\
& \quad +industry\_dummies + city\_dummies + \varepsilon_{i,t} \qquad (4)
\end{aligned}
$$

where the dependent variable is stock price crash risk (*i*.e., *NCSKEW* or *DUVOL*). $Treat_t$ is an indicator for the treatment and equals 1 (0) if a firm is in the treatment (control) group at year *t*. $Post_i$ is the time indicator which equals 1 (0) if a firm is in the three-year post- (pre-) event period that is from year *t* (year *t*-3) to year *t*+2 (year *t*-1). The coefficient on interaction term, $Treat_t \times Post_i$, captures changes in the stock price crash risk of treatment firms, relative to those of control firms, from the pre-event period to the post-event period. $Post_i$ is not included in the regression as this variable is potentially multicollinear with the year dummies.

Consistent with previous research (e.g., Kim et al., 2011; Chen et al., 2016; Jin et al., 2022), we control for a bunch of variables that may affect stock price crash risk, i.e., firm size (*size*), state ownership (*soe*), return on equity (*roe*), financial leverage (*lev*), sales growth (*salesgrowth*), financial health (*cashholdings*), CEO-chair(wo)man duality (*duality*), board size (*boardsize*), the largest shareholder's stock holdings (*top_shareholdings*), industrial concentration (*hhi*), CEOs' stock holdings (*ceoshare*), the average weekly stock returns (*ret*), the volatility of weekly stock returns (*sigma*), share turnover (*share_turnover*), and the volatility of returns on assets (*roa_volatility*). We also include year dummies, industry dummies, and city dummies (*year_dummies*, *industry_dummies*, and *city_dummies*) in our regressions. All variables are winsorized at the 1st and 99th percentiles to avoid the impact of outliers on our results, and are defined in Appendix 2. The standard errors of coefficients in the regressions are clustered at the firm level to control for potential heteroscedasticity and autocorrelation.

3.4. Propensity score matching

The potential systematic differences in firm characteristics between the treated firms and controlled firms may bias our analysis. To mitigate this concern, we perform the propensity score matching (PSM) and use the post-matched sample to run our DID regression. We do the matching year by year to ensure that our DID design based on the matched sample will compare the outcome of the treatment for the same treated firm, relative to that of its matched control firm, for the same year of interest. We match each treatment firm, with replacement, with a control firm by the year of establishing MSA in the city where the treatment firm is headquartered. A vector of matching covariates are selected as independent variables to run the following logit regression for the binary variable, *Treat*, to obtain the closest propensity score within a caliper of 1% in each year:

$$
\begin{aligned}
&Treat_t = \\
&\quad \beta_0 + \beta_1 size_{i,t} + \beta_2 roe_{i,t} + \beta_3 lev_{i,t} + \beta_4 salesgrowth_{i,t} + \beta_6 cashholdings_{i,t} \\
&\quad + \beta_6 boardsize_{i,t} + \beta_7 roa\_volatility_{i,t} + industry\_dummies + city\_dummies + \varepsilon_{i,t} \quad (5)
\end{aligned}
$$

The matching covariates include firm size (*size*), return on equity (*roe*), financial leverage (*lev*), sales growth (*salesgrowth*), financial health (*cashholdings*), board size (*boardsize*), the volatility of returns on assets (*roa_volatility*), as well as the industry dummies and city dummies. After the matching, we obtain the final sample, which comprises 7,072 firm-year observations corresponding to 1,156 unique firms, for our DID regression analysis.

To check the effectiveness of our matching, we perform a test of the common support in propensity-score matching. The result of the test is displayed in Figure 1. As shown in Figure 1-a, a certain difference exists in propensity scores between the treatment group and the control group prior to the matching. Figure 1-b reveals that after the matching, the distribution trends of the treatment group and the control group become similar. These results indicate that our matching substantively reduces the differences between the treated firms and the non-treated control firms.

To further check the covariate balance, we run the preceding logit regression, Model (5), by year based on the pre-matched and post-matched samples, respectively. Panel A (Panel B) of Table 1 reports the results for the pre-matched (post-matched) sample. While some covariates have statistically significant coefficients for the pre-matched sample, the coefficients for all covariates become statistically nonsignificant after the matching. These results further support the effectiveness of our propensity-score matching.

3.5. Descriptive statistics

Panel A of Table 2 reports the summary statistics of all variables, which are based on the sample after PSM and used in our regression analysis. The mean value of *NCSKEW* (*DUVOL*) is -0.243 (-0.195), with a standard deviation of 0.737 (0.506). The mean value of *Treat* is 0.511, indicating that approximately 51.1% of our sample firms are subject to digitalization-applied commercial reform and are classified into the treatment group, while the remaining 48.9% of firms do not experience such a reform and are classified into the control group. Panel B of Table 2 shows the Spearman correlation matrix of variables. *NCSKEW* and *DUVOL* are highly correlated, with the statistically significant correlation coefficient of 0.879, suggesting that these two variables capture the underlying same construct for stock price crash risk. The values of all other correlation coefficients are below 0.6, assuring that multicollinearity is of less concern in our regression analyses.

4. Empirical analysis of the effect of digitalization-applied commercial reform on stock price crash risk

4.1. Tests of parallel trends assumption

The validity of difference-in-differences research design relies crucially on the parallel trends assumption, which requires similar trends of the outcome variable (i.e., stock price crash risk) for both the treatment and control groups in the pre-event period (i.e., before the implementation of digitalization-involved commercial reform) (e.g., Beck et al., 2010; Roberts and Whited, 2013). To test this assumption, we first construct the following model to compare the stock price crash risk of treatment firms with that of control firms for our pre- versus post-event periods:

$$\begin{aligned}
&NCSKEW_{i,t} \text{ or } DUVOL_{i,t} = \\
&\quad \beta_0 + \beta_1 Treat_t \times Pre3 + \beta_2 Treated_t \times Pre2 + \beta_3 Treated_t \times Pre1 \\
&\quad + \beta_4 Treated_t \times Post1 + \beta_5 Treated_t \times Post2 + \beta_6 Treated_t \times Post3 + \beta_7 size_{i,t} \\
&\quad + \beta_8 soe_{i,t} + \beta_9 roe_{i,t} + \beta_{10} lev_{i,t} + \beta_{11} salesgrowth_{i,t} + \beta_{12} cashholdings_{i,t} \\
&\quad + \beta_{13} duality_{i,t} + \beta_{14} boardsize_{i,t} + \beta_{15} top\_shareholdings_{i,t} + \beta_{16} hhi_{i,t} \\
&\quad + \beta_{17} ceoshare_{i,t} + \beta_{18} ret_{i,t} + \beta_{19} sigma_{i,t} + \beta_{20} share\_turnover_{i,t} + \beta_{21} roa\_volatility_{i,t} \\
&\quad + year\_dummies + industry\_dummies + city\_dummies + \varepsilon_{i,t} \qquad (6)
\end{aligned}$$

where *Pre*3, *Pre*2, *Pre*1, *Post*1, *Post*2, and *Post*3 are the year dummies for the 6-year periods.

Panel A of Table 3 presents the results of running Model (6). The coefficients on interaction terms, *Treat*×*Pre*3, *Treat*×*Pre*2, and *Treat*×*Pre*1, are not statistically significant, supporting the parallel trends assumption for our DID research design. The coefficients on interaction terms, *Treat*×*Post*3, *Treat*×*Post*2, and *Treat*×*Post*1, are all negative and statistically significant. These results indicate that the commercial reform with the application of digitalization affects stock price crash risk in each year of our post-event sample period. From the magnitude of their coefficients, we may infer that the effect of digitalization-applied commercial reform is amplified over the post-event sample years.

We also show in Figure 2 the dynamic economic effects of digitalization-applied commercial reform in different years. It reveals that before the implementation of commercial reform, the estimated coefficient is close to 0, with no obvious difference over the years. However, after the implementation of reform, the policy effect becomes prominent. This finding lends further support to the parallel trends assumption and suggests that the reduced risk of stock price crashes is attributed to the commercial reform other than potential omitted time-series factors.

4.2. Empirical results of the difference-in-differences regression

Table B of Table 3 reports the results of our stacked difference-in-differences regression (i.e., Model (4)). The coefficients on *Treat*×*Post* are negative and statistically significant at the 1% level for both *NCSKEW* and *DUVOL*. The point estimate on *Treat*×*Post* is -0.161 (-0.159), which accounts for 21.85% (31.42%) of one standard deviation of *NCSKEW* (*DUVOL*) for the matched sample and is economically significant. These results reject the null hypothesis H1 and suggest that firms subject to the digitalization-applied commercial reform experience a decrease in stock price crash risk relative to those unaffected by the reform. In addition, the regression results for control variables are in line with those reported in prior studies (e.g., Kim et al., 2014; Piotroski et al., 2015; Chen et al., 2016).

4.3. Robustness Tests

4.3.1. Control for firm-fixed effects and within-city correlations of residuals

There might be some unobserved firm-specific characteristics that affect firms' stock price crash risk. To allay this concern, we include firm-fixed effects and run both the univariate and multivariate regressions on *Treat*×*Post* for *NCSKEW* and *DUVOL*. Panel A of Table 4 presents the results. All the coefficients on *Treat*×*Post* are negative with the statistical significance level of 1%. The point estimate on *Treat*×*Post* in our multivariate regression for *NCSKEW* (*DUVOL*) is -0.145 (-0.117), which accounts for 19.67% (23.12%) of one standard deviation of *NCSKEW* (*DUVOL*) for the matched sample and is economically significant. These results substantiate that our baseline DID regression results are immune to the bias associated with potential omitted time-invariant factors.

The residuals of observations might be correlated across firms and years within each city. Thus, in addition to the control of city-fixed effects, we also cluster the standard errors of coefficients by city. Panel B reports the results, which appear qualitatively identical to our baseline results.

4.3.2. Test of coefficient stability

Following Altonji et al. (2005), we analyze coefficient stability to evaluate whether potential omitted factors would have driven our baseline regression results. The econometric rationale behind this analysis is that if the regression model adequately controls for the main determinants of dependent variable, any newly added control variable should exhibit a minimal correlation with the already included explanatory variables, and the additional control should not significantly alter the stability of coefficient estimates for those existing explanatory variables. In this context, the higher the stability of coefficients for the explanatory variables following an addition of control variables, the lower the likelihood that the regression model omits any key variable. Based on this reasoning, we test the coefficient stabililty in the following ways. First, we rank the 15 control variables based on the economic magnitude of their coefficients in the baseline regression analysis,[5] and take the top 60% as the main control variables and the rest as the additional control variables. Second, we run a DID regression with the 9 main control variables, and progressively introduce each of the additional control variables into this regression.

The results are reported in Panel C of Table 4. The progressive addition of control variables has no substantial effect on the significance levels of the DID coefficient, substantiating its stability and insensitivity to additional controls. Meanwhile, the absolute values of the ratios of the standardized selection on "unobservables" to the standardized selection on "observables", reported in the Columns (3) and (6) of Panel C, are all well below 1%.[6] From these results, it could be inferred that any plausibly omitted variables in our baseline regression, including those determining the timing of the reform implemented by local MSAs, are likely to be weakly correlated with explanatory variables and thus should not bias our DID estimator substantively.

4.3.3. Placebo test

As with previous studies (e.g., Ferrara et al., 2012; Alder et al., 2016), we conduct a placebo test to check whether our baseline regression results are free from the potential confounding effect of random factors or omitted variables. To this end, we first randomly assign our control firms into the treatment and control groups to generate a fake treatment group, $Treat^{fake}$, and associated fake

[5] The economic magnitude of the coefficient is estimated by the percentage change in the sample mean of the dependent variable in response to a one-standard-deviation change in the control variable.

[6] The ratio is calculated as: $\left|\frac{\beta^{F}-\beta^{R}}{\beta^{F}}\right|$, where β^{F} is the estimated coefficient of the core explanatory variable (i.e., *Treat*×*Post* in our case) in the regression that includes a selected number of main control variables, and β^{R} is the estimated coefficient of the core explanatory variable in the regression that includes the selected main control variables as well as the progressively added control variables. The lower the ratio, the stronger the explanatory power of the main control variables, and thus the lesser extent to which any omitted variable would bias the results for the core explanatory variable. The ratio less than 1% implies that omitted variables are unlikely to overturn the results and inferences for the core regressor.

commercial reform time, $Post^{fake}$, for each year. We repeat this trial for 1,000 times to enhance the efficacy of our placebo test. Figure 3 displays the distribution and p values of estimated coefficients on the interaction term, $Treat^{fake}$×$Post^{fake}$. The placebo DID estimators for both *NCSKEW* and *DUVOL* are normally distributed and centered around 0. Almost all the placebo DID coefficients are positioned to the right of the baseline DID coefficient (as depicted by the vertical dotted line) and have p values higher than 0.1. In our one-sample t-test, the mean value of the placebo DID estimators shows no statistically significant difference from 0 (p = 0.234 and 0.211). It can be inferred from these results that the reduction in stock price crash risk is not accidental or driven by omitted factors; rather, it is attributed to the effectiveness of commercial reform.

4.3.4. Alternative measures of stock price crash risk

Following previous research (e.g., Hutton et al., 2009; Kim et al., 2011), we generate two alternative measures of stock price crash risk, *CRASH*1 and *CRASH*2, to re-test our main hypothesis. *CRASH*1 equals 1 if a firm experiences at least one crash week in the fiscal year, and 0 otherwise. *CRASH*2 equals the natural logarithm of 1 plus the frequency of crash weeks of the firm during a fiscal year. We report the results for this robustness check in Panel D of Table 4. Column (1) (Column (3)) shows the results from using *CRASH*1 (*CRASH*2) to test the parallel trends assumption. The coefficients on *Treat*×*Pre*3*, Treat*×*Pre*2, and *Treat*×*Pre*1 are statistically nonsignificant, indicating that the parallel trend assumption is satisfied for the DID regression analysis. Column (2) (Column (4)) reports the results from using *CRASH*1 (*CRASH*2) to run the DID regression. *Treat*×*Post* takes on significantly negative coefficients, reinforcing the notion that firms subject to the digitalization-involved commercial reform enjoy lower stock price crash risk.

4.4. Mechanism tests for the association between digitalization-involved commercial reform and *stock price crash risk*

As discussed in Section 2.2, the digitalization-applied commercial reform might enhance the information transparency and monitoring of corporate commercial activities, leading to the decrease in firms' stock price crash risk. Therefore, information transparency and monitoring are arguably two channels through which the digitalization-involved commercial reform reduces stock price crash risk. To lend credence to these mechanisms, we conduct two tests.

We first test the mediating role of information transparency, which is measured by media news (*Media_coverage*). *Media_coverage* is computed as the natural logarithm of a firm's total number of media news in a fiscal year. A higher value of *Media_coverage* indicates higher information transparency.

We next test whether the enhanced monitoring of firms' commercial activities is another mechanism. Given the difficulty of directly measuring the monitoring level, we use three outcome-based measures, that is, related party transactions (*Related_transaction*), abnormal accruals (*Ab_accrual*) and other accounts receivable (*Other_receivable*), to capture the strength of monitoring on firms' commercial activities. These measurements are in line with previous research (e.g., Dechow et al., 1995; Jiang et al., 2010; Kohlbeck and Mayhew, 2017; Brockman et al., 2019). *Related_transaction* is computed as the natural logarithm of 1 plus the non-market-price transactions of commodities and services between a firm and its closely related business parties (i.e., its parent company or subsidiaries) during a fiscal year. *Ab_accrual* is the abnormal accruals of a firm for a fiscal year, which is estimated using the modified Jones model (Dechow et al., 1995). Firms could inflate accruals to hoard bad news arising from commercial activities (e.g., He and Ren, 2023). Thus, the stronger the monitoring, the lower the abnormal accruals which are likely associated with opportunistic bad news hoarding by the firms. *Other_receivable* is calculated as the amount of other accounts receivable, divided by the total assets of the firm, at the end of a fiscal year. A higher balance of other accounts receivable is likely associated with a greater likelihood of asset losses that result from corporate malpractices or malfeasances (e.g., Jiang et al., 2010; Brockman et al., 2019). In short, a higher value of *Related_transaction, Ab_accrual* or *Other_receivable* implies a lower degree of the monitoring strength that a firm confronts. We perform the mediation analysis by running the following regressions:

$$\begin{aligned} & Media\_coverage, Relate\_transaction, Ab\_accrual, \text{or } Other\_receivable_{i,t} = \\ & \quad \beta_0 + \beta_1 Treat_t \times Post_i + \beta_2 Treated_t + \beta_3 size_{i,t} + \beta_4 soe_{i,t} + \beta_5 roe_{i,t} + \beta_6 lev_{i,t} \\ & \quad + \beta_7 salesgrowth_{i,t} + \beta_8 cashholdings_{i,t} + \beta_9 duality_{i,t} + \beta_{10} boardsize_{i,t} \\ & \quad + \beta_{11} top\_sharesholdings_{i,t} + \beta_{12} hhi_{i,t} + \beta_{13} ceoshare_{i,t} + \beta_{14} ret_{i,t} + \beta_{15} sigma_{i,t} \\ & \quad + \beta_{16} share\_turnover_{i,t} + \beta_{17} roa\_volatility_{i,t} + year\_dummies \\ & \quad + industry\_dummies + city\_dummies + \varepsilon_{i,t} \qquad (7) \end{aligned}$$

$$\begin{aligned} & NCSKEW \text{ or } DUVOL_{i,t} = \\ & \quad \beta_0 + \beta_1 Media\_coverage, Related\_transaction, Ab\_accrual, \text{or } Other\_receivable_{i,t} \\ & \quad + \beta_2 size_{i,t} + \beta_3 soe_{i,t} + \beta_4 roe_{i,t} + \beta_5 lev_{i,t} + \beta_6 salesgrowth_{i,t} + \beta_7 cashholdings_{i,t} \\ & \quad + \beta_8 duality_{i,t} + \beta_9 boardsize_{i,t} + \beta_{10} top\_shareholdings_{i,t} + \beta_{11} hhi_{i,t} + \beta_{12} ceoshare_{i,t} \\ & \quad + \beta_{13} ret_{i,t} + \beta_{14} sigma_{i,t} + \beta_{15} share\_turnover_{i,t} + \beta_{16} roa\_volatility_{i,t} \\ & \quad + year\_dummies + industry\_dummies + city\_dummies + \varepsilon_{i,t} \qquad (8) \end{aligned}$$

where the mediator variables are *Media_coverage*, *Related_transaction*, *Ab_accrual*, and *Other_receivable*, which are defined in Appendix 2. If the mediating effect exists, the coefficients of *Treat*×*Post* for *Media_coverage* (*Related_transaction*, *Ab_accrual*, and *Other_receivable*) in Equation (7) should be positive (negative) and statistically significant at conventional levels, while their coefficients in Equation (8) should be significantly negative (positive).

Panel A of Table 5 reports the results of the mechanism tests for the information channel. Both the coefficients of *Treat*×*Post* for the first-stage regressions (reported in Columns (1)) and the mediator (*Media_coverage*) for the second-stage regression (reported in Columns (2), (3)) are statistically significant at the 1% level with the predicted signs. These results support the conjecture that the digitalization-involved commercial reform lowers stock price crash risk by enhancing information transparency. Panel B shows the results of the mechanism test for the monitoring channel. The coefficients on *Treat*×*Post* in Columns (1), (4), and (7) for the first-stage regressions are negative and statistically significant. The coefficients for the mediators (*Related_transaction*, *Ab_accrual*, and *Other_receivable*) in Columns (2), (3), (5), (6), (8), and (9) for the second-stage regressions are positive and statistically significant at the 1% level. Combined, these results corroborate that the increased strength of monitoring is another channel through which the digitalization-involved commercial reform reduces stock price crash risk.[7]

4.5. Cross-sectional analyses of the association between digitalization-applied commercial reform and stock price crash risk

We also explore how our baseline results vary under different circumstances. Apart from the government, firms might reshape their commercial processes and models by utilizing digital technologies such as artificial intelligence, blockchain, cloud computing, or big data analytics. Adopting digital technologies enables firms to better transmit their internal information to the government authorities in real time. This enhanced transmission enables the government to efficiently access more comprehensive and accurate information about different aspects of the firm, such as internal operations, production, and sales, thereby facilitating the digitalization-involved commercial reform to take even stronger attenuating effect on stock price crash risk. In this regard, the favorable impact of the commercial reform on reducing crash risk is expected to be more

[7] Including the interaction term *Treat*×*Post* in the second-stage regression, the estimations in both mechanism tests yield similar results: the coefficients of *Treat*×*Post* and those of the mediators (i.e., *Media_coverage, Related_transaction*, *Ab_accrual*, and *Other_receivable*) remain statistically significant with the predicted signs.

pronounced for firms with a higher level of digitalization.

Innovation plays a crucial role in maintaining competitive advantages, achieving commercial success, and ensuring sustainable development (Le et al., 2006; Jiménez-Jiménez and Sanz-Valle, 2011). Yet, pursuing innovation not only requires long-term substantial investments but also involves significant uncertainty as the innovation outcomes are often unpredictable *inter alia* for reasons of the rapid developments of technologies by competitors and the unforeseeable changes in market demands. Hence, monitoring firms that invest largely in innovation becomes challenging. Meanwhile, managers who are more familiar with their firms enjoy the information advantages in the productivity and value of innovation projects, not least compared to external investors (Aboody and Lev, 2000). This information asymmetry makes it even more difficult to monitor these firms. Consequently, the crash risk of such firms is plausibly higher. Given that the application of digital technologies for commercial reform reduces the crash risk by enhancing information transparency and external monitoring, we expect this impact to be particularly stronger for firms with a high level of innovation. Accordingly, the negative association between digitalization-involved commercial reform and stock price crash risk should be more prominent for firms with intensive innovation activities.

Firms with strong internal governance have effective internal control mechanisms to handle diverse risks, improve the quality of information disclosures, and avoid information distortion. Moreover, strong corporate governance facilitates more effective monitoring of managers, which helps deter managers' self-serving behaviors and decreases the likelihood of them concealing negative news (e.g., Jin et al., 2022). In contrast, weak internal governance implies an opaque information environment and weak monitoring. Hence, we expect that the digitalization-applied commercial reform has a more pronounced mitigating effect on stock price crash risk for firms with weaker internal governance, as the application of digital technologies helps improve information transparency and strengthen external monitoring mechanisms for these firms.

To test the moderating effects, we create binary variables based on the full-sample medians of corporate digitalization (*Digit* and *Digit*1), corporate innovation (*Innovation* and *Innovation*1), and internal governance (*CG* and *CG*1), respectively. *Digit* equals the natural logarithm of the total number of words related to digital technologies in the annual report of a firm during a fiscal year;[8] *Digit*1 equals the digital-technology-related intangible assets disclosed in a firm's annual report, divided by the total intangible assets of the firm during a fiscal year.[9] *Innovation* is computed by the research and development (R&D) expenditures of a firm, divided by its total sales during a fiscal year; *Innovation*1 is calculated by the natural logarithm of the number of invention patents that are applied by a firm in a year and subsequently granted by the China National Intellectual Property Administration (CNIPA). *CG* is calculated as the number of independent directors, divided by the total number of directors of a firm, at the end of a fiscal year; *CG*1 is calculated as the number of shares held by the board members of a firm, divided by the number of its total shares outstanding, at the end of a fiscal year. The moderator variables (*Dum_Digit*, *Dum_Digit*1, *Dum_Innovation*, *Dum_Innovation*1, *Dum_CG*, and *Dum_CG*1) equal 1 if the values of *Digit, Digit*1, *Innovation*, *Innovation*1, *CG*, and *CG*1 are higher than their sample medians, respectively,

[8] We take the following steps to construct the variable for corporate digitalization. First, we sort out the annual reports of listed companies and extract all the text content by virtue of Python Crawler technologies. Second, we use python open source with "Jiaba" participle features to extract the text content, which involves the keywords of digital technologies based on the semantic system of national-level digital economy-related policy documents in China. The text content on digital technologies is shown in Appendix 3, which include artificial intelligence, blockchain, cloud computing, and big data analytics. Finally, we count the frequency of keywords on the digital technologies and take the natural logarithm of it as the indicator for corporate digitalization (*Digit*).

[9] Pursuant to the Chinese accounting standards for enterprises, investments in digital technologies are recorded as intangible assets. These assets are named with keywords that are related to digital technologies, such as "digital platforms", "digital management system", "intelligent automation", or associated patents. We classify these asset items as "digital-technology-related intangible assets".

and 0 otherwise. We then augment the baseline model (4) by including the moderator variable and its interaction with *Treat*×*Post*.

Table 6 shows the results of the moderation analysis. Panel A, Panel B, and Panel C report the moderating effects of firm-level digitalization, corporate innovation, and internal governance, respectively, when using *NCSKEW* and *DUVOL* as the proxies for stock price crash risk. The coefficients on ternary interaction terms are all statistically significant with the expected signs, indicating that the digitalization-involved commercial reform has a more prominent attenuating effect on stock price crash risk for firms with higher levels of digitalization and innovation and for those with weaker internal governance.

We further visualize the moderating effects of the three moderators in Figure 4, Figure 5, and Figure 6, respectively. The moderation effect is captured by the interaction terms between the moderator and the interaction term, *Treat*×*Post*. As depicted in Figures 4-6, the digitalization-involved commercial reform has a restraining effect on firms' stock price crash risk, regardless of the level of moderators. However, for firms with greater digitalization, higher innovation, and weaker internal governance, the mitigating effect of commercial reform on stock price crash risk is more evident. These results are thus consistent with our predictions.[10]

5. Conclusion

[10] In addition, we test whether our baseline results differ between state-owned firms and non-state-owned firms. To this end, we generate a moderator variable (*soe*) that indicates whether a firm is state-owned, augment Model (4) with the moderator variable (*soe*) and its interaction with *Treat*×*Post*, and run the augmented regression model. In results not tabulated, the coefficients on the ternary interaction term *Treat*×*Post*×*soe* are statistically nonsignificant, while those on the interaction term *Treat*×*Post* remain negative and statistically significant at the 1% level. This suggests that there is no statistically significant difference in the negative coefficients of *Treat*× *Post* between the state-owned and non-state-owned firms, and thus that the attenuating effect of commercial reform on crash risk does not vary with the firms' state ownership.

In recent years, the Chinese government has applied digital technologies in commercial reform that is aimed at optimizing commercial environments for sustainable economic growth. To assess the effectiveness of this digitalization-involved commercial reform, we examine its impact on firms' stock price crash risk. We provide robust evidence of a causal link between digitalization-involved commercial reform and a reduction in firms' stock price crash risk. Our mediating analyses reveal that the reform improves commercial information transparency as well as monitoring of corporate commercial activities and thereby lowers the stock price crash risk of firms. We also find that higher levels of corporate digitalization and innovation and weaker internal governance amplify the mitigating effect of digitalization-involved commercial reform on crash risk.

Our findings underline the positive impact of digitalization-involved commercial reform on information environments and emphasize its potential in facilitating well-organized commercial activities and mitigating risks. In this regard, the government should make good use of digital technologies, ideally in a way that minimizes their associated risks and costs, in order to improve firms' commercial information transparency and effectively monitor their commercial activities. In addition, our finding as to the strengthening moderating effect of firm-level digitalization also offers valuable implications for the government. To better realize the economic benefits of digitalization-involved commercial reform, the government may encourage firms to actively integrate digital technologies into corporate business structures and activities.

References

Aboody, D., Lev, B., 2000. Information asymmetry, R&D, and insider gains. The Journal of Finance. 55 (6), 2747-2766.

Acemoglu, D., 2002. Directed technical change. The Review of Economic Studies. 69 (4), 781-809.

Alder, S., Shao, L., Zilibotti, F., 2016. Economic reforms and industrial policy in a panel of Chinese cities. Journal of Economic Growth. 21 (4), 305-349.

Altonji, J. G., Elder, T. E., Taber, C. R. 2005. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. Journal of Political Economy, 113 (1), 151-184.

Baker, A. C., Larcker, D. F., Wang, C. C. Y., 2022. How much should we trust stacked difference-in-differences estimates? Journal of Financial Economics. 144 (2), 370-395.

Beck, T., Levine, R., Levkov A., 2010. Big bad banks? The winners and losers from bank deregulation in the United States. Journal of Finance. 65 (5), 1637-1667.

Biddle, G. C., Hilary, G., Verdi, R. S., 2009. How does financial reporting quality relate to investment efficiency? Journal of Accounting and Economics. 48 (2-3), 112-131.

Blichfeldt, H., Faullant, R., 2021. Performance effects of digital technology adoption and product & service innovation. A process-industry perspective. Technovation. 105, 102275.

Brockman, P., Firth, M., He, X., Rui, O., 2019. Relationship-based resource allocations: Evidence from the use of Guanxi during SEOs. Journal of Financial and Quantitative Analysis. 54 (3), 1193-1230.

Chen, W., Srinivasan, S., 2023. Going digital: implications for firm value and performance. Review of Accounting Studies. 1-47.

Chen, J., Hong, H., Stein, J. C., 2001. Forecasting crashes, trading volume, past returns, and conditional skewness in stock prices. Journal of Financial Economics. 61 (3), 345-381.

Chen, C., Kim, J. B., Yao, L., 2016. Earnings smoothing: Does it exacerbate or constrain stock price crash risk? Journal of Corporate Finance. 42, 36-54.

Chen, W., Zhang, L., Jiang, P., Meng, F., Sun, Q., 2022. Can digital transformation improve the information environment of the capital market? Evidence from the analysts' prediction behaviour. Accounting & Finance. 62 (2), 2543-2578.

Ciampi, F., Demi, S., Magrini, A., Marzi, G., Papa, A., 2021. Exploring the impact of big data analytics capabilities on business model innovation: The mediating role of entrepreneurial orientation. Journal of Business Research. 123, 1-13.

Cong, L. W., He, Z., 2019. Blockchain disruption and smart contracts. The Review of Financial Studies. 32 (5), 1754-1797.

Dechow, P. M., Sloan, R. G., Sweeney, A. P., 1995. Detecting earnings management. The Accounting Review. 70, 19-225.

Dimson, E., 1979. Risk measurement when shares are subject to infrequent trading. Journal of Financial Economics. 7 (2), 197-226.

Dinev, T., Hart, P., 2006. An extended privacy calculus model for e-commerce transactions. Information Systems Research. 17 (1), 61-80.

Drake, M. S., Myers, J. N., Myers, L. A., 2009. Disclosure quality and the mispricing of accruals and cash flow, Journal of Accounting, Auditing & Finance. 24 (3), 357-384.

Fan, J. P. H., Wong, T. J., 2005. Do external auditors perform a corporate governance role in emerging markets? Evidence from East Asia. Journal of Accounting Research. 43 (1), 35-72.

Ferrara, E. L., Duryea, S., Chong, A. E., 2012. Soap operas and fertility: Evidence from Brazil.

American Economic Journal: Applied Economics. 4 (4), 1-31.

Ferreira, J. J. M., Fernandes, C. I., Ferreira, F. A. F., 2019. To be or not to be digital, that is the question: Firm innovation and performance. Journal of Business research. 101, 583-590.

Gallery, G., Gallery, N., Supranowicz, M., 2008. Cash-based related party transactions in new economy firms. Accounting Research Journal. 21 (2), 147-166.

Gomber, P., Kauffman, R. J., Parker, C., Weber, B. W., 2018. On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. Journal of Management Information Systems. 35 (1), 220-265.

He, G., Bai, L., Ren, H. M., 2019. Analyst coverage and future stock price crash risk. Journal of Applied Accounting Research. 20 (1), 63-77.

He, G., Ren, H. M., 2023. Are financially constrained firms susceptible to a stock price crash? The European Journal of Finance. 29 (6), 612-637.

Hutton, A. P., Marcus A. J., Tehranian H., 2009. Opaque financial report, R2, and crash risk. Journal of Financial Economics. 94 (1), 67-86.

Jiang, G., Lee, C. M. C., Yue, H., 2010. Tunneling through intercorporate loans: The China experience, Journal of Financial Economics. 98 (1), 1-20.

Jiménez-Jiménez, D., Sanz-Valle, R., 2011. Innovation, organizational learning, and performance. Journal of Business Research. 64, 408-417.

Jin, L., Myers, S. C., 2006. R2 around the world: New theory and new tests. Journal of Financial Economics. 79 (2), 257-292.

Jin, H. M., Su, Z. Q., Wang, L., Xiao, Z., 2022. Do academic independent directors matter? Evidence from stock price crash risk, Journal of Business Research. 144, 1129-1148.

Kohlbeck, M., Mayhew, B. W., 2017. Are related party transactions red flags? Contemporary Accounting Research. 34 (2), 900-928.

Kothari, S. P., Shu, S., Wysocki, P. D., 2009. Do managers withhold bad news? Journal of Accounting Research. 47, 241-276.

Kim, J. B., Li, Y., Zhang, L., 2011. Corporate tax avoidance and stock price crash risk: Firm-level analysis. Journal of Financial Economics. 100 (3), 639-662.

Kim, Y., Li, H., Li, S., 2014. Corporate social responsibility and stock price crash risk. Journal of Banking and Finance. 43 (1), 1-13.

Lai, S. M., Liu, C. L., Wang, T., 2014. Increased disclosure and investment efficiency. Asia-Pacific Journal of Accounting & Economics. 21 (3), 308-327.

Le, S. A., Walters, B., Kroll, M., 2006. The moderating effects of external monitors on the relationship between R&D spending and firm performance, Journal of Business Research. 59 (2), 278-287.

Lee, H. L., Lee, H., 2015. Effect of information disclosure and transparency ranking system on mispricing of accruals of Taiwanese firms, Review of Quantitative Finance and Accounting. 44, 445-471.

Leuven, E., Sianesi, B., 2018. Psmatch2: Stata module to perform full mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components.

Loebbecke, C., Picot, A., 2015. Reflections on societal and business model transformation arising from digitalization and big data analytics: A research agenda. The Journal of Strategic Information Systems. 24 (3), 149-157.

Luo, Y., 2022. A general framework of digitization risks in international business. Journal of International Business Studies. 53 (2), 344-361.

Matarazzo, M., Penco, L., Profumo, G., Quaglia, R., 2021. Digital transformation and customer value creation in made in Italy SMES: A dynamic capabilities perspective, Journal of Business Research. 123, 642-656.

Xu, L., Liu, Q., Li, B., Ma, C., 2022. Fintech business and firm access to bank loans. Accounting and Finance. 62 (4), 4381-4421.

Piotroski, J. D., Wong, T. J., 2012. Institutions and information environment of Chinese listed firms. Capitalizing China. University of Chicago Press. 201-242.

Piotroski, J. D., Wong, T. J., Zhang, T., 2015. Political incentives to suppress negative information: Evidence from Chinese listed firms. Journal of Accounting Research. 53 (2), 405-459.

Roberts, M. R., Whited T. M., 2013. Endogeneity in empirical corporate finance. Handbook of the Economics of Finance. (1), 493-572.

Rosati, P., Gogolin, F., Lynn, T., 2022. Cyber-security incidents and audit quality. European Accounting Review. 31 (3), 701-728.

Roth, J., Sant'Anna, P. H., Bilinski, A., Poe, J., 2023. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. Journal of Econometrics. 235 (2), 2218-2244.

Wu, L., Lou, B., Hitt, L., 2019. Data analytics supports decentralized innovation. Management Science. 65 (10), 4863-4877.

Appendix 1: Sample selection

The sample selection procedure	No. of observations	No. of firms
Observations of the population of companies listed on the Shenzhen or Shanghai Stock Exchanges for the period 2011-2019	26,345	4,016
Less: observations of firms labeled with ST, ST *, or PT	(1,935)	(234)
Less: observations of firms in the financial industry	(512)	(89)
Less: observations of firms cross-listed overseas	(35)	(9)
Less: observations of loss firms	(58)	(18)
Less: observations with missing values in regressors	(7,568)	(1,089)
Sample before propensity-score matching	16,237	2,577
Final sample after propensity-score matching	7,072	1,156

Appendix 2: Summary of variable definitions

Variables	Definitions
NCSKEW	A measure of stock price crash risk that captures the negative skewness of firm-specific weekly stock returns over a fiscal year. See Equation (2) for detail.
DUVOL	The down-to-up volatility measure of stock price crash risk, calculated as the natural logarithm of the ratio of the standard deviation of firm-specific weekly stock returns in the "down" weeks to that in the "up" weeks. See Equation (3) for detail.
*CRASH*1	1 if a firm has at least one crash week in a fiscal year, and 0 otherwise. The crash week is defined as a week when the firm-specific weekly stock return falls by 3.2 standard deviations of the weekly returns for the year.
*CRASH*2	The natural logarithm of 1 plus the frequency of crash weeks of a firm during a fiscal year.
Treat	1 (0) for a treatment (control) firm. The treatment firm is defined as subject to the digitalization-involved commercial reform in which the Market Supervision Administration was established to introduce digital commercial registration system for improving information environments and monitoring on commercial activities of firms. The control firm is defined as not subject to the digitalization-involved commercial reform in the six-year period centered at the beginning of the year of the reform for the treatment firm, nor before the period.
Post	1 (0) if a treatment firm is in the three-year period since (before) the digitalization-involved commercial reform took place.
Related_transaction	The natural logarithm of 1 plus the non-market-price transactions of commodities and services between a firm and its closely related business parties (i.e., its parent company or subsidiaries) during a fiscal year.
Other_receivable	The amount of other accounts receivable of a firm, divided by the total assets of the firm, at the end of a fiscal year.
Media_coverage	The natural logarithm of the total number of media news about a firm in a fiscal year.
Ab_accrual	The abnormal accruals of a firm for a fiscal year, which are estimated by using the modified Jones model (Dechow et al., 1995).
Digit	The natural logarithm of the total number of words related to digital technologies in the annual report of a firm during a fiscal year, and 0 if there is no such word in the annual report.
*Digit*1	The digital-technology-related intangible assets, divided by the total intangible assets of a firm, during a fiscal year.
Innovation	The R&D expenditures by a firm, divided by the total sales of the firm, during a fiscal year.
*Innovation*1	The natural logarithm of the number of invention patents that are applied by a firm in a year and subsequently granted by the China National Intellectual Property Administration.
CG	The number of independent directors, divided by the total number of directors on the board of a firm, at the end of a fiscal year.
*CG*1	The number of shares held by the board members of a firm, divided by the number of its total shares outstanding, at the end of a fiscal year.
size	The natural logarithm of the total assets of a firm at the end of a fiscal year.
soe	1 if a firm is a state-owned enterprise (i.e., the firm of which the largest ultimate shareholder pertains to a government entity), and 0 otherwise.

roe	Return on equity, calculated as the net profit of a firm for a fiscal year, divided by the total assets of the firm at the end of the fiscal year.
lev	The total debt of a firm, divided by the total assets of the firm, at the end of a fiscal year.
salesgrowth	The difference between the firm's sales for the current fiscal year and the sales for the previous year, divided by the sales for the previous year.
cashholdings	The cash flows of a firm, divided by the total assets of the firm, at the end of a fiscal year.
duality	1 if the CEO of a firm and the chairman/chairwoman of the board are the same person for a fiscal year.
boardsize	The natural logarithm of the total number of board members of a firm at the end of a fiscal year.
top_shareholdings	The number of shares held by the largest shareholder of a firm, divided by the number of its total shares outstanding, at the end of a fiscal year.
hhi	The Herfindahl-Hirschman Index computed on firms' sales for each industry in a fiscal year; industries are classified based on the industrial classification guidance released by the China Securities Regulatory Commission in 2012.
ceoshare	The percentage of outstanding shares owned by a firm's CEO at the end of a fiscal year.
ret	The mean of firm-specific weekly stock returns in a fiscal year.
sigma	The standard deviation of firm-specific weekly stock returns in a fiscal year.
share_turnover	The detrended stock trading volume, calculated as the average monthly share turnover for the current fiscal year minus the average monthly share turnover for the previous fiscal year. The monthly share turnover is the monthly trading volume divided by the number of the total floating shares in the month.
roa_volatility	The standard deviation of a firm's returns on assets for the recent five fiscal years.

Appendix 3: Glossary of corporate digitalization

Digitalization	Specific digital technologies
Artificial intelligence technology	Artificial intelligence, business intelligence, image understanding, investment decision support system, intelligent data analysis, machine learning, deep leaning, intelligent robotics, semantic search, biometric technology, face recognition, voice recognition, identity verification, autonomous diving, and natural language processing
Blockchain technology	Blockchain, digital currency, distributed computing, differential privacy technology, and smart financial contract
Cloud computing technology	Cloud computing, stream computing, graph computing, in-memory computing, multi-party security computing, brain-like computing, green computing, cognitive computing, fusion architecture, billion level concurrency, exabyte storage, Internet of things, and information physics system
Big data technology	Big data, data mining, text mining, data visualization, heterogeneous data, credit reporting, augmented reality, mixed reality, and virtual reality

Table 1: Propensity-score matching between the treatment and control firms

Panel A: Logit regressions run by year for estimating propensity scores based on the pre-matched sample

Variables	(1) 2014	(2) 2015	(3) 2016	(4) 2017
size$_t$	-0.1531***	-0.0831***	-0.0912*	-0.0426
	(-2.7627)	(-4.4918)	(-1.6741)	(-0.5204)
roe$_t$	1.8439***	-0.0548	0.4761	0.4937***
	(3.0658)	(-0.1033)	(0.6149)	(3.4030)
lev$_t$	0.4215	-0.0789	-0.1989	-0.1372
	(1.2213)	(-0.2174)	(-0.5679)	(-0.2663)
salesgrowth$_t$	-0.0281	-0.0068	0.0072***	0.0007
	(-0.7767)	(-0.7332)	(3.0177)	(0.0658)
cashholdings$_t$	-4.6840	-1.3945	4.3861*	7.0830***
	(-1.3649)	(-0.4237)	(1.6510)	(4.5412)
boardsize$_t$	-0.0558	-0.0547	0.2190***	0.1685
	(-0.1877)	(-0.1928)	(7.7989)	(0.4102)
roa_volatility$_t$	4.3793	3.0264	-1.3763	-6.5241
	(1.5004)	(1.0426)	(-0.6181)	(-1.6002)
Observations	2,022	2,125	2,306	2,137
Pseudo R^2	0.010	0.009	0.013	0.007
Industry-fixed effects	included	included	included	included
City-fixed effects	included	included	included	included

Notes: Panel A of Table 1 reports the results of the logit regression, which is run by year for estimating propensity scores based on the pre-matched sample. The sample period ranges from 2011 to 2019. We use seven covariates - *size*, *roe*, *lev*, *salesgrowth*, *cashholdings*, *boardsize*, and *roa_volatility*. The definitions of all variables are provided in Appendix 2. The treatment indicator variable, *Treat*, equals 1 (0) for a treatment (control) firm. The treatment firm is defined as subject to the digitalization-involved commercial reform in which the Market Supervision Administration was established to introduce digital commercial registration system for improving information environments and monitoring on commercial activities of firms. The control firm is not subject to the digitalization-involved commercial reform in the six-year period centered on the beginning of the year of reform for the treatment firm, nor before the period. Industry-fixed effects and city-fixed effects are controlled in each regression, but their results are not reported for simplicity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel B: Tests of covariate balance for the post-matched sample

Variables	(1) 2014	(2) 2015	(3) 2016	(4) 2017
size$_t$	0.0154	-0.0461	-0.0507	-0.0357
	(0.2269)	(-0.7385)	(-0.8618)	(-0.3790)
roe$_t$	0.4841	0.1686	0.2704	-0.0250
	(0.6626)	(0.3175)	(0.3626)	(-0.0194)
lev$_t$	-0.2889	0.2375	0.2313	0.3527
	(-0.7415)	(0.6256)	(0.6219)	(0.5529)
salesgrowth$_t$	0.0357	0.0038	-0.0003	0.0907
	(0.9044)	(0.6704)	(-0.0645)	(1.3699)
cashholdings$_t$	0.6760	0.2070	0.9401	-0.8769
	(0.6642)	(0.2117)	(1.0147)	(-0.5123)
boardsize$_t$	0.1398	0.0814	0.1754	-0.3518
	(0.3888)	(0.2485)	(0.5541)	(-0.6909)
roa_volatility$_t$	0.3230	-0.6001	-0.1815	0.8991
	(0.4462)	(-0.8156)	(-0.1275)	(0.4323)
Observations	1,032	1,246	1,122	1,024
Pseudo R^2	0.004	0.006	0.004	0.022
Industry-fixed effects	included	included	included	included
City-fixed effects	included	included	included	included

Notes: Panel B of Table 1 reports the results from testing the covariate balance for the matched sample used in the difference-in-differences regression of stock price crash risk. We use seven covariates - *size*, *roe*, *lev*, *salesgrowth*, *cashholdings*, *boardsize*, and *roa_volatility*. The definitions of all variables are provided in Appendix 2. The treatment indicator variable, *Treat*, equals 1 (0) for a treatment (control) firm. The treatment firm is defined as subject to the digitalization-involved commercial reform in which the Market Supervision Administration was established to introduce digital commercial registration system for improving information environments and monitoring on commercial activities of firms. The control firm is not subject to the digitalization-involved commercial reform in the six-year period centered at the beginning of the year of the reform for the treatment firm, nor before the period. We follow Leuven and Sianesi (2018) to match each treatment firm, with replacement, with a control firm by using the closest propensity score within a caliper of 1% for each year. Industry-fixed effects and city-fixed effects are controlled in each regression, but their results are not reported for simplicity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 2: Univariate statistics

Panel A: Summary statistics of variables

Variables	N	Mean	Min.	10%	25%	Median	75%	90%	Max.	Std. Dev.
NCSKEW	7,072	-0.243	-2.788	-1.150	-0.643	-0.212	0.189	0.597	2.267	0.737
DUVOL	7,072	-0.195	-1.686	-0.790	-0.486	-0.162	0.165	0.474	1.429	0.506
*CRASH*1	7,072	0.481	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.500
*CRASH*2	7,072	0.552	0.000	0.000	0.000	0.000	0.423	2.156	4.587	0.429
Treat	7,072	0.511	0.000	0.000	0.000	1.000	1.000	1.000	1.000	0.334
Ab_accrual	7,072	-0.017	-0.326	-0.096	-0.045	-0.006	0.030	0.053	0.066	0.062
Media_coverage	7,072	4.279	0.693	3.132	3.555	4.246	4.344	5.123	8.344	1.047
Related_transaction	7,072	6.159	0.000	0.000	0.000	0.000	18.000	25.279	28.788	10.782
Other_receivable	7,072	0.347	0.000	0.096	0.201	0.345	0.490	0.599	0.806	0.187
Digit	7,072	1.028	0.000	0.000	0.000	0.693	1.792	2.996	6.252	1.278
*Digit*1	7,072	0.003	0.000	0.000	0.000	0.001	0.003	0.007	0.045	0.007
Innovation	7,072	0.040	0.000	0.002	0.012	0.032	0.049	0.082	1.259	0.049
*Innovation*1	7,072	0.652	0.000	0.000	0.000	0.000	1.099	2.079	8.034	1.007
CG	7,072	0.405	0.000	0.073	0.216	0.413	0.587	0.711	0.890	0.232
CG1	7,072	0.076	0.000	0.000	0.013	0.054	0.091	0.133	0.652	0.125
size	7,072	22.275	18.964	20.775	21.336	22.082	23.017	24.044	26.297	1.281
soe	7,072	0.411	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.492
roe	7,072	0.058	-1.595	0.005	0.031	0.070	0.113	0.164	0.377	0.151
lev	7,072	0.412	0.044	0.156	0.277	0.454	0.565	0.665	0.901	0.189
salesgrowth	7,072	0.395	-0.772	-0.181	-0.030	0.135	0.424	0.992	12.455	1.131
cashholdings	7,072	0.044	-0.208	-0.037	0.006	0.043	0.084	0.126	0.256	0.068
duality	7,072	0.248	0.000	0.000	0.000	0.000	0.000	1.000	1.000	0.432
boardsize	7,072	2.152	1.609	1.946	2.079	2.197	2.197	2.398	2.708	0.198
top_shareholdings	7,072	36.186	8.260	17.710	24.195	34.650	46.435	56.850	75.790	14.978
hhi	7,072	0.054	0.001	0.007	0.018	0.038	0.073	0.122	0.304	0.053
ceoshare	7,072	0.003	0.000	0.000	0.001	0.002	0.004	0.007	0.017	0.041
ret	7,072	0.001	-0.034	-0.010	-0.006	-0.001	0.006	0.014	0.081	0.011
sigma	7,072	0.062	0.018	0.038	0.045	0.056	0.071	0.097	0.243	0.026
share_turnover	7,072	-0.019	-0.251	-0.079	-0.033	-0.009	0.006	0.026	0.152	0.049
roa_volatility	7,072	0.041	0.001	0.006	0.010	0.019	0.037	0.078	0.748	0.073

Notes: Panel A of Table 2 reports the descriptive statistics of all variables used in the multivariate tests of the association between the digitalization-involved commercial reform and stock price crash risk. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2. The sample period ranges from 2011 to 2019. Observations that have missing values in any of the regressors are excluded from the samples used for the multivariate tests.

Panel B: Correlation matrix

Variables	NCSKEW	DUVOL	Treat×Post	size	soe	roe	lev	salesgrowth	cashholdings	duality	boardsize	top_shareholdings	hhi	ceoshare	ret	sigma	share_turnover	roa_volatility
NCSKEW	1.000																	
DUVOL	0.879***	1.000																
Treat×Post	0.005	0.003	1.000															
size	0.097***	0.121***	-0.043***	1.000														
soe	0.111***	0.115***	-0.100***	0.410***	1.000													
roe	0.103	0.111***	-0.047***	0.590***	0.350***	1.000												
lev	-0.011	-0.001	0.008	0.086***	-0.008	-0.101***	1.000											
salesgrowth	-0.004	-0.003	-0.004	0.027***	0.013	0.092***	0.040***	1.000										
cashholdings	-0.006	-0.003	0.030***	0.052***	0.029***	-0.148***	0.283***	-0.097***	1.000									
duality	-0.049***	-0.053***	0.050***	-0.209***	-0.292***	-0.166***	-0.008	-0.020**	-0.024***	1.000								
boardsize	0.051***	0.053***	-0.033***	0.277***	0.294***	0.175***	0.039***	-0.028***	0.054***	-0.182***	1.000							
top_shareholdings	0.067***	0.061***	-0.012	0.225***	0.208***	0.106***	0.096***	0.014*	0.076***	-0.043***	0.028***	1.000						
hhi	-0.010	-0.008	0.005	-0.034***	-0.055***	0.080***	-0.042***	0.132***	-0.259***	0.022***	-0.059***	-0.028***	1.000					
ceoshare	-0.050	-0.052***	0.042***	0.023***	-0.144***	-0.067***	0.003	-0.015*	0.001	0.069***	-0.013	0.000	0.088***	1.000				
ret	0.122***	0.124***	0.134***	-0.043***	-0.058***	-0.035***	0.070***	0.041***	0.086***	0.022***	-0.058***	-0.005	-0.008	0.004	1.000			
sigma	0.094***	0.093***	0.100***	-0.212***	-0.161***	-0.082***	-0.109***	0.058***	-0.045***	0.082***	-0.143***	-0.089***	0.057***	0.013	0.593***	1.000		
share_turnover	0.060***	0.063***	0.023***	0.225***	0.167***	0.171***	-0.057***	0.024***	0.073***	-0.105***	0.058***	-0.031***	-0.030***	0.028***	0.383***	0.243***	1.000	
roa_volatility	-0.006	0.000	0.006	-0.090***	-0.037***	-0.073***	-0.104***	-0.012	-0.008	0.012	-0.039***	-0.034***	0.061***	0.046***	0.057***	0.081***	0.047***	1.000

Notes: Panel B of Table 2 provides the Spearman correlation coefficients for all variables involved in the baseline regression regarding the relationship between digitalization-involved commercial reform and stock price crash risk. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 3: Baseline regression

Panel A: Multivariate test of the parallel trends assumption

Variables	(1) Dependent variable = $NCSKEW_t$	(2) Dependent variable = $DUVOL_t$
Treat×*Pre*3	0.022	0.040
	(0.447)	(1.220)
Treat×*Pre*2	0.037	0.034
	(0.762)	(0.984)
Treat×*Pre*1	0.045	0.031
	(0.982)	(0.943)
Treat×*Post*1	-0.085**	-0.065**
	(-2.278)	(-2.549)
Treat×*Post*2	-0.100**	-0.076**
	(-2.133)	(-2.364)
Treat×*Post*3	-0.111**	-0.085***
	(-2.321)	(-2.605)
$size_t$	0.045***	0.043***
	(5.426)	(7.369)
soe_t	0.087***	0.055***
	(5.382)	(4.845)
roe_t	0.054	0.029
	(1.181)	(0.967)
lev_t	-0.142**	-0.082*
	(-1.964)	(-1.649)
$salesgrowth_t$	-0.019***	-0.013***
	(-2.655)	(-2.914)
$cashholdings_t$	-0.122	-0.096
	(-1.200)	(-1.327)
$duality_t$	-0.020	-0.014
	(-1.304)	(-1.299)
$boardsize_t$	0.058*	0.032
	(1.713)	(1.331)
$top\_shareholdings_t$	0.001***	0.001**
	(3.006)	(2.074)
hhi_t	-0.192	-0.134
	(-1.616)	(-1.615)
$ceoshare_t$	-0.121	-0.077
	(-0.637)	(-1.206)
ret_t	10.719***	8.970***
	(10.401)	(11.665)
$sigma_t$	5.559***	3.106***
	(11.402)	(9.029)
$share\_turnover_t$	-0.001	0.007
	(-0.009)	(0.070)
$roa\_volatility_t$	-0.053	0.024
	(-0.494)	(0.350)
Constant	-1.191***	-1.009***
	(-6.077)	(-7.336)
Observations	7,072	7,072
Adj. R^2	0.098	0.100
Year-fixed effects	included	included
Industry-fixed effects	included	included
City-fixed effects	included	included

Notes: Table A of Table 3 presents the results of the multivariate test of the parallel trends assumption for the difference-in-differences regression of the association between digitalization-involved commercial reform (*Treat×Post*) and stock price crash risk (*NCSKEW* and *DUVOL*). The treatment indicator variable, *Treat*, equals 1 (0) for a treatment (control) firm. The treatment firm is defined as subject to the digitalization-involved commercial reform in which the Market Supervision Administration was established to introduce digital commercial registration system for improving information environments and monitoring on commercial activities of firms. The control firm is not subject to the digitalization-involved commercial reform in the six-year period centered at the beginning of the year of the reform for the treatment firm, nor before the period. *Pre*3, *Pre*2, *Pre*1, *Post*1, *Post*2, and *Post*3 are the year dummies for the 6-year periods. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. Year dummies, industry dummies, and city dummies are included in each regression, but their results are not reported for brevity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2.

Panel B: Difference-in-differences (DID) regression as to the association between digitalization-involved commercial reform and stock price crash risk

Variables	(1) Dependent variable = *NCSKEW*$_t$	(2) Dependent variable = *DUVOL*$_t$
Treat×Post	-0.161***	-0.159***
	(-12.579)	(-20.129)
Treat	0.639***	0.692***
	(10.961)	(19.039)
size$_t$	0.046***	0.043***
	(5.513)	(7.470)
soe$_t$	0.087***	0.055***
	(5.378)	(4.834)
roe$_t$	0.051	0.027
	(1.122)	(0.894)
lev$_t$	-0.145	-0.084
	(-1.005)	(-1.102)
salesgrowth$_t$	-0.019***	-0.013***
	(-2.665)	(-2.917)
cashholdings$_t$	-0.123	-0.097
	(-1.209)	(-1.335)
duality$_t$	-0.019	-0.013
	(-1.270)	(-1.249)
boardsize$_t$	0.058*	0.031
	(1.693)	(1.309)
top_shareholdings$_t$	0.001***	0.001**
	(3.040)	(2.115)
hhi$_t$	-0.196	-0.137*
	(-1.644)	(-1.655)
ceoshare$_t$	-0.122***	-0.078***
	(-4.697)	(-4.268)
ret$_t$	10.735***	8.974***
	(10.451)	(11.701)
sigma$_t$	5.553***	3.106***
	(11.385)	(9.031)
share_turnover$_t$	-0.004	0.005
	(-0.028)	(0.051)
roa_volatility$_t$	-0.051	0.026
	(-0.473)	(0.371)
Constant	-1.304***	-1.097***
	(-6.865)	(-8.192)
Observations	7,072	7,072
Adj. R^2	0.119	0.120
Year-fixed effects	included	included
Industry-fixed effects	included	included
City-fixed effects	included	included

Notes: Table B of Table 3 reports the OLS regression results for the association between digitalization-involved commercial reform (*Treat×Post*) and stock price crash risk (*NCSKEW* and *DUVOL*). *Treat* equals 1 (0) for a treatment (control) firm. The treatment firm is defined as subject to the digitalization-involved commercial reform in which the Market Supervision Administration was established to introduce digital commercial registration system for improving information environments and monitoring on commercial activities of firms. The control firm is not subject to the digitalization-involved commercial reform in the six-year period centered at the beginning of the year of the reform for the treatment firm, nor before the period. *Post* is the time indicator variable that equals 1 (0) if a treatment firm is in the three-year period since (before) the digitalization-involved commercial reform took place. The interaction term, *Treat×Post*, captures the impact of digitalization-involved commercial reform on stock price crash risk. The sample period ranges from 2011 to 2019. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2. Year dummies, industry dummies, and city dummies are included in each regression, but their results are not reported for brevity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 4: Robustness tests of baseline results

Panel A: Inclusion of firm-fixed effects in the DID regression

Variables	(1) Dependent variable =$NCSKEW_t$	(2) Dependent variable = $DUVOL_t$	(3) Dependent variable = $NCSKEW_t$	(4) Dependent variable = $DUVOL_t$
Treat×Post	-0.174***	-0.147***	-0.145***	-0.117***
	(-22.781)	(-33.297)	(-22.579)	(-26.476)
$size_t$			0.709***	0.576***
			(22.177)	(29.019)
soe_t			-0.008	0.030**
			(-0.357)	(2.088)
roe_t			0.074	0.041
			(0.691)	(0.806)
lev_t			0.126*	0.060
			(1.722)	(1.042)
$salesgrowth_t$			-0.053	-0.040
			(-0.654)	(-0.744)
$cashholdings_t$			-0.025***	-0.015***
			(-2.745)	(-2.841)
$duality_t$			-0.202*	-0.076
			(-1.817)	(-0.932)
$boardsize_t$			0.006	0.005
			(0.199)	(0.224)
$top\_shareholdings_t$			0.003	-0.014
			(0.040)	(-0.285)
hhi_t			0.005***	0.002***
			(3.614)	(2.750)
$ceoshare_t$			-0.085	-0.062
			(-0.645)	(-0.804)
ret_t			0.058	0.041*
			(1.527)	(1.678)
$sigma_t$			10.659***	9.020***
			(8.953)	(10.032)
$share\_turnover_t$			5.681***	3.072***
			(9.464)	(6.808)
$roa\_volatility_t$			-0.078	-0.047
			(-0.438)	(-0.410)
Constant	0.228***	0.148***	-0.205	-0.820**
	(58.269)	(54.314)	(-0.390)	(-2.467)
Observations	7,072	7,072	7,072	7,072
Adj. R^2	0.067	0.255	0.113	0.121
Year-fixed effects	included	included	included	included
Firm-fixed effects	included	included	included	included
City-fixed effects	included	included	included	included

Notes: Table A of Table 4 reports the firm-fixed-effects difference-in-differences regression results for the association between digitalization-involved commercial reform (*Treat×Post*) and stock price crash risk (*NCSKEW* and *DUVOL*). Columns (1) and (2) report the results of the univariate regression that includes *Treat×Post* and excludes the control variables. Columns (3) and (4) report the results of the multivariate regression that includes *Treat×Post* and the control variables. The sample period ranges from 2011 to 2019. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2. Year dummies, firm dummies, and city dummies are included in each regression, but their results are not reported for brevity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel B: Clustering the standard errors of coefficients by city in the DID regression

	(1) Dependent variable	(2) Dependent variable	(3) Dependent variable	(4) Dependent variable
Variables	$=NCSKEW_t$	$= DUVOL_t$	$= NCSKEW_t$	$= DUVOL_t$
Treat×Post	-0.174***	-0.147***	-0.161***	-0.159***
	(-24.511)	(-33.297)	(-16.597)	(-37.783)
$Treat_t$			0.639***	0.692***
			(10.824)	(16.386)
$size_t$			0.046***	0.043***
			(4.771)	(6.716)
soe_t			0.087***	0.055***
			(4.917)	(4.378)
roe_t			0.051	0.027
			(1.267)	(0.993)
lev_t			-0.145**	-0.084*
			(-2.019)	(-1.828)
$salesgrowth_t$			-0.019**	-0.013***
			(-2.444)	(-2.918)
$cashholdings_t$			-0.123	-0.097
			(-1.176)	(-1.327)
$duality_t$			-0.019	-0.013
			(-1.080)	(-1.002)
$boardsize_t$			0.058*	0.031
			(1.892)	(1.537)
$top\_shareholdings_t$			0.001***	0.001**
			(3.224)	(2.094)
hhi_t			-0.196*	-0.137**
			(-1.899)	(-1.989)
$ceoshare_t$			-0.122***	-0.078***
			(-4.307)	(-4.097)
ret_t			10.735***	8.974***
			(9.183)	(10.077)
$sigma_t$			5.553***	3.106***
			(11.466)	(9.159)
$share\_turnover_t$			-0.004	0.005
			(-0.031)	(0.062)
$roa\_volatility_t$			-0.051	0.026
			(-0.520)	(0.373)
Constant	0.230***	0.148***	-1.304***	-1.097***
	(88.381)	(54.314)	(-5.883)	(-7.039)
Observations	7,072	7,072	7,072	7,072
Adj. R^2	0.067	0.256	0.119	0.120
Year-fixed effects	included	included	included	included
Industry-fixed effects	included	included	included	included
City-fixed effects	included	included	included	included

Notes: Table B of Table 4 reports the OLS regression results for the association between digitalization-involved commercial reform (*Treat×Post*) and stock price crash risk (*NCSKEW* and *DUVOL*). Columns (1) and (2) report the results of the univariate regression that includes *Treat×Post* and excludes the control variables. Columns (3) and (4) report the results of the multivariate regression that includes *Treat×Post* and the control variables. The sample period ranges from 2011 to 2019. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2. Year dummies, industry dummies, and city dummies are included in each regression, but their results are not reported for brevity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by city. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel C: Test of coefficient stability

	Dependent variable =*NCSKEW*$_t$				Dependent variable = *DUVOL*$_t$		
Variables	(1)	(2)	(3)	Variables	(4)	(5)	(6)
	DID Coefficients	t-stat.	selection ratio		DID coefficients	t-stat.	selection ratio
Main control variables	-0.167***	-5.960		*main control variables*	-0.160***	-8.296	
+*roa_volatility*$_t$	-0.168***	-5.971	0.0085	+*size*$_t$	-0.172***	-8.734	0.0019
+*size*$_t$	-0.159***	-6.061	0.0021	+*duality*$_t$	-0.172***	-8.731	0.0019
+*salesgrowth*$_t$	-0.151***	-5.973	0.0017	+*boardsize*$_t$	-0.168***	-8.768	0.0052
+*duality*$_t$	-0.161***	-5.924	0.0043	+*salesgrowth*$_t$	-0.162***	-8.811	0.0033
+*share_turnover*$_t$	-0.159***	-10.058	0.0009	+*lev*$_t$	-0.173***	-8.753	0.0017
+*top_shareholdings*$_t$	-0.161***	-10.102	0.0006	+*top_shareholdings*$_t$	-0.159***	-8.551	0.0010

Notes: Table C of Table 4 reports the results from testing the stability of the coefficient of *Treat×Post* in the difference-in-differences regression of stock price crash risk (*NCSKEW* and *DUVOL*). Column (1) and (2) (Column (4) and (5)) present the coefficients and t value of *Treat×Post* in the regression of *NCSKEW* (*DUVOL*), after controlling for the 9 main control variables (i.e., the control variables ranked in the top 60% based on the economic magnitude of their coefficients in the baseline regression analysis), and progressively introduce each of the other 6 control variables into the regression (i.e., the control variables ranked in the bottom 40% based on the economic magnitude of their coefficients in the baseline regression). For instance, in the first (second) row of Column (1), the coefficient of *Treat×Post* in the regression of *NCSKEW* is -0.167 (-0.168) when *roa_volatility* (both *roa_volatility and size*) is (are) included in the regression along with the 9 main control variables (i.e., *ret*, *sigma*, *hhi*, *lev*, *cashholdings*, *ceoshare*, *soe*, *boardsize*, and *roe*). In the first (second) row of Column (4), the coefficient of *Treat×Post* in the regression of *DUVOL* is -0.160 (-0.172) when si*ze* (both *size and duality*) is (are) included in the regression along with the 9 main control variables (i.e., *sigma*, *cashholdings*, *ret*, *roe*, *roa_volatility*, *share_turnover*, *ceoshare*, *soe*, and *hhi*). In Column (3) (Column (6)), the selection ratio (i.e., the absolute value of the ratio of the standardized selection on "unobservables" to the standardized selection on "observables") is calculated as the absolute value of the difference in the estimated coefficients *Treat×Post* between the *NCSKEW* (*DUVOL*) regression with the 9 main control variables and the *NCSKEW* (*DUVOL*) regression with all controls that include the progressively added control variable(s), divided by the coefficient *Treat×Post* estimated from the regression with the 9 main control variables. For instance, in the first (second) row of Column (3), the selection ratio for *Treat×Post* in the regression of *NCSKEW* is 0.0085 (0.0021) when *roa_volatility* (both *roa_volatility and size*) is (are) included along with the 9 main control variables in the regression. In the first (second) row of Column (6), the selection ratio for *Treat×Post* in the regression of *DUVOL* is 0.0019 (0.0019) when si*ze* (both *size and duality*) is (are) included along with the 9 main control variables in the regression.

Panel D: Alternative measures of stock price crash risk

Variables	(1) Dependent variable = $CRASH1_t$	(2) Dependent variable = $CRASH1_t$	(3) Dependent variable = $CRASH2_t$	(4) Dependent variable = $CRASH2_t$
Treat×*Pre*3	-0.006		0.000	
	(-0.169)		(0.215)	
Treat×*Pre*2	0.001		-0.001	
	(0.034)		(-0.639)	
Treat×*Pre*1	0.011		-0.001	
	(0.317)		(-0.835)	
Treat×*Post*1	-0.068***		-0.005**	
	(-2.629)		(-2.289)	
Treat×*Post*2	-0.079**		-0.008***	
	(-2.379)		(-2.634)	
Treat×*Post*3	-0.094***		-0.011***	
	(-2.781)		(-3.079)	
Treat×*Post*		-0.160***		-0.104**
		(-24.904)		(-2.278)
Treat		0.555***		0.005**
		(27.580)		(2.367)
$size_t$	0.023***	0.028***	0.006***	0.006***
	(3.890)	(4.824)	(5.990)	(5.764)
soe_t	0.050***	0.053***	0.009***	0.008***
	(4.086)	(4.467)	(4.164)	(4.032)
roe_t	0.064**	0.069**	0.996***	0.995***
	(1.977)	(2.174)	(225.676)	(216.508)
lev_t	-0.132***	-0.155***	0.017*	0.018*
	(-2.598)	(-3.040)	(1.877)	(1.959)
$salesgrowth_t$	-0.012***	-0.011**	0.000	0.001
	(-2.750)	(-2.539)	(0.636)	(0.906)
$cashholdings_t$	-0.152*	-0.146*	-0.001	-0.004
	(-1.927)	(-1.899)	(-0.093)	(-0.422)
$duality_t$	-0.020*	-0.022**	-0.002*	-0.003**
	(-1.750)	(-1.967)	(-1.909)	(-2.100)
$boardsize_t$	0.031	0.038	-0.006	-0.005
	(1.199)	(1.481)	(-1.581)	(-1.322)
$top\_shareholdings_t$	0.001***	0.001***	-0.000*	-0.000**
	(2.788)	(2.999)	(-1.890)	(-2.050)
hhi_t	-0.254***	-0.241***	-0.021**	-0.020**
	(-2.737)	(-2.618)	(-2.405)	(-2.106)
$ceoshare_t$	-0.068***	-0.077***	-0.007***	-0.008***
	(-3.605)	(-4.057)	(-3.100)	(-3.084)
ret_t	5.358***	5.327***	-0.056	-0.082
	(7.487)	(7.463)	(-1.029)	(-1.443)
$sigma_t$	3.019***	3.113***	0.067**	0.094***
	(10.155)	(10.606)	(2.244)	(2.978)
$share\_turnover_t$	0.068	0.067	0.040***	0.037***
	(0.607)	(0.604)	(4.801)	(4.374)
$roa\_volatility_t$	-0.003	-0.019	0.013*	0.016**
	(-0.045)	(-0.267)	(1.752)	(1.995)
Constant	-0.229	-0.442***	-1.084***	-1.095***
	(-1.625)	(-3.357)	(-50.929)	(-49.335)
Observations	7,072	7,072	7,072	7,072
Adj. R^2	0.062	0.054	0.168	0.167
Year-fixed effects	included	included	included	included
Industry-fixed effects	included	included	included	included
City-fixed effects	included	included	included	included

Notes: Table Dof Table 4 reports the results of the test that uses alternative measures of stock price crash risk (i.e., *CRASH*1 and *CRASH*2). Columns (1) and (3) report the results of the parallel trends assumption test using alternative measures as to *CRASH*1 and *CRASH*2, respectively. Columns (2) and (4) report the results of baseline regression using alternative measures as to *CRASH*1 and *CRASH*2, respectively. The sample period ranges from 2011 to 2019. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2. Year dummies, industry dummies, and city dummies are included in each regression, but their results are not reported for brevity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 5: Tests of the mechanisms through which the digitalization-involved commercial reform reduces stock price crash risk

Panel A: The information channel

Variables	(1) Dependent variable = *Media_coverage*$_t$	(2) Dependent variable = *NCSKEW*$_t$	(3) Dependent variable = *DUVOL*$_t$
Treat×Post	0.263***		
	(9.248)		
Treat	-0.252***		
	(-9.073)		
Media_coverage		-0.188***	-0.142***
		(-5.808)	(-6.442)
size$_t$	0.033***	0.052***	0.048***
	(7.461)	(6.249)	(8.284)
soe$_t$	-0.090***	0.070***	0.042***
	(-8.835)	(4.340)	(3.717)
roe$_t$	-0.206***	0.012	-0.002
	(-8.882)	(0.273)	(-0.079)
lev$_t$	0.105***	-0.125*	-0.069
	(3.988)	(-1.732)	(-1.400)
salesgrowth$_t$	0.005**	-0.018**	-0.012***
	(2.106)	(-2.551)	(-2.786)
cashholdings$_t$	0.030	-0.117	-0.092
	(0.726)	(-1.161)	(-1.286)
duality$_t$	0.002	-0.019	-0.013
	(0.297)	(-1.248)	(-1.228)
boardsize$_t$	0.041**	0.065*	0.037
	(2.211)	(1.927)	(1.571)
top_shareholdings$_t$	0.011***	0.004***	0.002***
	(40.158)	(6.028)	(5.714)
hhi$_t$	0.094**	-0.178	-0.124
	(2.040)	(-1.498)	(-1.499)
ceoshare$_t$	0.065***	-0.110***	-0.069***
	(5.376)	(-4.216)	(-3.765)
ret$_t$	1.877***	11.087***	9.241***
	(7.785)	(10.786)	(12.034)
sigma$_t$	0.825***	5.708***	3.223***
	(5.843)	(11.714)	(9.390)
share_turnover$_t$	-0.960***	-0.184	-0.131
	(-21.677)	(-1.251)	(-1.222)
roa_volatility$_t$	-0.072*	-0.064	0.015
	(-1.943)	(-0.603)	(0.224)
Constant	2.844***	-0.770***	-0.691***
	(29.404)	(-3.629)	(-4.692)
Observations	7,072	7,072	7,072
Adj. R^2	0.473	0.101	0.103
Year-fixed effects	included	included	included
Industry-fixed effects	included	included	included
City-fixed effects	included	included	included

Notes: Panel A of Table 5 reports the results as to the test of the information channel through which the digitalization-involved commercial reform reduces stock price crash risk. Column (1) reports the results of the regression of media news (*Media_coverage*) on digitalization-involved commercial reform (*Treat×Post*). Columns (2) and (3) report the results of the baseline regression that is augmented by *Media_coverage* but excludes *Treat×Post* and *Treat.* The sample period ranges from 2011 to 2019. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2. Year dummies, industry dummies, and city dummies are included in each regression, but their results are not reported for brevity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel B: The monitoring channel

Variables	(1) Dependent variable = *Related_transaction*$_t$	(2) Dependent variable = *NCSKEW*$_t$	(3) Dependent variable = *DUVOL*$_t$	(4) Dependent variable = *Ab_accrual*$_t$	(5) Dependent variable = *NCSKEW*$_t$	(6) Dependent variable = *DUVOL*$_t$	(7) Dependent variable = *Other_receivable*$_t$	(8) Dependent variable = *NCSKEW*$_t$	(9) Dependent variable = *DUVOL*$_t$
Treat×Post	-0.153***			-0.003***			-0.0760***		
	(-5.345)			(-4.563)			(-5.543)		
Treat	0.123***			0.002***			-0.0830***		
	(5.243)			(3.897)			(-4.591)		
Related_transaction		0.018**	0.012**						
		(2.182)	(2.034)						
Ab_accrual					12.294***	11.011***			
					(5.230)	(6.925)			
Other_receivable								0.244***	0.197***
								(5.195)	(5.989)
size$_t$	0.957***	0.028**	0.032***	-0.006***	0.123***	0.112***	0.010***	0.043***	0.041***
	(56.503)	(2.468)	(3.912)	(-155.824)	(7.707)	(10.068)	(6.681)	(5.242)	(7.171)
soe$_t$	0.024	0.087***	0.054***	0.000	0.086***	0.054***	0.028***	0.080***	0.049***
	(0.637)	(5.352)	(4.809)	(1.030)	(5.276)	(4.743)	(9.221)	(4.974)	(4.352)
roe$_t$	0.372***	0.044	0.022	0.028***	-0.292***	-0.280***	-0.012	0.054	0.029
	(4.039)	(0.971)	(0.739)	(134.692)	(-3.770)	(-5.252)	(-1.433)	(1.184)	(0.964)
lev$_t$	-1.170***	-0.126*	-0.071	-0.001***	-0.137*	-0.077	-0.031**	-0.137*	-0.078
	(-9.203)	(-1.734)	(-1.429)	(-2.797)	(-1.878)	(-1.540)	(-2.345)	(-1.914)	(-1.583)
salesgrowth$_t$	-0.079***	-0.017**	-0.012***	0.000	-0.019***	-0.013***	-0.003**	-0.018***	-0.013***
	(-5.921)	(-2.464)	(-2.702)	(0.436)	(-2.679)	(-2.954)	(-2.236)	(-2.578)	(-2.812)
cashholdings$_t$	1.957***	-0.158	-0.120*	0.002***	-0.146	-0.115	-0.049**	-0.111	-0.087
	(10.121)	(-1.547)	(-1.652)	(3.474)	(-1.429)	(-1.587)	(-2.461)	(-1.097)	(-1.208)
duality$_t$	-0.029	-0.018	-0.013	0.000	-0.020	-0.014	-0.002	-0.019	-0.013
	(-1.024)	(-1.223)	(-1.212)	(1.076)	(-1.316)	(-1.308)	(-0.769)	(-1.248)	(-1.224)
boardsize$_t$	0.295***	0.053	0.028	-0.000	0.058*	0.031	0.004	0.057*	0.031
	(3.741)	(1.547)	(1.172)	(-0.673)	(1.709)	(1.298)	(0.663)	(1.677)	(1.287)
top_shareholdings$_t$	0.003***	0.001***	0.001**	-0.000	0.001***	0.001**	-0.000***	0.002***	0.001**
	(2.746)	(2.936)	(2.016)	(-1.496)	(3.095)	(2.169)	(-3.913)	(3.223)	(2.329)
hhi$_t$	-1.685***	-0.162	-0.114	-0.000	-0.187	-0.129	-0.136***	-0.160	-0.107
	(-8.291)	(-1.349)	(-1.359)	(-0.748)	(-1.566)	(-1.563)	(-5.653)	(-1.347)	(-1.301)
ceoshare$_t$	-0.217***	-0.118***	-0.076***	0.000	-0.124***	-0.079***	-0.017***	-0.118***	-0.075***
	(-4.525)	(-4.560)	(-4.142)	(0.479)	(-4.775)	(-4.368)	(-3.184)	(-4.552)	(-4.103)
ret$_t$	3.148***	10.686***	8.943***	-0.026***	11.019***	9.222***	-4.964***	11.950***	9.959***
	(2.842)	(10.406)	(11.650)	(-6.313)	(10.715)	(12.013)	(-26.487)	(11.157)	(12.541)
sigma$_t$	-3.321***	5.610***	3.143***	-0.006***	5.620***	3.171***	-1.633***	5.951***	3.427***
	(-5.420)	(11.468)	(9.108)	(-3.404)	(11.534)	(9.221)	(-20.181)	(12.165)	(9.978)
share_turnover$_t$	0.592***	-0.015	-0.002	-0.004***	0.040	0.046	-0.124***	0.026	0.030
	(3.432)	(-0.102)	(-0.015)	(-5.754)	(0.274)	(0.435)	(-4.265)	(0.182)	(0.284)
roa_volatility$_t$	0.050	-0.050	0.026	0.001**	-0.068	0.011	-0.022	-0.045	0.030
	(0.368)	(-0.470)	(0.375)	(2.475)	(-0.649)	(0.161)	(-1.219)	(-0.424)	(0.438)
Constant	-1.426***	-1.278***	-1.080***	0.131***	-2.913***	-2.533***	0.258***	-1.368***	-1.148***
	(-3.882)	(-6.723)	(-8.059)	(147.677)	(-8.537)	(-10.543)	(7.646)	(-7.244)	(-8.661)
Observations	7,072	7,072	7,072	7,072	7,072	7,072	7,072	7,072	7,072
Adj. R^2	0.798	0.098	0.100	0.859	0.100	0.103	0.435	0.100	0.103
Year-fixed effects	included	included	included	included	included	included	included	included	included
Industry-fixed effects	included	included	included	included	included	included	included	included	included
City-fixed effects	included	included	included	included	included	included	included	included	included

Notes: Panel B of Table 5 reports the results of the test of the monitoring channel through which the digitalization-involved commercial reform reduces stock price crash risk. Column (1) reports the results of the regression of related party transactions (*Related_transaction*) on digitalization-involved commercial reform (*Treat×Post*). Columns (2) and (3) report the results of the baseline regression that is augmented by *Related_transaction* but excludes *Treat×Post* and *Treat*. Column (4) reports the results of the regression of abnormal accruals (*Ab_accrual*) on digitalization-involved commercial reform (*Treat×Post*). Columns (5) and (6) report the results of the baseline regression that is augmented by *Ab_accrual* but excludes *Treat×Post* and *Treat*. Column (7) reports the results of the regression of other accounts receivable (*Other_receivable*) on digitalization-involved commercial reform (*Treat×Post*). Columns (8) and (9) report the results of the baseline regression that is augmented by *Other_receivable* but excludes *Treat×Post* and *Treat*. The sample period ranges from 2011 to 2019. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2. Year dummies, industry dummies, and city dummies are included in each regression, but their results are not reported for brevity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 6: The moderation analysis of the association between digitalization-applied commercial reform and stock price crash risk

Panel A: The moderating effect of corporate digitalization

	(1) Dependent variable	(2) Dependent variable	(3) Dependent variable	(4) Dependent variable
Variables	$= NCSKEW_t$	$= DUVOL_t$	$= NCSKEW_t$	$= DUVOL_t$
Treat×Post×Dum_Digit	-0.088**	-0.052**		
	(-2.449)	(-2.131)		
*Treat×Post×Dum_Digit*1			-0.071**	-0.053**
			(-2.136)	(-2.049)
Dum_Digit	-0.026***	-0.022**		
	(-2.764)	(-1.996)		
*Dum_Digit*1			-0.039***	-0.036***
			(-2.582)	(-3.417)
Treat×Post	-0.071***	-0.046***	-0.203***	-0.111***
	(-11.363)	(-18.575)	(-10.654)	(-17.060)
Treat	0.721***	0.739***	0.709***	0.728***
	(12.355)	(20.031)	(12.136)	(19.564)
$size_t$	0.053***	0.048***	0.051***	0.046***
	(6.016)	(7.716)	(5.825)	(7.519)
soe_t	0.084***	0.052***	0.084***	0.052***
	(5.193)	(4.594)	(5.246)	(4.650)
roe_t	0.074	0.041	0.079	0.044
	(1.352)	(1.121)	(1.428)	(1.213)
lev_t	-0.542***	-0.275***	-0.552***	-0.281***
	(-4.117)	(-2.910)	(-4.192)	(-2.981)
$salesgrowth_t$	-0.025***	-0.017***	-0.025***	-0.017***
	(-2.989)	(-3.214)	(-3.015)	(-3.255)
$cashholdings_t$	0.243*	0.082	0.265*	0.099
	(1.692)	(0.818)	(1.849)	(0.987)
$duality_t$	-0.018	-0.013	-0.018	-0.012
	(-1.193)	(-1.196)	(-1.190)	(-1.178)
$boardsize_t$	0.091*	0.061	0.091*	0.061
	(1.685)	(1.620)	(1.679)	(1.634)
$top\_shareholdings_t$	0.002***	0.001**	0.002***	0.001***
	(3.252)	(2.568)	(3.349)	(2.712)
hhi_t	-0.023	-0.036	-0.042	-0.052
	(-0.186)	(-0.414)	(-0.330)	(-0.588)
$ceoshare_t$	-0.122***	-0.078***	-0.121***	-0.078***
	(-4.398)	(-4.030)	(-4.392)	(-4.019)
ret_t	4.871***	5.474***	4.815***	5.435***
	(4.193)	(6.534)	(4.132)	(6.465)
$sigma_t$	6.871***	3.699***	6.835***	3.675***
	(11.961)	(9.425)	(11.864)	(9.328)
$share\_turnover_t$	1.147***	0.957***	1.163***	0.971***
	(3.434)	(3.959)	(3.479)	(4.016)
$roa\_volatility_t$	-0.117	-0.023	-0.110	-0.018
	(-1.049)	(-0.305)	(-0.986)	(-0.235)
Constant	-1.655***	-1.315***	-1.589***	-1.263***
	(-7.695)	(-8.656)	(-7.446)	(-8.381)
Observations	7,072	7,072	7,072	7,072
Adj. R^2	0.100	0.103	0.101	0.100
Year-fixed effects	included	included	included	included
Industry-fixed effects	included	included	included	included
City-fixed effects	included	included	included	included

Notes: Panel A of Table 6 reports the results for the moderating effect of corporate digitalization (*Digit* and *Digit*1) on the association between digitalization-involved commercial reform and stock price crash risk (*NCSKEW* and *DUVOL*). The moderating effect is captured by the interaction term between the indicator for corporate digitalization (i.e., *Dum_Digit* and *Dum_Digit*1) and *Treat×Post*. *Dum_Digit* (*Dum_Digit*1) equals 1 if the value of *Digit* (*Digit*1) is higher than its full-sample median, and 0 otherwise. Columns (1) and (2) report the moderating effect of *Dum_Digit*. Columns (3) and (4) report the moderating effect of *Dum_Digit*1. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2. Year dummies, industry dummies, and city dummies are included in each regression, but their results are not reported for brevity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel B: The moderating effect of corporate innovation

Variables	(1) Dependent variable = $NCSKEW_t$	(2) Dependent variable = $DUVOL_t$	(3) Dependent variable = $NCSKEW_t$	(4) Dependent variable = $DUVOL_t$
Treat×Post×Dum_Innovation	-0.089**	-0.079***		
	(-2.367)	(-3.018)		
*Treat×Post×Dum_Innovation*1			-0.148**	-0.097**
			(-2.260)	(-1.999)
Dum_Innovation	-0.025	-0.015		
	(-1.453)	(-1.270)		
*Dum_Innovation*1			-0.017	0.008
			(-0.658)	(0.434)
Treat×Post	-0.078***	-0.040***	-0.091***	-0.105***
	(-9.592)	(-16.803)	(-10.868)	(-18.670)
Treat	0.629***	0.686***	0.638***	0.693***
	(10.658)	(18.592)	(10.905)	(18.958)
$size_t$	0.044***	0.042***	0.045***	0.042***
	(5.370)	(7.335)	(5.398)	(7.178)
soe_t	0.084***	0.053***	0.088***	0.055***
	(5.223)	(4.671)	(5.425)	(4.857)
roe_t	0.039	0.019	0.049	0.027
	(0.839)	(0.612)	(1.084)	(0.894)
lev_t	-0.145**	-0.084*	-0.148**	-0.087*
	(-1.999)	(-1.685)	(-2.047)	(-1.758)
$salesgrowth_t$	-0.018***	-0.013***	-0.019***	-0.013***
	(-2.603)	(-2.851)	(-2.640)	(-2.900)
$cashholdings_t$	-0.127	-0.099	-0.121	-0.097
	(-1.248)	(-1.374)	(-1.190)	(-1.334)
$duality_t$	-0.017	-0.012	-0.019	-0.013
	(-1.134)	(-1.096)	(-1.292)	(-1.276)
$boardsize_t$	0.059*	0.032	0.059*	0.032
	(1.728)	(1.334)	(1.720)	(1.316)
$top\_shareholdings_t$	0.001***	0.001**	0.001***	0.001**
	(3.021)	(2.098)	(3.047)	(2.127)
hhi_t	-0.203*	-0.142*	-0.199*	-0.138*
	(-1.703)	(-1.713)	(-1.672)	(-1.672)
$ceoshare_t$	-0.117***	-0.074***	-0.123***	-0.078***
	(-4.511)	(-4.077)	(-4.747)	(-4.299)
ret_t	10.706***	8.955***	10.760***	8.981***
	(10.416)	(11.679)	(10.472)	(11.714)
$sigma_t$	5.570***	3.118***	5.554***	3.116***
	(11.389)	(9.044)	(11.389)	(9.070)
$share\_turnover_t$	-0.012	-0.000	-0.009	0.001
	(-0.084)	(-0.001)	(-0.062)	(0.005)
$roa\_volatility_t$	-0.048	0.028	-0.052	0.025
	(-0.449)	(0.407)	(-0.482)	(0.361)
Constant	-1.266***	-1.071***	-1.303***	-1.077***
	(-6.643)	(-7.988)	(-6.747)	(-7.923)
Observations	7,072	7,072	7,072	7,072
Adj. R^2	0.099	0.101	0.098	0.100
Year-fixed effects	included	included	included	included
Industry-fixed effects	included	included	included	included
City-fixed effects	included	included	included	included

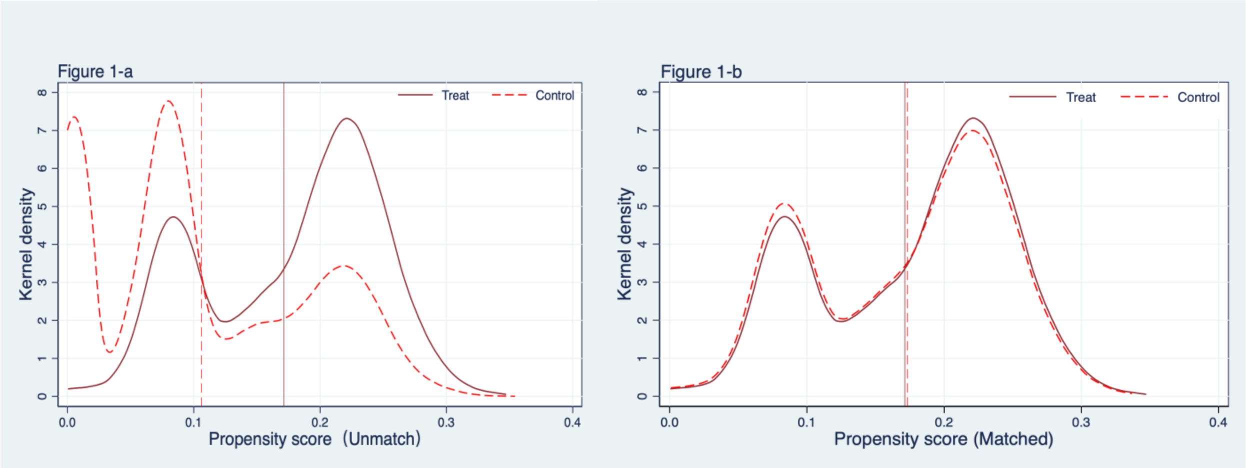
Notes: Panel B of Table 6 reports the results for the moderating effect of corporate innovation (*Innovation* and *Innovation*1) on the association between digitalization-involved commercial reform and stock price crash risk (*NCSKEW* and *DUVOL*). The moderating effect is captured by the interaction term between the indicator for corporate innovation (i.e., *Dum_Innovation* and *Dum_Innovation*1) and *Treat×Post*. *Dum_Innovation* (*Dum_Innovation*1) equals 1 if the value of *Innovation* (*Innovation*1) is higher than its full-sample median, and 0 otherwise. Columns (1) and (2) report the moderating effect of *Dum_Innovation*. Columns (3) and (4) report the moderating effect of *Dum_Innovation*1. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2. Year dummies, industry dummies, and city dummies are included in each regression, but their results are not reported for the sake of brevity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel C: The moderating effect of corporate governance

Variables	(1) Dependent variable = *NCSKEW*$_t$	(2) Dependent variable = *DUVOL*$_t$	(3) Dependent variable = *NCSKEW*$_t$	(4) Dependent variable = *DUVOL*$_t$
Treat×Post×Dum_CG	0.082**	0.053**		
	(2.205)	(2.031)		
*Treat×Post×Dum_CG*1			0.123***	0.063**
			(3.293)	(2.393)
Dum_CG	-0.006	-0.004		
	(-0.339)	(-0.335)		
*Dum_CG*1			0.019	0.012
			(1.228)	(1.052)
Treat×Post	-0.183***	-0.121***	-0.181***	-0.114***
	(-11.066)	(-18.520)	(-10.754)	(-17.942)
Treat	0.634***	0.689***	0.630***	0.686***
	(10.747)	(18.735)	(10.704)	(18.719)
size$_t$	0.045***	0.043***	0.046***	0.043***
	(5.057)	(6.958)	(5.585)	(7.526)
soe$_t$	0.087***	0.055***	0.086***	0.054***
	(5.387)	(4.846)	(5.352)	(4.814)
roe$_t$	0.048	0.025	0.045	0.023
	(1.043)	(0.818)	(0.983)	(0.765)
lev$_t$	-0.145**	-0.084*	-0.132*	-0.077
	(-2.000)	(-1.695)	(-1.829)	(-1.557)
salesgrowth$_t$	-0.019***	-0.013***	-0.019***	-0.013***
	(-2.687)	(-2.936)	(-2.682)	(-2.931)
cashholdings$_t$	-0.124	-0.097	-0.148	-0.111
	(-1.215)	(-1.343)	(-1.449)	(-1.528)
duality$_t$	-0.019	-0.013	-0.018	-0.013
	(-1.241)	(-1.222)	(-1.218)	(-1.210)
boardsize$_t$	0.058*	0.032	0.057*	0.031
	(1.704)	(1.321)	(1.687)	(1.300)
top_shareholdings$_t$	0.001***	0.001**	0.001***	0.001**
	(3.048)	(2.125)	(3.012)	(2.093)
hhi$_t$	-0.198*	-0.138*	-0.181	-0.129
	(-1.664)	(-1.674)	(-1.519)	(-1.556)
ceoshare$_t$	-0.122***	-0.078***	-0.122***	-0.078***
	(-4.695)	(-4.263)	(-4.730)	(-4.285)
ret$_t$	10.774***	8.999***	10.701***	8.953***
	(10.487)	(11.734)	(10.429)	(11.672)
sigma$_t$	5.546***	3.101***	5.574***	3.117***
	(11.376)	(9.018)	(11.445)	(9.075)
share_turnover$_t$	-0.008	0.003	-0.003	0.006
	(-0.057)	(0.026)	(-0.024)	(0.052)
roa_volatility$_t$	-0.051	0.025	-0.054	0.023
	(-0.477)	(0.367)	(-0.509)	(0.341)
Constant	-1.285***	-1.085***	-1.322***	-1.107***
	(-6.380)	(-7.697)	(-6.946)	(-8.255)
Observations	7,072	7,072	7,072	7,072
Adj. R^2	0.098	0.100	0.099	0.100
Year-fixed effects	included	included	included	included
Industry-fixed effects	included	included	included	included
City-fixed effects	included	included	included	included

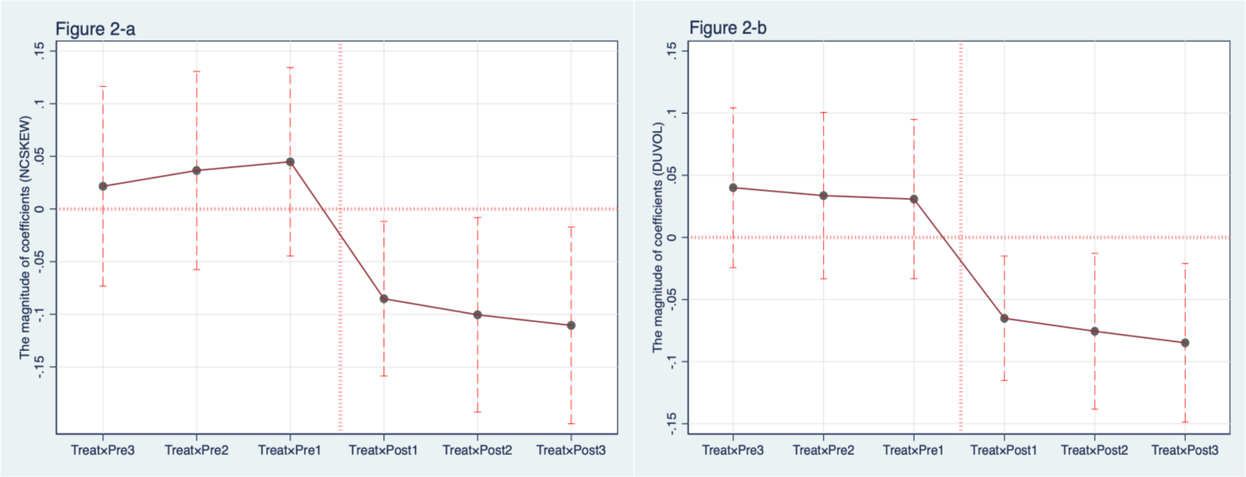
Notes: Panel C of Table 6 reports the results for the moderating effect of corporate governance (*CG* and *CG*1) on the association between digitalization-involved commercial reform and stock price crash risk (*NCSKEW* and *DUVOL*). The moderating effect is captured by the interaction term between the indicator for corporate governance (i.e., *Dum_CG* and *Dum_CG*1) and *Treat×Post*. *Dum_CG* (*Dum_CG*1) equals 1 if the value of *CG* (*CG*1) is higher than its full-sample median, and 0 otherwise. Columns (1) and (2) report the moderating effect of *Dum_CG*. Columns (3) and (4) report the moderating effect of *Dum_CG*1. The sample period ranges from 2011 to 2019. All the continuous variables are winsorized at the 1 and 99 percentage points, respectively, and are defined in Appendix 2. Year dummies, industry dummies, and city dummies are included in each regression, but their results are not reported for brevity. The t-statistics are based on robust standard errors adjusted for heteroskedasticity and clustered by firm. *, **, and *** indicate the two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Figure 1: Kernel density distribution of propensity matching



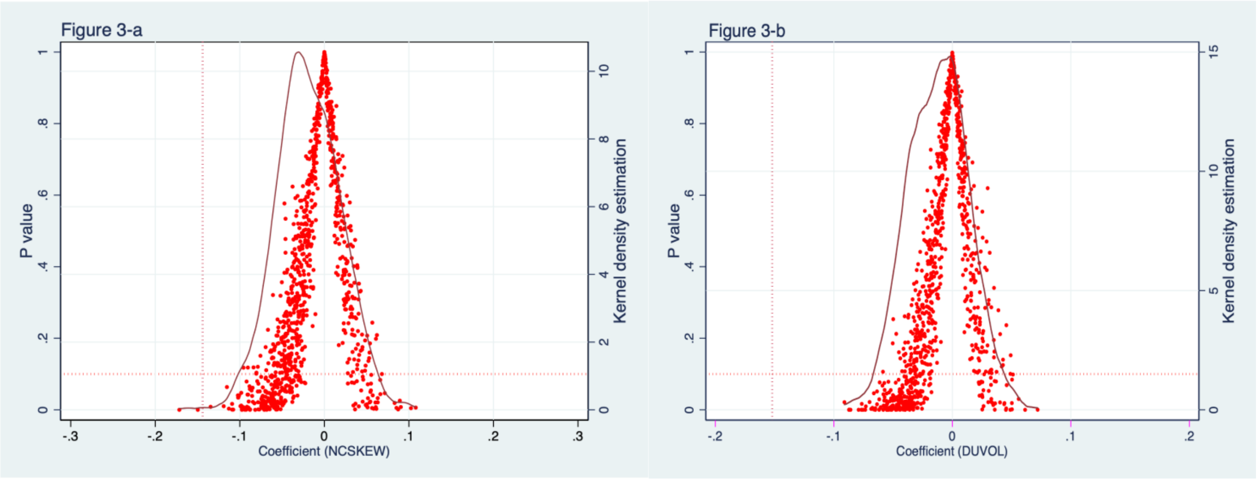
Notes: Figure 1 shows the distribution, in the form of kernel density curve, of propensity scores for the treatment group and control group before and after the matching. The horizontal axis represents the propensity scores; the vertical axis represents the probability density. The left (right) figure shows the distribution of propensity scores before (after) the matching. The sample period ranges from 2011 to 2019. The treatment indicator variable, *Treat*, equals 1 (0) for a treatment (control) firm. The treatment firm is defined as subject to the digitalization-involved commercial reform in which the Market Supervision Administration was established to introduce digital commercial registration system for improving information environments and monitoring on commercial activities of firms. The control firm is not subject to the digitalization-involved commercial reform in the six-year period centered at the beginning of the year of the reform for the treatment firm, nor before the period. The solid (dashed) curves represent the distribution of propensity scores for the treatment (control) firms. We follow Leuven and Sianesi (2018) to match each treatment firm, with replacement, with a control firm by using the closest propensity score within a caliper of 1% for each year.

Figure 2: Parallel trend test



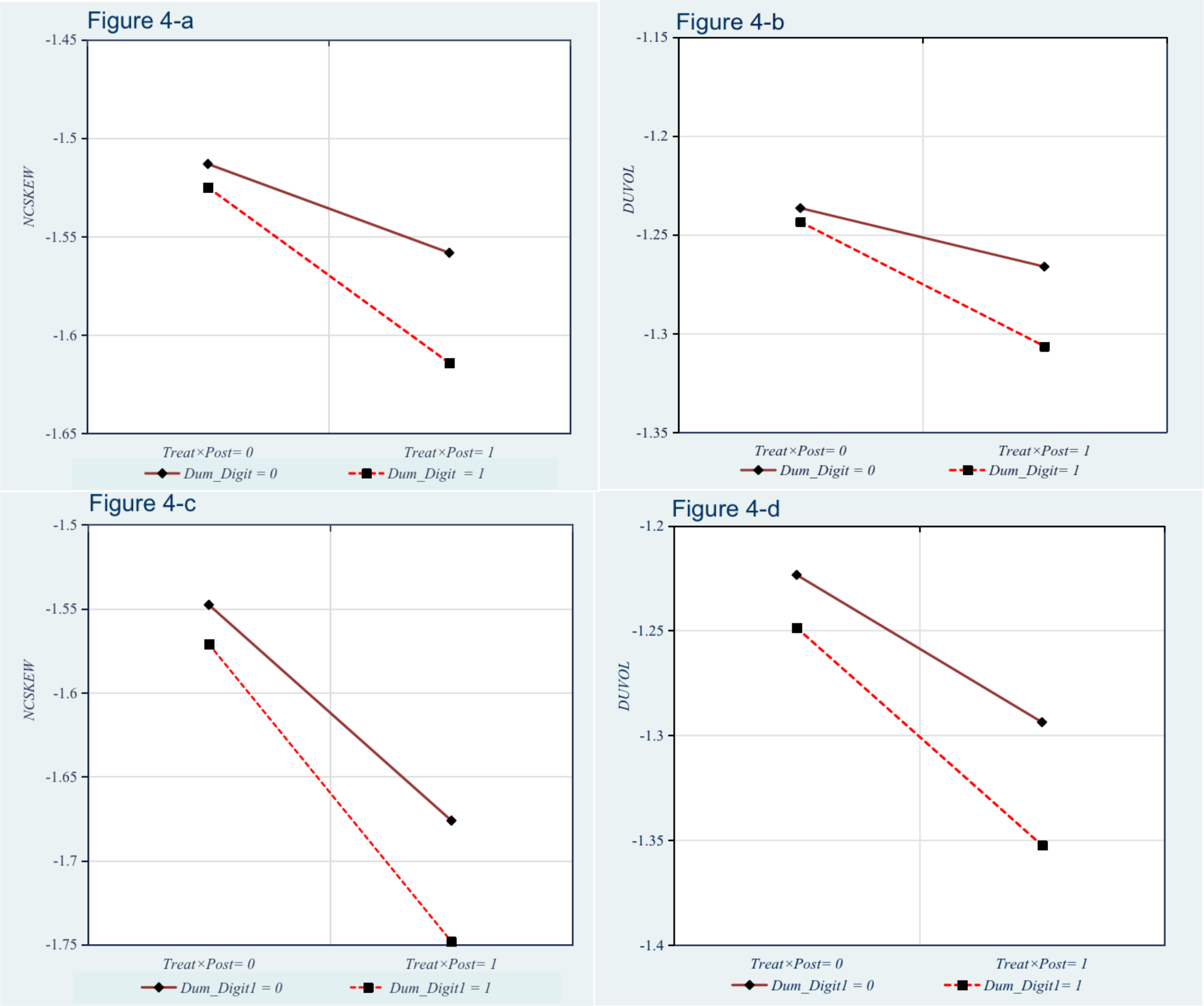
Notes: Figure 2 presents the results of the coefficient test of the parallel trends assumption for the difference-in-differences (DID) regression estimation. Specifically, Figure 2-a (2-b) shows the graphical diagnostic of parallel trend assumption for the DID regression where *NCSKEW* (*DUVOL*) is the dependent variable. The horizontal axis represents the interaction terms between *Treat* and *Pre** (*Post**); the vertical axis represents the magnitude of the coefficient of *Treat*×*Pre** (*Treat*×*Post**). The short dashed lines, which are perpendicular to the horizontal axis, are the corresponding 95% confidence interval for each coefficient. We consider a 6-year period and report the coefficients of *Treat*×*Pre** (*Treat*×*Post**), which are estimated from the regression model (6). *Pre** and *Post** include *Pre*3, *Pre*2, *Pre*1, *Post*1, *Post*2*,* and *Post*3, which are the year dummies for the 6-year period. The standard errors of the coefficients are adjusted for heteroskedasticity and clustered by firm. All the continuous variables are winsorized at 1 and 99 percentage points, respectively, and are defined in Appendix 2.

Figure 3: Placebo test



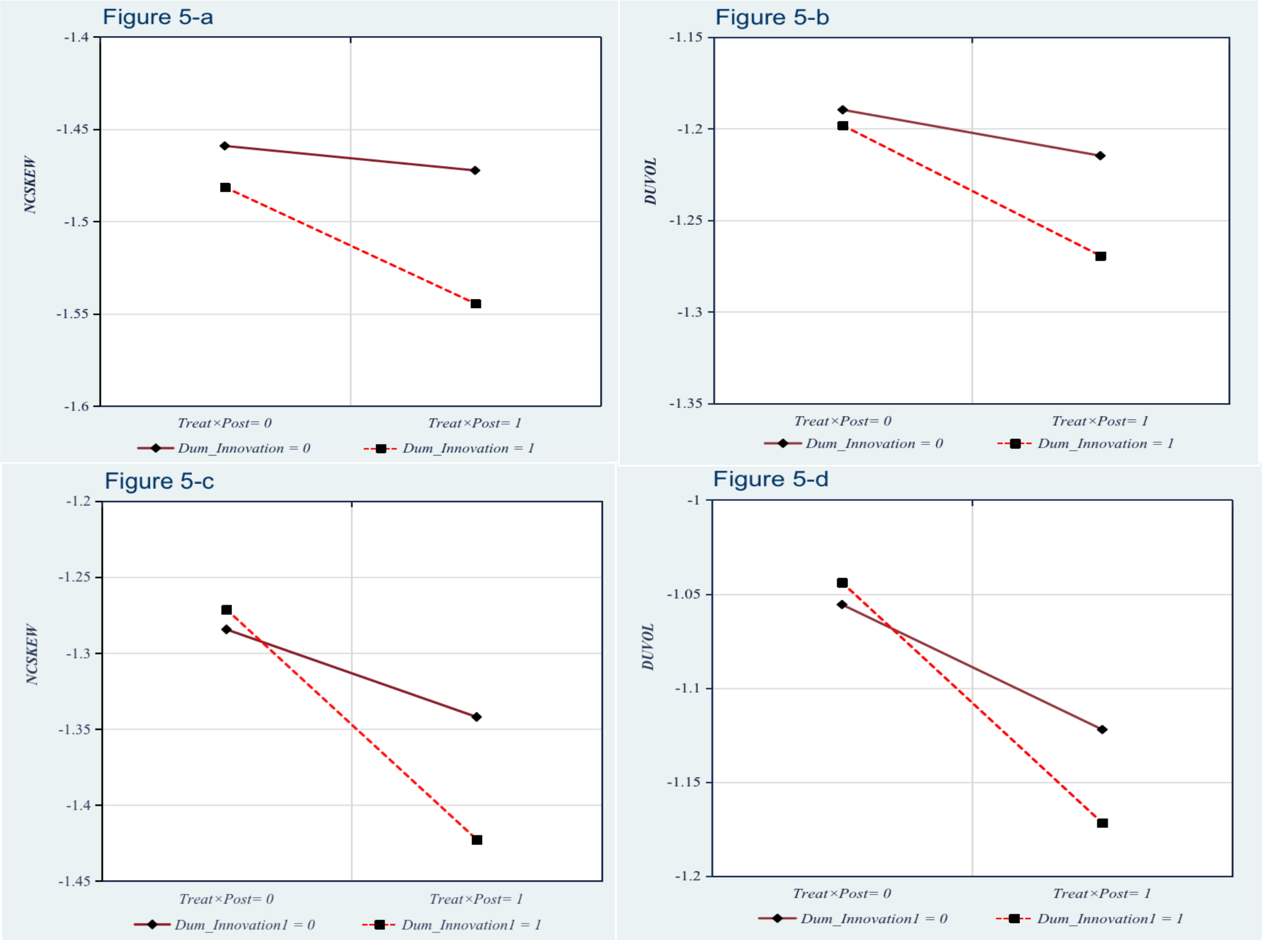
Notes: Figure 3 plots the cumulative distribution density of the 1,000 coefficient estimates in a placebo test. We randomly assign observations, which are not subject to the digitalization-involved commercial reform, to generate a fake treatment group *Treat^fake^* and associated fake reform time *Post^fake^* for each year and repeat this trial for 1,000 times to obtain 1,000 DID estimators for the interaction term $Treat^{fake} \times Post^{fake}$. The horizontal axis represents the magnitude of the estimated coefficients of the interaction term $Treat^{fake} \times Post^{fake}$; the vertical axis represents its corresponding p values on statistical significance and kernel density estimates, respectively. The dots (solid curve) represent (s) the distribution of kernel density (p values) of the estimated coefficients in the placebo test; the left (right) figure shows such results for *NCSKEW* (*DUVOL*). The dotted vertical line represents the estimated coefficient on *Treat*×*Post* for the baseline regression of *NCSKEW* (*DUVOL*), corresponding with the result of Column (1) (Column (2)), under Table B of Table 3.

Figure 4: The moderating effect of corporate digitalization



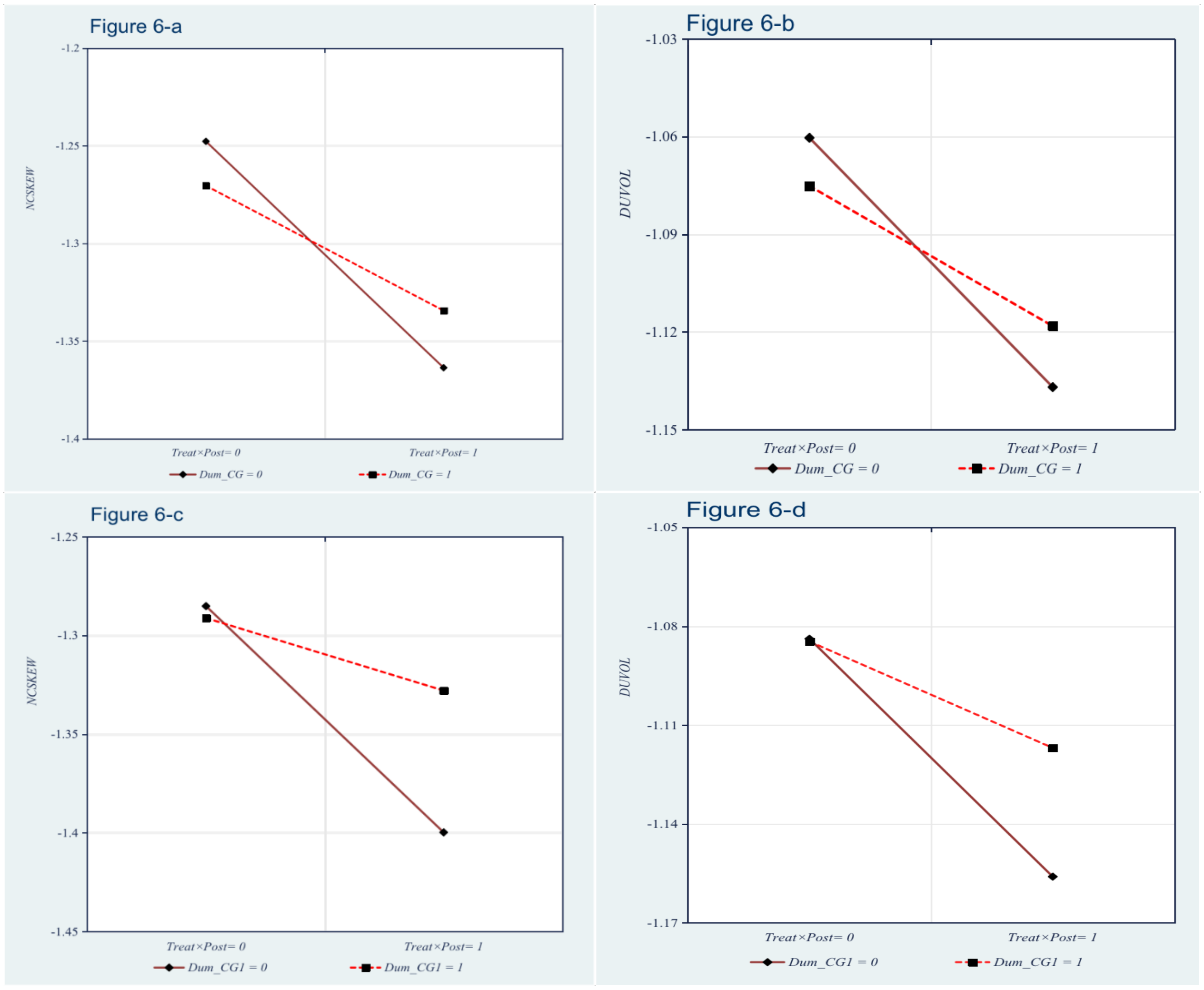
Notes: Figure 4 shows the diagram as to the linear interaction effect of corporate digitalization (*Digit* and *Digit*1) on the association between digitalization-involved commercial reform and stock price crash risk. The interaction effect is captured by the ternary interaction term between the indicator variable for corporate digitalization *Dum_Digit* (*Dum_Digit*1) and the DID interaction term *Treat*×*Post*. *Dum_Digit* (*Dum_Digit*1) equals 1 if the value of *Digit* (*Digit*1) is higher than its full-sample median, and 0 otherwise. The horizontal axis represents the value of the interaction term *Treat*×*Post*. The vertical axis represents the levels of stock price crash risk (i.e., *NCSKEW* and *DUVOL* for the left figure and right figure, respectively).

Figure 5: The moderation effect of corporate innovation



Notes: Figure 5 shows the diagram as to the linear interaction effect of corporate innovation (*Innovation* and *Innovation*1) on the association between digitalization-involved commercial reform and stock price crash risk. The interaction effect is captured by the interaction term between the indicator variable for corporate innovation *Dum_Innovation* (*Dum_Innovation*1) and the DID interaction term *Treat*×*Post*. *Dum_Innovation* (*Dum_Innovation*1) equals 1 if the value of *Innovation* (*Innovation*1) is higher than its full-sample median, and 0 otherwise. The horizontal axis represents the value of the interaction term *Treat*×*Post*. The vertical axis represents the levels of stock price crash risk (i.e., *NCSKEW* and *DUVOL* for the left figure and right figure, respectively).

Figure 6: The moderating effect of corporate governance



Notes: Figure 6 shows the diagram as to the linear interaction effect of corporate governance (*CG* and *CG*1) on the association between digitalization-involved commercial reform and stock price crash risk. The interaction effect is captured by the interaction term between the indicator variable for corporate governance *Dum_CG* (*Dum_CG*1) and the DID interaction term *Treat*×*Post*. *Dum_CG* (*Dum_CG*1) equals 1 if the value of *CG* (*CG*1) is higher than its full-sample median, and 0 otherwise. The horizontal axis represents the value of the interaction term *Treat*×*Post*, and the vertical axis represents the levels of the stock price crash risk (i.e., *NCSKEW* and *DUVOL* for the left figure and right figure, respectively).

Smart Homes: Enhancing Lives or Creating Challenges? Insights from People with Vulnerabilities

Dinara Davlembayeva
Newcastle University Business School
Joanne Swaffield
Newcastle University Business School
Savvas Papagiannidis
Newcastle University Business School
Davit Marikyan
Newcastle University Business School
Diana Gregory-Smith
Newcastle University Business School

Research in Progress

Abstract

Smart homes promise to enhance quality of life, yet evidence about their perceptions and the implications for vulnerable populations—who could potentially benefit most from this technology—remains inconclusive. Addressing this knowledge gap, this study adopts an exploratory approach to investigate smart home perceptions among individuals from diverse social backgrounds, focusing on the social characteristics that influence willingness to adopt. Through qualitative, field-based insights, we identified seven key factors—compatibility, cost, aesthetics, comfort, uncertain usability, complexity and energy savings—as central to smart home adoption, each influenced by users' personal circumstances and vulnerabilities. These findings were integrated with existing literature to conceptualise a research model addressing the task-technology fit of smart homes for users with varying vulnerabilities. This model lays the foundation for further quantitative testing, offering insights for creating more inclusive smart home environments and promoting broader adoption.

Keywords: smart homes, adoption, vulnerabilities, task-technology fit.

1.0 Introduction

A smart home is "*a residence equipped with a high-tech network, linking sensors and domestic devices, appliances, and features that can be remotely monitored, accessed or controlled, and provide services that respond to the needs of its inhabitants*" (Balta-Ozkan et al., 2013, p. 364). Smart home adoption occurs under the premise that this technology promises to enhance quality of life and simplify household tasks by automating and centralising control over home environments, including security, lighting, and energy-efficient devices (Marikyan et al., 2019). However, the intended benefits of smart homes often clash with unintended consequences or contradictions.

Despite their convenience, smart devices can introduce uncertainties and challenges, such as installation difficulties and disruptions to habitual daily routines (White & Miller, 2024), raising concerns about the technology's implications for vulnerability and accessibility, potentially counteracting the well-being of home inhabitants (Brown & Markusson, 2019; Shirani et al., 2020; White & Miller, 2024).

The literature acknowledges the contradictory impacts of smart homes and the varied perceptions of their benefits among vulnerable populations (Shirani et al., 2020; Sovacool et al., 2021). Vulnerable people are those who are in a dynamic state of being and are exposed to emotional, physical, or other harmful forces (Calo, 2016). Researchers have noted that smart homes expose users to potential security and privacy risks (Hammi et al., 2022). Furthermore, the distribution of benefits is arguably inequitable, depending on user characteristics—such as age, income level, and technical skill (Brown & Markusson, 2019; Sovacool et al., 2021). Studies suggest that while smart homes can empower elderly individuals with health conditions, they may be challenging for those facing financial struggles and who lack technical expertise (Sovacool et al., 2021). Conversely, older adults with medical needs may lack the tech-savviness and confidence to fully realise the benefits of smart homes (Balta-Ozkan et al., 2013; Brown & Markusson, 2019), exacerbating vulnerability among those who could benefit most from this technology (Shirani et al., 2020).

Existing evidence on the perceptions and adoption of smart homes among vulnerable individuals is inconclusive on two counts. First, there is limited understanding of the perceptions and challenges associated with smart home adoption beyond the binary division of elderly (often financially stable, with medical needs but less tech-savvy) versus younger populations (typically more innovative, risk-tolerant, and tech-literate but less financially secure). Second, the literature lacks insights into the perceived suitability of smart homes and key features for people with diverse vulnerabilities.

Given these gaps in the literature, this study first adopts an exploratory inductive approach to gather field-based insights into smart home perceptions among individuals from various social groups. The aim is not only to identify users' perceptions but also to discern the social characteristics that shape these perceptions and their willingness to adopt smart homes. Field-based findings revealed seven key factors—*compatibility, cost, aesthetics, comfort, uncertain usability, complexity and energy savings*—deemed significant for smart home adoption, each influenced by personal circumstances and vulnerabilities. Insights from the qualitative empirical phase were then synthesised with existing literature on smart homes and vulnerable populations to conceptualise a research model on the task-technology fit perceptions among smart home users with different vulnerabilities. This model will be further tested and generalised using a large sample size. Combining qualitative and secondary data for quantitative testing enables the conceptualisation of the relationships (cf. Fattoum et al., 2024; Morgan et al., 2005) that lack sufficient evidence in existing research.

2.0 Exploratory stage

2.1 Inductive approach

The inductive exploratory stage concerned the investigation of consumer responses to eight different energy-efficient technologies, such as heat-retaining window coating (a coating applied to windows to improve their thermal performance), infrared panels (heat emitting panels warming up objects rather than air in a room), and air bricks (ventilation holes automatically opening and shutting to improve air flow in a building). Participants were invited to visit a purpose-built test-site in the northeast of England (Futures Close), where they could experience the technologies installed in demonstrator houses. Futures Close consists of five types of property (nine houses in total), each built to original building standards and designed to be representative of UK housing stock from 1910 through to the modern day.

Participants engaged in short introductory interviews, followed by a tour of the demonstrator houses. They were shown the technologies and invited to ask questions and provide any feedback. Members of the research team listened to the discussions during the tour and noted any relevant comments. The three-hour site visit concluded with a short focus group discussion and those who took part received a £60 shopping voucher to compensate for the time spent on participation in the study.

Five site visits took place over three days in September 2024, with participants split into groups of between four and six (27 participants in total). This included 17 men and 15 women of various ages. The participants came from different areas across the north east and represented a diverse range of personal circumstances (e.g., household make-up, financial security, employment, health conditions).

The notes taken during the tour and focus groups were transcribed using the methodology developed by Strauss and Corbin (1998). Two researchers independently analysed the collected qualitative data to arrive at coded themes to ensure the reliability of findings and eliminate the possibility of biases in data interpretation. The researchers began by coding the words and concepts expressed by participants, followed by coding patterns in the text that conveyed similar meanings. The codes were then grouped into second-order categories based on underlying relationships. Deviations in the produced codes were then discussed to arrive at a final unified coding scheme.

2.2 Field-based insights

Initial findings identified seven factors influencing willingness to adopt, relating to the product's *physical and practical characteristics*, the *sensory experiences* associated with adoption, and *uncertain usability*.

Physical and practical aspects include *compatibility* with the home and *cost*. Physical compatibility refers to the fabric or layout of the property, which in some cases made the technologies unsuitable for certain homes. Cost concerns were varied, encompassing not only the initial investment but also installation, maintenance, and replacement costs, the latter connected to uncertainty about the product's durability. Participants also mentioned running costs, potential unexpected bills if issues arose, the expense of redecorating after retrofitting, and considerations around return on investment (ROI).

Sensory experiences refer to the product's *aesthetics* and *comfort* during use. Participants discussed the visual appeal and how well products fit within the home's aesthetic, considering both appearance and alignment with the house's style. Comfort perceptions also played a role; for example, one product was described as both "stuffy" and "cozy." Interestingly, these sensory impressions influenced perceptions of practical factors. For instance, one participant initially dismissed a technology as "a nightmare to put in" but, after experiencing the comfort it provided, reassessed its value and viewed the potential disruption as "worth it."

Finally, participants expressed concerns about the *usability*, *complexity*, and *energy efficiency* of products. Complexities around installing, using, and repairing technologies—or finding qualified help to do so—were significant considerations. Participants were worried about installation and the resulting cost, disruptions and the time required. As a way to avoid such disruptions, some participants were more comfortable with add-on technologies that did not require "fundamental change" or "breaking the fabric" of the house. Also, many wanted more data supporting energy-saving claims and preferred to read reviews from others who had installed and used the products.

These uncertainties appeared to stem from participants' vulnerabilities, including limited *knowledge*, *age-related considerations* and *monetary concerns*. Many participants were unfamiliar with the products outside of this project, while others felt overwhelmed by the range of available options. Trust was also a significant concern; participants were cautious about companies' motives and skeptical of marketing claims. Monetary concerns often correlated with age: some older participants, while financially secure enough to afford the initial investment, worried that, at their stage of life, they might not see a return on that investment.

3.0 The synthesis of field-based insights with theory: Conceptualising task-technology fit factors and vulnerabilities

This study uses the Task-Technology Fit (TTF) framework by Goodhue and Thompson (1995) to explore how the perceptions of smart homes by people with vulnerabilities influence the assessment of technology fit to their personal goals. Task-technology fit refers to the degree to which technology aids in performing specific tasks (Goodhue & Thompson, 1995). A user's evaluation of task-technology fit is shaped by both the characteristics of the task and the technology's attributes. Task characteristics are assessed based on factors that may influence a user's reliance on specific features of the technology, while technology characteristics relate to the unique attributes or functionalities of the technology itself (Goodhue & Thompson, 1995).

Consistent with the TTF framework, several key considerations regarding smart home technology identified at the exploratory stage of this study include: cost (covering installation, operating expenses, and financial risks), usability/complexity (ease of use and learning curve), comfort (convenience provided by the technology), sustainability (energy-saving benefits), and aesthetics, and product compatibility as context-specific factors. Cost, usability/complexity, comfort, and sustainability align with existing research on smart home adoption (e.g., Hubert et al., 2019; Marikyan et al., 2019; Marikyan et al., 2021; Papagiannidis & Davlembayeva, 2022). In particular, high upfront costs can deter adoption, while affordability and transparent pricing can enhance accessibility and appeal (Li et al., 2021). Complex systems are discouraging, especially for less tech-savvy individuals (Harris et al., 2022). Therefore, usability is an important driver as it indicates that smart home technology seamlessly fits their skill levels and does not require cognitive effort (Marikyan et al., 2021). Comfort, often tied to convenience (e.g., automation of mundane tasks, voice control, and remote access to devices), underscores the value consumers place on technologies that simplify daily routines, and enhance user experience and satisfaction (Balta-Ozkan et al., 2013; Baudier et al., 2020). Also, the features optimising the use of resources and energy consumption appeal to users seeking to save on utility bills and minimise ecological footprint (Papagiannidis & Davlembayeva, 2022). Additionally, previous literature highlights the role of perceived usefulness, defined as "*the degree to which an individual believes that using the system will help attain gains in job performance*"(Venkatesh et al., 2003), in shaping perceptions of technology fit for household tasks (Marikyan et al., 2021).

Our synthesis of field-based insights and research findings suggests that three types of vulnerabilities—conditions that expose individuals to potential harm—may influence perceptions of technology suitability for household tasks. The first type of vulnerability arises from limited familiarity with and knowledge of technology. A lack of skills and understanding can lead to fear (Brown & Markusson, 2019) and low self-confidence in using technology effectively to improve quality of life (Balta-Ozkan et al., 2013; Sovacool et al., 2021). The second type stems from limited financial resources, where smart homes are perceived as non-essential luxuries, potentially increasing user dependence on external experts (Sovacool et al., 2021). The third type of vulnerability is related to health limitations, which are often not examined independently of age groups. However, as Shirani et al. (2020) emphasise, “it is important that a focus on elderly people does not result in overlooking the needs of other potentially vulnerable groups, such as those with disabilities and families with young children”. Many smart home devices currently on the market do not adequately address the needs and preferences of vulnerable individuals, exacerbating the digital divide and limiting the potential benefits of smart homes (Shirani et al., 2020).

The resulting conceptual model is presented in figure 1.

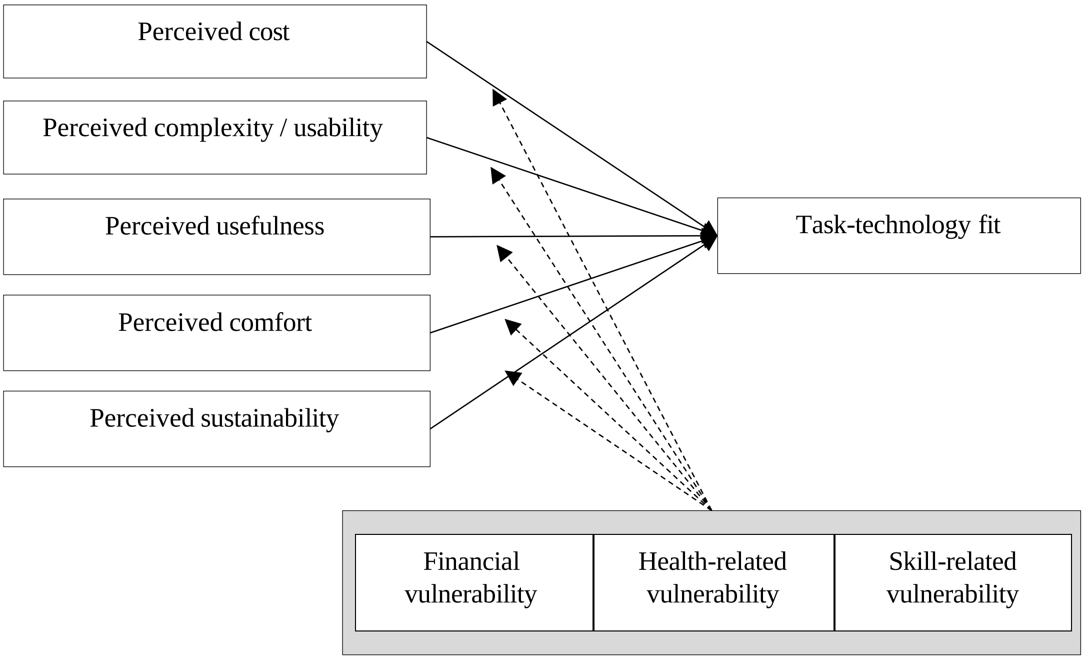


Figure 1. The perception of task-technology fit factors by people with vulnerabilities

4.0 Conclusions, future steps and study implications

This study used an exploratory approach to identify the factors central to the adoption of smart homes and synthesised the findings with the existing evidence to conceptualise how the perception of these factors is shaped by users' individual circumstances and vulnerabilities. The next stages of this study will focus on validating the research model on smart home responses and perceived fit using a sample of approximately 500 individuals facing various financial, health-related, and skill or knowledge-based vulnerabilities. These findings aim to expand the literature on smart home adoption, which has predominantly examined usage patterns by age groups (e.g., Sovacool et al., 2021), thereby limiting insights into how individuals with different types of limitations might respond to such technology.

This research has practical implications for policymakers and social services by identifying the potential barriers to adopting smart home technologies. These insights could facilitate targeted support measures, such as subsidies for low-income users, technical assistance programs, and guidelines that ensure ethical, accessible design. When smart home technology is perceived as overly complex, intrusive, or irrelevant, vulnerable users may hesitate to adopt it or discontinue its use soon after installation. A clearer understanding of their concerns and preferences can encourage sustained adoption by developing solutions that fit their lifestyles and meet genuine needs.

While smart home technology has the potential to improve quality of life, if it is not designed with vulnerable users in mind, it risks reinforcing social inequities. Recognising and addressing the unique needs of vulnerable groups can help distribute the benefits of this technology more equitably, ensuring that all population segments can access advances in safety, convenience, and overall well-being.

References

Balta-Ozkan, N., Davidson, R., Bicket, M., & Whitmarsh, L. (2013). Social barriers to the adoption of smart homes. *Energy Policy*, *63*, 363-374. https://doi.org/https://doi.org/10.1016/j.enpol.2013.08.043

Baudier, P., Ammi, C., & Deboeuf-Rouchon, M. (2020). Smart home: Highly-educated students' acceptance. *Technological Forecasting and Social Change*, *153*, 119355.

Brown, C. J., & Markusson, N. (2019). The responses of older adults to smart energy monitors. *Energy Policy*, *130*, 218-226.

Calo, R. (2016). Privacy, vulnerability, and affordance. *DePaul L. Rev.*, *66*, 591.

Fattoum, A., Chari, S., & Shaw, D. (2024). Configuring systems to be viable in a crisis: The role of intuitive decision-making. *European Journal of Operational Research*, *317*(1), 205-218.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly*, 213-236.

Hammi, B., Zeadally, S., Khatoun, R., & Nebhen, J. (2022). Survey on smart homes: Vulnerabilities, risks, and countermeasures. *Computers & Security*, *117*, 102677. https://doi.org/https://doi.org/10.1016/j.cose.2022.102677

Harris, M. T., Blocker, K. A., & Rogers, W. A. (2022). Older adults and smart technology: facilitators and barriers to use. *Frontiers in Computer Science*, *4*, 835927.

Hubert, M., Blut, M., Brock, C., Zhang, R. W., Koch, V., & Riedl, R. (2019). The influence of acceptance and adoption drivers on smart home usage. *European Journal of Marketing*, *53*(6), 1073-1098.

Li, W., Yigitcanlar, T., Erol, I., & Liu, A. (2021). Motivations, barriers and risks of smart home adoption: From systematic literature review to conceptual framework. *Energy Research & Social Science*, *80*, 102211.

Marikyan, D., Papagiannidis, S., & Alamanos, E. (2019). A systematic review of the smart home literature: A user perspective. *Technological Forecasting and Social Change*, *138*, 139-154.

Marikyan, D., Papagiannidis, S., & Alamanos, E. (2021). “Smart Home Sweet Smart Home”: An Examination of Smart Home Acceptance. *International Journal of E-Business Research (IJEBR)*, *17*(2), 1-23.

Morgan, N. A., Anderson, E. W., & Mittal, V. (2005). Understanding firms’ customer satisfaction information usage. *Journal of Marketing*, *69*(3), 131-151.

Papagiannidis, S., & Davlembayeva, D. (2022). Bringing smart home technology to peer-to-peer accommodation: Exploring the drivers of intention to stay in smart accommodation. *Information Systems Frontiers*, *24*(4), 1189-1208.

Shirani, F., Groves, C., Henwood, K., Pidgeon, N., & Roberts, E. (2020). ‘I'm the smart meter’: Perceptions of smart technology amongst vulnerable consumers. *Energy Policy*, *144*, 111637. https://doi.org/https://doi.org/10.1016/j.enpol.2020.111637

Sovacool, B. K., Martiskainen, M., & Furszyfer Del Rio, D. D. (2021). Knowledge, energy sustainability, and vulnerability in the demographics of smart home technology diffusion. *Energy Policy*, *153*, 112196. https://doi.org/https://doi.org/10.1016/j.enpol.2021.112196

Strauss, A., & Corbin, J. (1998). Basics of qualitative research techniques.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 425-478.

White, B., & Miller, M. (2024). Ambient smart environments: affordances, allostasis, and wellbeing. *Synthese*, *204*(2), 48.

Measuring Vulnerability in Financial Services Customers: A Scale for Proactive Solutions

Research In progress

Hai Nguyen, Tommi Tapanainen, Suhwa Ou, Chandan Padhan

Hai Nguyen, University of South Wales, hai.nguyen1@southwales.ac.uk Tommi Tapanainen, Busan National University, tojuta@gmail.com Suhwa Ou, Soochow University, suhua.scu@gmail.com
Chandan Padhan, UK Health Security Agency (UKHSA) chandanpadhan@outlook.com

Abstract (around 150 words)

Financial organizations face significant risks when they neglect vulnerable customers, including increased default rates, regulatory challenges, and damage to their reputation. Traditional, reactive service models, which primarily address vulnerabilities during in-person interactions, are becoming less effective as branch offices decline. As a result, proactive measures are now essential to support these customers and protect financial institutions. While previous research on the digital divide in financial services has explored vulnerability, it has often focused on specific groups, such as those with financial limitations. However, vulnerability is a multifaceted issue that encompasses a broader range of factors. Given the growing importance of digital transformation in the financial industry, various data sources such as transactional, behavioural, external, and socio-demographic datasets can be processed through data analytics and AI to identify vulnerability. Therefore, developing a comprehensive vulnerability scale would help financial institutions proactively assist vulnerable individuals, improving both customer support and the institution's long-term objectives.

Keywords: Vulnerability, Financial Services, Inclusion Scores, Proactive Approach, Digital Transformation, Fintech

1.0 Introduction

Vulnerability in financial services or financial inclusion/exclusion refer to the availability and use of financial services (Allen et al., 2016). Surveys of vulnerability in financial services show that half or more of UK adults exhibited vulnerability and low financial resilience, and that the proportion of people with low financial resilience had grown significantly since 2020 (Financial Lives Survey, 2022). Indeed, the situation is reflected around the world, not only in advanced countries such as the United States (CFPB, 2019) and Australia (NAB, 2019) but also in emerging markets such as Nigeria (Mogaji et al., 2021).

For financial organizations, vulnerable customers poses several risks, including increased financial risks such as higher default rates and non-performing loans, regulatory and compliance challenges, and potential legal and reputational consequences. Neglecting vulnerable customers can lead to customer attrition, loss of trust, and decreased profitability.

Financial organizations have traditionally operated within a reactive service model, addressing vulnerabilities as they become evident when customers visit the branch-office and have a meeting face-to-face. However, the reduction in branch offices is compounding the problem, with fewer opportunities to even identify the vulnerable customers. Therefore, intervention is needed to mitigate financial vulnerability, not only to support customers but also to benefit financial organizations. Addressing customer demands through identifying vulnerable individuals and meeting their needs with proactive banking practices has become more crucial than ever.

In this research we aim to answer the following research questions:
Research Question 1: *Who are the vulnerable customers of financial institutions, particularly in light of the rapid shift of financial services to the digital world?*
Research Question 2: *How can financial organizations develop a vulnerability score as a proactive measure to strengthen their financial inclusion strategies?*

2.0 Literature review and research problems

2.1 Who are the vulnerable customers of financial institutions?

Vulnerability is a multifaceted issue with overlapping dimensions (see Table 1, below). Anyone can become vulnerable, and it is in fact expected that most people will experience some form of vulnerability during their lifetime (de la Cuesta et al., 2021; Thomson et al., 2020).

Table 1: Vulnerable Group Review
(Tentative results of the first-round literature review)

Vulnerable groups	**Definition**
Financial	Financial vulnerability refers to customers facing financial

Vulnerability	hardship, poverty, and limited access to traditional banking products, making informed decisions, and navigating complex systems challenging (Greene & Stavins, 2021; Mogaji et al., 2021; Nam & Lee, 2023; Rosenbaum et al., 2017).
Digital Vulnerability	Customers with limited technology access and low digital literacy have difficulties to use digital services from private and public providers (Nam & Lee, 2023).
Geographic Vulnerability	Customers residing in remote or underserved areas may have limited access to physical banking services, such as branches or ATMs. This can create barriers to accessing financial services, particularly for those who prefer or require in-person interactions or have limited access to digital channels (Conrad et al., 2019; Fernández-Gutiérrez & Ashton, 2023; Guerra-Leal et al., 2021; Roa et al., 2021).
Age-related Vulnerability	Younger customers face financial risks due to lack of experience and literacy (Roa et al., 2021), while older adults face challenges like cognitive decline and physical disabilities, hindering effective financial management (Conrad et al., 2019; Fernández-Gutiérrez & Ashton, 2023).
Health and Disability	Customers with health conditions, disabilities, or cognitive impairments may face barriers in accessing and navigating banking services, including mobility issues, complex information comprehension, and communication challenges. (Rosenbaum et al., 2017).
Low literacy or limited proficiency	Customers may face challenges in understanding financial terminology, product information, and communication materials, hindering informed decision-making and service access (Bathula & Gupta, 2021; Conrad et al., 2019).
Psychosocial factors and lack of cognitive ability	Warmath et al. (2023) suggest that various psychosocial factors, such as self-control, conscientiousness, goal confidence, and future orientation, relate to a person’s vulnerability to financial hardship.
Gender-based	There is research linking females with lower likelihood of using

vulnerability	digital banking, particularly in developing or emerging economies such as India (Bathula & Gupta, 2021) and Mexico (Guerra-Leal et al., 2021)
Situational Vulnerability	Customers may experience temporary vulnerability due to life events or transitions, such as job loss, illness, bereavement, or taking on caring responsibilities. These situations can impact their financial stability, ability to manage banking affairs, and overall well-being, making it challenging to navigate banking services during such periods (Bathula & Gupta, 2021; Salisbury et al., 2023).

2.2 The existing reactive approach to identifying vulnerability

The current vulnerability management practices in several financial organizations, such as Barclays (KPMG, 2021), primarily rely on manual, reactive approaches. Vulnerability markers are applied only after customer interactions or reported incidents, which means opportunities for early intervention are often missed. This delay can hinder the identification and support of vulnerable customers, prolonging their exposure to financial risks (Mogaji et al., 2021). Additionally, relying solely on manual processes and superficial interpretations of customer interactions may overlook subtle indicators of vulnerability (González et al., 2022; Rosenbaum et al., 2017), leading to the neglect of underlying issues (Alalwan et al., 2018). Furthermore, detecting vulnerability can be particularly challenging in cases where a customer shows low financial engagement. For example, Vik et al. (2024) found that even if an individual opened a bank account, it didn’t necessarily indicate active use of the account.

2.3 Toward a proactive approach to identify vulnerability

Financial organizations have been applying various statistical techniques and machine learning to create credit scores. These scores transform relevant data into numerical measures widely used to evaluate business, real estate, and consumer loans. Credit scoring is one of the most successful applications of research modelling in finance and banking, as evidenced by the increasing number of scoring analysts in the industry (Abdun and Pointon, 2011). However, credit scores primarily focus on payment history, credit utilization, length of credit history, types of credit used, recent credit inquiries, debt-to-income ratio, and public records, concentrating mainly on the customers' credit situation.

To proactively identify and support vulnerable customers, financial organizations could develop a vulnerability score, similar to a credit score, but incorporating a wider range of data sources to capture a broader definition of vulnerability. This vulnerability score would assess an individual’s or entity’s risk of facing adverse outcomes due to various vulnerability factors.

In fact, the finance industry’s ongoing digital transformation, encompassing mobile banking apps and other Fintech services, holds significant potential for enhancing support for vulnerable customers. Numerous data sources such as transactional, behavioural, external, and socio-demographic datasets, can be processed through data analytics and AI to enable the identifying of vulnerability. Usage of automation would also reduce costs involved with simply scaling up the reactive approach to a collaborative effort within banks. In other words, digital transformation can equip financial organizations with a more proactive strategy.

3.0 Research methods

To address the research questions, the following research methods are applied (Campbell et al., 2002; Leavy, P., 2022; Rowe & Wright, 2001).

Stage 1: Creating the Initial Feature List

The first step in building vulnerability scores is creating an initial feature list that captures multi-dimensional aspects of vulnerability, as shown in Table 1. This comprehensive list is derived from a literature review and international initiatives, such as the Digital Financial Services Consumer Competency Framework (Fundira et al., 2023; ITU, 2021).

Stage 2: Pilot Survey to Validate and Test the Vulnerability Scores

A pilot survey, designed to collect data on various aspects of individuals' financial lives, will be conducted to validate the initial feature list. This survey will be administered in collaboration with local banks in selected countries, and statistical analysis techniques will help identify the most predictive features.

Stage 3: Domain Expert Seminar to Review and Finalize the Vulnerability Scores

A seminar will be held to gather feedback from domain experts on the inclusion scores generated from the literature review and survey data (Campbell et al., 2002; Rowe & Wright, 2001). Experts in social work, healthcare, finance, and data analysis will review the selected features to ensure they are relevant across social, business, and data perspectives. Diversity and inclusion professionals, along with representatives from underserved areas, will participate to ensure the scores are inclusive and relevant to the target audience.

Stage 4: Validation with Secondary Data and Practical Application

Following the seminar, a larger-scale survey will be conducted to further verify the inclusion scores. Additionally, data firms may validate these scores using secondary data sources, such as administrative records, third-party data, or publicly available datasets. Financial organizations, as well as other businesses like insurance companies and healthcare providers, may then consider integrating these inclusion scores into their scoring systems.

4.0 Expected results and contributions

In terms of theoretical implications, this research aims to establish a foundation for serving vulnerable customers by identifying its critical areas and developing a validated inclusion score. We contribute to research in the Digital Divide, particularly in relation to finance, by concretely specifying a complex individual profile that combines features of different vulnerabilities or “divides”, rather than a single statistic such as access to Internet and/or low income and/or digital literacy. The Digital Divide research has repeatedly emphasized the overlapping vulnerabilities at the national and macroeconomic levels (e.g. Khera et al., 2022), but often limiting to descriptive rather than prescriptive conclusions. Hence, our research can offer ways forward.

In terms of practical implications, we will provide recommendations for policymakers and financial institutions on how to use these scores to assist vulnerable populations and implement early warning systems to proactively identify and support at-risk individuals. Additionally, other service sectors, such as insurance and healthcare, could adopt this approach to verify the inclusion score for their specific needs. Finally, the results of this study could be adapted for global application, as inclusion is a global issue.

References

Abdun, H.A. & Pointon, J. (2011). *Credit scoring, statistical techniques and evaluation criteria: a review of the literature*. Intelligent systems in accounting, finance and management, 18(2-3), 59-88.

Alalwan, A.A., Dwivedi, Y.K., Rana, N.P. & Algharabat, R. (2018). *Examining factors influencing Jordanian customers' intentions and adoption of Internet banking: Extending UTAUT2 with risk*. Journal of Retailing and Consumer Services, 40, 125-138.

Allen, F., Demirguc-Kunt, A., Klapper, L., & Martinez Peria, M. S. (2016). *The foundations of financial inclusion: Understanding ownership and use of formal accounts*. Journal of Financial Intermediation, 27, 1–30.

Bathula, S. & Gupta, A. (2021). The determinants of financial inclusion and digital financial inclusion in India: A comparative study. The Review of Finance and Banking, 13(2), 109-120. http://dx.doi.org/10.24818/rfb.21.13.02.02.

Campbell, S. M., Braspenning, J. A., Hutchinson, A., & Marshall, M. (2002). *Research methods used in developing and applying quality indicators in primary care*. Quality and Safety in Health Care, 11(4), 358-364.

Cnaan, R.A., Scott, M.L., Heist, H.D. & Moodithaya, M.S. (2023). *Financial inclusion in the digital banking age: Lessons from rural India*. Journal of Social Policy, 52(3), 520-541.

Conrad, A., Neuberger, D., Peters, F. & Rösch, F. (2019). *The Impact of Socio-Economic and Demographic Factors on the Use of Digital Access to Financial Services*. Credit and Capital Markets, 52(3), 295–321.

de la Cuesta-González, M., Fernandez-Olit, B., Orenes-Casanova, I., & Paredes-Gazquez, J. (2022). *Affective and cognitive factors that hinder the banking relationships of economically vulnerable consumers*. International Journal of Bank Marketing, 40(7), 1337-1363.

Fernández-López, S., Álvarez-Espiño, N., Rey-Ares, L. & Castro-González, S. (2023). *Consumer financial vulnerability: Review, synthesis, and future research agenda*. Journal of Economic Surveys, 38, 1045-1084.

Fundira, M., Edoun, E.I. & Pradhan, A. (2023). *Adapting to the digital age: Investigating the frameworks for financial services in modern communities*. Business Strategy and Development, 7, e303, https://doi.org/10.1002/bsd2.303

Greene, C., & Stavins, J. (2021). *Income and banking access in the USA: The effect on bill payment choice*. Journal of Payments Strategy and Systems, 15(3), 244–249.

Guerra-Leal, E.M., Arredondo-Trapero, F.G. & Vazquez-Parra, J.C. (2021). *Financial inclusion and digital banking on an emergent economy*. Review of Behavioral Finance, 15(2), 257-272.

ITU (2021). *Digital Financial Services Consumer Competency Framework.* International Telecommunication Union report. https://figi.itu.int/wp-content/uploads/2021/04/Digital-Financial-Services-Consumer-Competency-Framework-1.pdf

Khera, P., Ng, S., Ogawa, S. & Sahay, R. (2022). *Measuring Digital Financial Inclusion in Emerging Market and Developing Economies: A New Index*. Asian Economic Policy Review, 17, 213-230. doi: 10.1111/aepr.12377

Leavy, P. (2022). *Research design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches*. Guilford Publications.

Mogaji, E., Adeola, O., Hinson, R. E., Nguyen, N. P., Nwoba, A. C., & Soetan, T. O. (2021). *Marketing bank services to financially vulnerable customers: evidence from an emerging economy*. International Journal of Bank Marketing, 39(3), 402-428.

Monferrer Tirado, D., Vidal-Meliá, L., Cardiff, J., & Quille, K. (2023). *Vulnerable customers' perception of corporate social responsibility in the banking sector in a post-crisis context*. International Journal of Bank Marketing.

Nam, Y & Lee, S.T. (2023). *Behind the growth of FinTech in South Korea: Digital divide in the use of digital financial services*. Telematics and Informatics, 81, 101995.

Rosenbaum, M.S., Seger-Guttmann, T. & Gilardo , M. (2017). Commentary: Vulnerable consumers in service settings. Journal of Services Marketing, 31(4-5), 309-312.

Rowe, G., & Wright, G. (2001). *Expert opinions in forecasting: the role of the Delphi technique*. Principles of forecasting: A handbook for researchers and practitioners, 125-144.

Salisbury, L.C., Nenkov, G.Y., Blanchard, S.J., Hill, R.P., Brown, A.L. & Martin, K.D. (2023). *Beyond Income: Dynamic Consumer Financial Vulnerability*. Journal of Marketing, 87(5), 657-678.

Trivedi, S.K. (2020). *A study on credit scoring modelling with different feature selection and machine learning approaches*. Technology in Society, 63, 101413.

Vik, P.M., Kamerade, D. & Dayson, K.T. (2024). *The Link Between Digital Skills and Financial Inclusion—Evidence from Consumers Survey Data from Low-Income Areas*. Journal of Consumer Policy, 47, 373-393.

Warmath, D., O'Connor, G.E., Wong, N. & Newmeyer, C. (2022). *The role of social psychological factors in vulnerability to financial hardship*. Journal of Consumer Affairs, 56, 1148-1177.

Yang, B., Wang, X., Wu, T. & Deng, W. (2023). *Reducing farmers' poverty vulnerability in China: The role of digital financial inclusion*. Review of Development Economics, 27(3), 1445-1480.

Resisting Deletion: Algorithmic Governance and The Right to Be Forgotten

Research In progress

Klara Källström, Marie Eneman, Jan Ljungberg
Faculty of Science and Technology, University of Gothenburg, Sweden

Abstract

As digital images become fluid, mobile, and embedded within algorithmic infrastructures, this paper examines the politics of visibility in algorithmic governance, focusing on digital archives, data retention, and the governance of deletion. Using the Swedish police's adoption of Clearview AI's facial recognition technology as a case study, it explores how networked images resist erasure, challenging the enforcement of the Right to Be Forgotten (RTBF) under the EU's General Data Protection Regulation (GDPR). By analyzing policy documents, institutional workflows, and bureaucratic decision-making, the paper asks: How do archival practices influence the ability to uphold the right to be forgotten within algorithmic governance? The findings reveal tensions between public and private data infrastructures, where automated decision-making, predictive analytics, and archival mechanisms entrench networked images in state operations. This study calls for a re-evaluation of privacy, digital archives, and algorithmic governance, as deletion remains a politically and technically contingent act within distributed systems.

Keywords: Algorithmic governance, Distributed agency, Networked images, Politics of visibility, Facial recognition, Archival infrastructures, The Right to Be Forgotten (RTBF), Data surveillance

1.0 Introduction

The implementation of digital systems in public administration is shaped by negotiations, institutional constraints, and classification struggles (Bowker & Star, 1999). As technological infrastructures become increasingly interwoven with commercial actors, critical issues of control, power, and transparency emerge, reflecting the co-production of technology and governance (Jasanoff, 2004) and the growing reliance on predictive AI models in regulatory and bureaucratic decision-making (Amoore, 2024).

This paper examines the Swedish police's use of a controversial facial recognition application, situating it within the broader challenges of algorithmic governance, legal ambiguities, and the intersection of public and private interests (Kosta, 2022). The resulting authority statements reveal not only the infrastructural messiness of law enforcement's adoption of new technologies but also the tensions between technical systems and governance frameworks (Ananny & Crawford, 2018).

Building on these governance challenges, this paper analyzes the sociotechnical dimensions of Swedish authorities' use of facial recognition technology, particularly by conceptualizing the image as a digital object—fluid, mobile, and embedded within algorithmic infrastructures (Dewdney & Sluis, 2023; Paglen, 2019). As governance infrastructures become increasingly data-driven, images shift from stable representations to networked entities that circulate through algorithmic systems, shaping decision-making processes and classification mechanisms (Rouvroy & Berns, 2013).

Automated decision-making and predictive analytics (Amoore, 2020; Stahl et al., 2023) increasingly shape data governance, raising critical privacy and epistemological concerns in the generation and interpretation of visual data (Crawford, 2021; Pasquale, 2015). As algorithmic images proliferate, this study examines the Right to Be Forgotten (RTBF), a provision of the EU's General Data Protection Regulation (GDPR), which aims to mitigate the enduring risks of privacy breaches. However, data infrastructures often resist erasure, as deletion becomes a political act embedded within archival logics and classification systems (Bowker, 2005; Kosta, 2022; Mantelero, 2013; Star & Ruhleder, 1996).

Official statements and archival practices within Swedish government institutions are analyzed to understand how bureaucratic infrastructures and classifications determine what is remembered and forgotten, shaping the production, preservation, and transformation of knowledge over time (Bowker, 2005; Jasanoff, 2004). Ultimately, the networked image is conceptualized both as a technological infrastructure and as a dynamic of social relations (Dewdney & Sluis, 2023), reflecting the entanglement of digital surveillance, governance, and data flows in contemporary state operations (Ananny & Crawford, 2018).

Thus, the guiding question for this paper is: How do archival practices shape the ability to enforce the *right to be forgotten* within algorithmic governance?

Rather than operating as a neutral legal safeguard, the GDPR and its Right to Be Forgotten (RTBF) provision must be understood within the sociotechnical infrastructures that mediate its enforcement (Bennett & Raab, 2020; Bowker & Star, 1999; Jasanoff, 2004). While GDPR is designed to enhance individuals' control over personal data and streamline regulatory compliance (Council of the European Union, 2015), its implementation is shaped by institutional constraints, bureaucratic classification systems, and algorithmic data retention structures (Kosta, 2022).

Among GDPR's key provisions, RTBF allows individuals, under certain conditions, to request the removal of personal information from internet searches and databases (Council of the European Union, 2018). However, data infrastructures often resist erasure, as deletion becomes a politically and technically contingent act, embedded within archival logics (Bowker, 2005). The findings of this paper suggest that RTBF enforcement is not simply a legal issue but an infrastructural one, requiring a deeper examination of how digital archives, classification regimes, and algorithmic processing shape the governance of visibility and deletion in state operations (Mantelero, 2013; Rouvroy & Berns, 2013).

2.0 Theoretical foundation

The concept of *informational privacy* in digital environments is often framed as granting individuals the ability to dissociate from past actions or data that might otherwise persist indefinitely (Floridi, 2015; Richards, 2022; Véliz, 2020). Central to this perspective is the notion of *personal data*, which serves as the foundation of one's digital identity (Floridi, 2014; Richards, 2022; Véliz, 2020). However, this framing assumes a unified and autonomous individual, whose identity is intrinsically tied to data ownership. Gilles Deleuze's concept of the *dividual* disrupts this notion, proposing that digital environments do not represent a cohesive, singular self, but instead function through fragmented and divisible entities (Deleuze, 1992). Unlike the individual, the dividual is not defined by a stable identity but by the data points, metrics, and traces it leaves behind (Amoore, 2020; Bowker & Star, 1999).

If the dividual is composed of dispersed, analyzable fragments, then informational privacy cannot simply be understood as an individual's control over their data. Instead, privacy becomes a question of how these fragments are aggregated, interpreted, and used to classify, predict, and govern behavior (Amoore, 2024; Lyon, 2003). This shifts the focus from individual data ownership to the infrastructural systems that process and regulate data flows, raising critical concerns about whether personal data adequately captures the complexity of identity and privacy in algorithmic governance (Jasanoff, 2004).

When personal data is incorporated into larger, algorithm-driven systems, the dividual is increasingly shaped by automated infrastructures that operate beyond individual control (Amoore, 2020). This raises concerns about how predictive analytics and generative AI models influence governance frameworks, reinforcing algorithmic decision-making as a mode of control (Amoore, 2024; Pasquale, 2015). Concerns thus emerge regarding the capacity to regulate how this information is processed and made actionable (Ananny & Crawford, 2018). In environments dominated by automated classification systems, infrastructures often operate autonomously, shaping political and ethical decisions without direct human oversight (Bowker, 2005; Suchman, 1994). This complicates efforts to manage or delete data, as it becomes less about erasing a single identity and more about controlling the governance of fragmented data points across global information ecosystems (Star & Ruhleder, 1996).

As algorithms and data infrastructures increasingly determine what is remembered and forgotten, privacy concerns become even more intricate. The Right to Be Forgotten (RTBF) assumes that deleting data restores privacy, but this is insufficient in a networked environment where data fragments circulate across multiple systems. The act of deletion itself is mediated by infrastructures of classification and bureaucratic logic (Bowker & Star, 1999). Nissenbaum's (2004) concept of contextual integrity offers a lens through which to address these challenges, arguing that privacy is not about absolute control over data, but about ensuring that information flows adhere to the norms and expectations of the context in which they were generated.

The norms governing contextual integrity may become increasingly strained when data fragments are reassembled and reinterpreted by algorithms across disparate contexts. For example, data collected for one purpose may be recombined and used to generate predictive risk assessments, violating contextual boundaries and expectations (Pasquale, 2015). The dividual, as a composite of fragmented data points, thus challenges traditional notions of privacy that assume the integrity of a singular, autonomous self, underscoring the need to rethink privacy governance within distributed systems (Amoore, 2020; Jasanoff, 2004).

In this framework, privacy and data governance must account for the dividual as both a product and a target of algorithmic systems (Latour, 2005). Contextual integrity, therefore, highlights the need to regulate not just data flows but also the algorithmic processes that reconstitute and repurpose personal data in ways that undermine privacy and agency (Nissenbaum, 2004, 2019; Pasquale, 2015).

Historically, archival practices have been shaped by state control, science, and surveillance, influencing contemporary government approaches to data management and categorization (Bowker, 2005; Geoghegan, 2023; Jasanoff, 2004; Rouvroy & Berns, 2013;). The interplay between documentation, surveillance, and archiving—especially with photographic images—has long enabled authorities to classify individuals, reinforce bureaucratic structures, and govern populations (Lyon, 2003, 2014; Sekula, 1986). The advent of photographic classification in criminology and policing established state surveillance frameworks that persist today, particularly in biometric data collection and facial recognition technologies (Crawford, 2021). These historical legacies continue to shape contemporary governance, reinforcing power asymmetries in digital privacy, data retention, and algorithmic decision-making (Ananny & Crawford, 2018).

The shift from traditional archives to expansive digital repositories—capable of storing vast amounts of data indefinitely—introduces new regulatory and infrastructural complexities (Geoghegan, 2023; Gitelman, 2013; Bowker & Star, 1999). Unlike physical archives, which were limited by spatial constraints, digital archives expand exponentially, raising concerns about control, access, and retention policies (Bowker, 2005). This evolution necessitates a re-evaluation of data governance strategies, particularly in how deletion, access, and archival practices intersect with algorithmic infrastructures, surveillance, and bureaucratic classification systems (Bowker & Star, 1999).

3.0 Research design

3.1 Setting

Clearview AI’s platform leverages a vast database of over three billion images scraped from the internet without consent, raising significant concerns about data governance, classification, and surveillance infrastructures (Rezende, 2020; Shepherd, 2024). The controversy surrounding Clearview AI intensified in 2020 when BuzzFeed News leaked its customer list, revealing that law enforcement agencies in the United States, Canada, and several European countries, including Sweden, had used the application. In countries such as Canada, Finland, and Sweden, some police officers accessed Clearview AI’s system without institutional approval, highlighting the porous boundaries between state surveillance and commercial data extraction (Amoore, 2020; Eneman et al., 2022; Shepherd, 2024).

In February 2021, the Swedish Authority for Privacy Protection (IMY) determined that the Swedish Police had violated the Swedish Criminal Data Act by using Clearview AI without conducting a mandatory data protection impact assessment (Dnr 4756-21). This decision underscores the regulatory tensions between emerging biometric surveillance technologies and existing legal frameworks (Mantelero, 2013). The IMY subsequently ordered that affected individuals be notified and instructed that any data transferred to Clearview AI be erased (Eneman et al., 2022; Kosta, 2022)

3.2 Document Collection and Analysis

The empirical material for this study consists of public documents from three main authorities: the IMY (Swedish Authority for Privacy Protection), the Police Authority, and the Administrative Court in Stockholm. In Sweden, public access to government records is governed by the Principle of Public Access to Information, a foundational right established in the Freedom of the Press Act (194:105). This principle aims to ensure transparency in state operations but is increasingly complicated by the shift from analog documentation to digital infrastructures of governance (Ananny & Crawford, 2018; Bowker, 2005).

The documents analyzed in this study (Bowen, 2009) provide insight into how emerging facial recognition technologies intersect with legal reasoning, institutional classification practices, and bureaucratic decision-making processes (Eneman et al., 2022).

Authority	Type of document
The Swedish Authority for Privacy Protection (IMY)	*Investigation into the use of Clearview AI* (Dnr DI-2020-2719, date: 2020-03-05) *Request for supplementation* (Dnr DI-2020-2719, date: 2020-03-30) *Decision after the inspection* (Dnr DI-2020-2719, date: 2021-02-10)
The Swedish Police Authority	*Investigation into the use of Clearview AI* (Dnr A126.614/2020, date: 2020-03-19) *Request for supplementation* (Dnr A126.614/2020, date: 2020-05-07) *Appeal regarding the IMY's decision* (Dnr A126.614/2020, date: 2021-03-01) *Completion of previously filed appeal regarding the IMY's decision* (Dnr A126.614/2020, date: 2021-03-05)
The Administrative Court in Stockholm	*Decision of the Administrative Court* (Dnr 4756-21, date: 2021-09-30)
The Court of Appeal in Stockholm	*Decision of the Court of Appeal* (Dnr 7678-21, date: 2022-11-07)

Table 1. Overview of the collected documents

4.0 Concluding Remarks

Our analysis of the police's use of Clearview AI's facial recognition technology highlights critical tensions at the intersection of archival practices, privacy rights, and algorithmic governance. These challenges are further complicated by the increasing entanglement of public and private sectors, where data infrastructures, proprietary technologies, and commercial interests blur traditional regulatory boundaries.

A key concern is the feasibility of enforcing data deletion orders once images have undergone algorithmic processing. When facial recognition or data-mining techniques are applied, images transition from static records to networked entities embedded within decision-making infrastructures (Amoore, 2020; Amoore, 2024). This raises fundamental questions about data erasure in distributed systems, as deletion from a single source does not eliminate its traces within broader algorithmic ecosystems, where information is duplicated, transformed, and integrated into predictive analytics (Amoore, 2020; Amoore, 2024).

This creates a double bind: the imperatives of security, efficiency, and technological innovation come into conflict with the structural realities of data governance, raising concerns about what can truly be erased and who has the authority to decide what remains visible or forgotten. Statements from authorities reflect these tensions: How can data protection laws such as the GDPR enforce erasure when digital objects are deeply embedded in complex, multi-layered networks of automated classification and decision-making? These challenges underscore the limitations of existing regulatory frameworks, as data governance infrastructures increasingly operate outside more established forms of legal and institutional boundaries (Crawford, 2021; Mantelero, 2013; Kosta, 2022; Amoore, 2024).

The IMY's findings emphasize the need for archival practices to evolve alongside shifting notions of privacy and visibility, particularly as governance mechanisms struggle to account for the fluid and networked nature of contemporary digital repositories. This necessitates a reassessment of data management approaches, ensuring that institutions develop frameworks for governing digital archives that take into account individual privacy, algorithmic governance, and the persistent entanglement of public and private infrastructures.

References

Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others.* Duke University Press.

Amoore, L. (2024). *A world model: On the political logics of generative AI.* Duke University Press.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Bennett, C. J., & Raab, C. D. (2020). *The governance of privacy: Policy instruments in global perspective* (2nd ed.). MIT Press

Bowker, G. C. (2005). *Memory practices in the sciences.* MIT Press.

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences.* MIT Press.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence.* Yale University Press. https://doi.org/10.12987/9780300252392

Dencik, L., Hintz, A., Redden, J., & Treré, E. (2022). *Data justice.* SAGE.

Deleuze, G. (1992). Postscript on the societies of control. *October, 59*, 3–7.

Dewdney, A., & Sluis, K. (Eds.). (2023). *The networked image in post-digital culture.* Routledge. https://doi.org/10.4324/9781003221496

Eneman, M., Ljungberg, J., Raviola, E., & Rolandsson, B. (2022). The sensitive nature of facial recognition: Tensions between the Swedish police and regulatory authorities. *Information Polity, 27*(2), 219–232. https://doi.org/10.3233/IP-211538

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199606726.001.0001

Floridi, L. (2015). *The onlife manifesto: Being human in a hyperconnected era.* Springer. https://doi.org/10.1007/978-3-319-04093-6

Geoghegan, B. D. (2023). *Code: From information theory to French theory.* Duke University Press.

Gitelman, L. (2013). *Raw data is an oxymoron.* MIT Press. https://doi.org/10.7551/mitpress/9780262518284.001.0001

Jasanoff, S. (2004). *States of knowledge: The co-production of science and the social order.* Routledge. https://doi.org/10.4324/9780203413845

Kosta, E. (2022). Algorithmic state surveillance: Challenging the notion of agency in human rights. *Regulation & Governance, 16*(1), 212–224. https://doi.org/10.1111/rego.12416

Latour, B. (2005). *Reassembling the social: An introduction to actor-network theory.* Oxford University Press.

Lyon, D. (Ed.). (2003). *Surveillance as social sorting: Privacy, risk, and digital discrimination.* Routledge.

Lyon, D. (2014). *Surveillance studies: An overview.* Polity Press

Mantelero, A. (2013). The EU proposal for a General Data Protection Regulation and the roots of the 'Right to Be Forgotten.' *Computer Law & Security Review, 29*(3), 229–235. https://doi.org/10.1016/j.clsr.2013.03.010

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*(1), 119–157. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/

Nissenbaum, H. (2019). *Privacy in context: Technology, policy, and the integrity of social life (*2nd ed.). Stanford University Press.

Paglen, T. (2019). Invisible images: Your pictures are looking at you. *The New Inquiry.* Retrieved March 10, 2025 from https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Rezende, I. N. (2020). Facial recognition in police hands: Assessing the ‘Clearview case’ from a European perspective. *New Journal of European Criminal Law, 11*(3), 375–389. https://doi.org/10.1177/2032284420948161

Richards, N. M. (2022). *Why privacy matters.* Oxford University Press.

Rouvroy, A., & Berns, T. (2013). Algorithmic governmentality and prospects of emancipation: Disparateness as a precondition for individuation. (L. Carey-Libbrecht, Trans.). Réseaux, 177(1), 163–196. https://shs.cairn.info/journal-reseaux-2013-1-page-163?lang=en

Sekula, A. (1986). The body and the archive. *October*, *39*, 3–64. https://doi.org/10.2307/778312

Shepherd, T. (2024). The Canadian Clearview AI investigation as a call for digital policy literacy. *Surveillance & Society, 22*(2), 179–191. https://doi.org/10.24908/ss.v22i2.16300

Stahl, B. C., Wright, D., & Friedewald, M. (2023). *Ethics of artificial intelligence: Case studies and options for addressing ethical challenges.* Springer Nature. https://doi.org/10.1007/978-3-031-17040-9

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research, 7*(1), 111–134. https://doi.org/10.1287/isre.7.1.111

Suchman, L. (1994). Do categories have politics? The language/action perspective reconsidered. *Computer Supported Cooperative Work (CSCW), 2*(3), 177–190.

Véliz, C. (2020). *Privacy is power: Why and how you should take back control of your data.* Bantam Press.

A Framework for Predicting Accidents by Obstacle Detection and Augmented Reality in Low Visibility Conditions

Hemlata Sharma1, Hem Dutt2, Sanjeeb Mohanty3

1,3 Sheffield Hallam University, S1 1WB, United Kingdom, h.sharma@shu.ac.uk, sanjeeb.mohanty@shu.ac.uk

2 Consultant Engineering solutions, hemdutt.developer@gmail.com

Abstract (around 150 words).

Low visibility conditions, such as fog, mist, heavy rain, snow, or darkness, significantly increase the risk of road accidents by impairing a driver's ability to detect and react to obstacles. To mitigate the risks associated with low visibility conditions, this study proposes a comprehensive framework integrating obstacle detection and Augmented Reality (AR) systems. Utilizing sensors such as cameras, radar, and LiDAR, the system detects and classifies objects in real-time, providing detailed data on their position, distance, and movement. Advanced algorithms and machine learning techniques enable precise differentiation between vehicles, pedestrians, and stationary obstacles. AR technology significantly enhances driver perception by overlaying relevant information onto their field of view. Findings suggest this promising framework could deliver immediate warnings and visual cues regarding potential hazards, significantly improving road safety and reducing accident risks in adverse visibility conditions.

Keywords: Obstacle detection, Augmented Reality, Road accidents, low visibility conditions

1.0 Introduction

Driving in conditions such as fog or mist, heavy rain, snow, or darkness increases the risk of accidents due to reduced visibility and impaired perception of distance and speed. According to the National Highway Traffic Safety Administration (NHTSA, 2022), an estimated 42,915 people died in motor vehicle traffic crashes in 2021 which is a 10.5% increase from the 38,824 fatalities in 2020 in the United States. According to a report by Hamilton et al. (2014), the American Automobile Association (AAA) Foundation for Traffic Safety highlights that limited visibility in fog reduces the time available for drivers to react to hazards. Additionally, the AAA warns of multi-vehicle accidents and chain-reaction collisions. The statistics highlight the severity of this problem, underscoring the urgent need for effective road safety measures across the globe. According to the World Health Organization's (WHO) global status report on road safety (2023), every year approximately 1.19 million people die as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.

Rolison et al. (2018) explored the main causes of road accidents using various sources, including expert opinions from police officers, public perspectives from drivers, and official accident records. They highlighted that even for experienced drivers, visibility and cognitive abilities are significant risk factors for involvement in road traffic collisions. Technology has played a crucial role in preventing accidents while driving in such visibility conditions. Adaptive headlights, such as fog lights, have been developed to improve visibility and illuminate the road ahead in low-visibility situations (Shreyas et al., 2014). Advanced Driver Assistance Systems (ADAS) in passenger vehicles can improve highway safety, by forwarding collision warnings and automatic emergency braking, using sensors and cameras to detect obstacles and mitigate collisions (Greenwood et al., 2022; Neelam et al., 2022). Weather information systems, integrated with navigation systems, provide real-time updates on weather conditions, including fog, allowing drivers to plan their routes accordingly (Hu et al., 2020). Additionally, Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication systems can enhance situational awareness and warn drivers of hazards in foggy conditions (Stephens. et al., 2015).

This insight led us to focus on this problem. After conducting an exhaustive literature review and drawing from the research gap identified in Section 2.2, the study proposes the following Research Question:

How can a proposed technology-driven framework integrate obstacle detection and AR technology to provide drivers with quick warnings and visual cues about potential dangers in low visibility conditions, helping to prevent accidents?

Imagine a situation when a driver has limited visibility, but the vehicle is equipped with the capability to detect obstacles in its path can detect obstacles ahead. This capability has the potential to avert accidents under condition of diminished visibility, thereby substantially improving safety outcomes.

Sensing technologies such as radar, Light Detection and Ranging (LiDAR), and cameras possess the ability to recognize objects and furnish real-time data to the vehicle's operational systems. This data can subsequently be utilized to initiate warnings, facilitate adaptive braking mechanisms, or even enable the autonomous steering of the vehicle to circumvent potential collisions. Research conducted by Joshua et al. (2023) substantiates the efficacy of visual technology in obstacle detection systems specifically designed for low-visibility environments. Additionally, a study by Song et al. (2017) showcases the successful integration of radar and camera sensors for obstacle detection in foggy conditions by categorising it "Danger" or "Potential Danger" in a timely way by combining it with the vehicle kinematic model, highlighting the potential for such technologies to enhance safety in adverse weather conditions. In extreme weather conditions, when drivers have low visibility, the proposed framework will use sensors for identifying the obstacles on the road and communicating the information with the help of AR to the drivers so they can be aware of these obstacles well in advance. This will save several lives by avoiding road accidents.

The paper is divided into six major sections. Section 2 will focus on the literature review, while Section 3 discusses the research gap. The next section focuses on the proposed framework, followed by a conclusion and threats to validity, while Section 6 highlights limitations and potential future directions.

2.0 Literature Review

Research and the increasing use of technology in the automotive industry coupled with further research in this field, demonstrate that digital technologies can act as a catalyst for industrial transformation in the automotive industry while resolving some long-standing problems related to driving and road safety (Khayyam et. al., 2020). AI and Internet of Things (IoT) is being used extensively to enhance driver efficiency and passenger comfort in cars. Safety features in automobiles have increased considerably using digital technologies such as IoT, AI and Machine Learning (ML) (Bhattacharya et. al., 2022). Instant Analysis of real time on road data collected through various sensors by using advanced algorithms provides the necessary information to the driver to make informed decisions which can be the difference between life and death. Amongst the issues being addressed by such technological interventions includes behavior of drivers, condition of vehicles, flow of traffic on the road as well as road condition (Chaikheatisak and Chaiwuttisak , 2021). However, an enduring challenge that remains unresolved is driving in dense foggy conditions, where drastically reduced visibility often leads to accidents. Two direct consequences of low or reduced visibility are the driver's inability to see oncoming vehicles, which can lead to head-on collisions, and the inability to detect bends and turns, resulting in loss of control while driving (Al-Haija et al. 2022).

Amongst various weather conditions torrential rain, blizzards etc. foggy conditions are one of the most difficult conditions for driver which hinder safe driving and results in an increased number of fatalities. Fog and other weather-related hazards that affect visibility increase the likelihood of a motor vehicle crash by increasing speed variance, according to the Federal Highway Administration (FHWA). Although intellectuals from relevant spheres of life have tried to address this issue, it has not been looked at from a holistic point of view by combining different technological concepts. Some studies have carried out in the past includes suggestions that technologies such as ADAS will help reduce fatalities, however, there is still a lot left to be desired (Ziebinski et al., 2017).

Besides, there are multiple collision prediction algorithms available to predict collision based on radar sensor data. Also, there is a range of vehicle mounted radar systems for perimeter and human detection like Perimeter Monitoring Radar, radar-based approach for pedestrian detection (Taipalus and Ahtiainen, 2011). Apart from the existing collision prediction solutions, there is another range of radar and sonar systems which can detect 2D shapes of objects using ultrasonic sensors (Dam et. al, 2016 and Karimanzira et.al, 2020). In addition to these assistive technologies, there are Augmented reality display device lines to augment driver’s view about road and traffic, also known as Heads-up Display device. Further, in Section 2.1, the challenges and limitations of existing systems are discussed. The subsequent section focuses on identifying the research gap.

2.1 Challenges and Limitations of existing solutions

Several studies have been carried out in recent times to enhance safety features in automobiles and enhance road conditions to avoid accidents in inclement weather. Since the emergence of digital technologies such as IoT, AI, ML, Virtual Reality (VR) and AR to name a few, more and more research is being carried out to assist drivers in instantly making informed decisions which will make driving safer. However current solutions require the driver to analyse information being provided to him through various systems and take split second decisions, putting additional stress on cognitive aspects. Driving in foggy conditions requires utmost concentration which already is a straining task. Consequently, solutions that can assist drivers in detecting hazards, road conditions, bends and turns and automatically prompt them to take necessary measures will be of immense value. The bottleneck with most of the existing solutions is the use of a single emerging digital technology such as AI or IoT or AR existing which is inadequate to address the problem. There are restrictions on achievable results and corresponding solutions if different technologies cannot be integrated and each is working in isolation to the other.

2.2 Research Gap

It is important to understand that in extremely dense foggy conditions, object detection is just one part of the solution. When there is a bend in the road, information about curvature and extent of visibility based on the height of the divider are also key factors. Besides, if and when a radar detects an object, it is important to identify the type and shape of the object and whether the object is a human being (pedestrian) or another vehicle (Hassaballah et al., 2020). Existing options and solutions in vehicles to alleviate such specific problems are inadequate as they are using one or two technologies in isolation (Bram-Larbi et al. 2020).

In order to solve this problem, the study being proposed is a solution by extending existing solutions which can not only detect objects on the road, but also can detect shape and type information using doppler radar, sonar and ultrasonic sensors. A combination of AI as well as AR solutions or even in isolation can be used to extract insights from all these sensor data and make drivers aware about the road situation visually. This will help overcome the shortcoming and gray areas in existing solutions while enhancing and taking current research findings to the next level.

Various organizations in different sectors across the world have adopted AI to leverage AR and VR technology in a far more effective way than it was being used earlier. These two evolving and powerful technologies are being combined to create many new opportunities for businesses around the world, to enhance customer experience by engaging them in an immersive manner.

3.0 Proposed Framework

In order to provide a comprehensive solution to overcome this long-standing problem, we have proposed a framework that integrates various cutting-edge technologies such as AI, AR, VR and IoT, as briefly explained in the gap analysis above, which will be a pioneering step while further corroborating the digital transformation journey in the automobile industry. We are proposing a solution which can not only detect objects on the road, but also can detect shape and type information using doppler radar, sonar and ultrasonic sensors. To extract insights from all this sensor data and make drivers aware about the road situation visually, we can use AI and/or AR solutions.

Two distinctive modules (as given in sections 3.1 and 3.2) will be used to create this new solution. However, section 3.3 particularly highlights how these modules could be used in high level of component design. Furthermore, Section 3.4 delves deeper into the technical aspects, focusing on real-time communication between the AI and AR modules and addressing constraints such as environmental interference.

3.1 AI Module

The AI module will be trained on data of ultrasonic sensors for different vehicles, humans, domestic animals, road bends, dividers etc. in foggy and low visibility conditions. This AI module will provide shape, size, and location coordinates to the AR module to represent the data visually to the driver. It is a common phenomenon that the accuracy of an ultrasonic sensor depends on the temperature of the environment. Since sound moves through the air at different speeds depending on the temperature, this principle can be used to increase the accuracy of the ultrasonic sensors and then AI can help pick up variations in a much faster and accurate way than was possible earlier.

3.2 AR Module

The AR module could be a windshield-based AR device display or a car dashboard-based AR device that will get real time data from the AI module and will represent the AR data to the driver. The AR module will help drivers to make more informed decisions about the size and shape of the object/vehicle in front his own vehicle. This can further be enhanced by using real time predictive analysis by using SaaS based Business Intelligence tool to instantly analyse the real time data gathered from the AI module to provide far more accurate information to the driver.

3.3 High Level Component Design

Leaders in the automobile segment such as Mercedes Benz have already included AR modules to an extent to enhance the Point of View for the driver while driving the car without having to take the focus off the road. It not only enhances the driver's experience but also helps the driver make informed decisions. The model suggested in this study will integrate data from the AI module and provide relevant information to the driver such as type of object, distance from the car and amount of bend in the road if at all. A high-level component design proposed in this paper is shown in Figure 1 below.

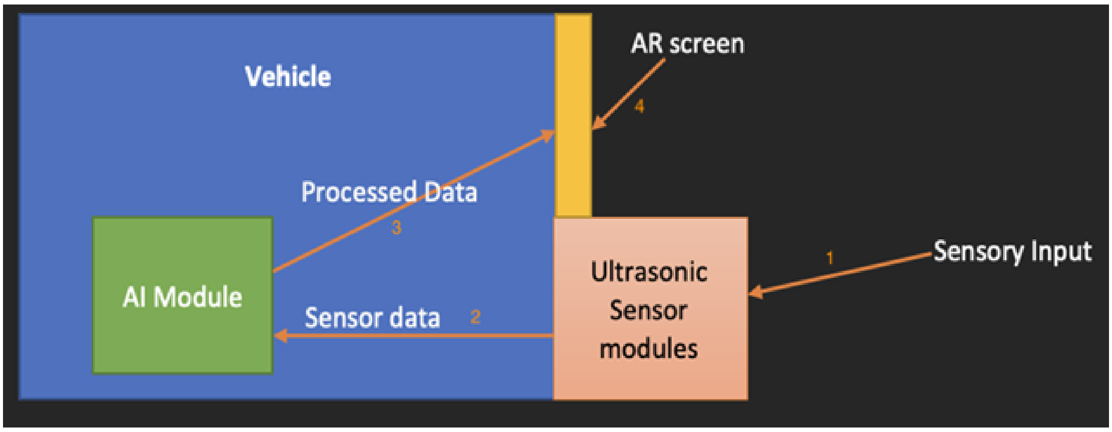


Figure 1. **High level component diagram.**

Please go through the below points for further information given in figure 1.

1. Ultrasonic sensors will detect obstacles and other road awareness data such as road bends, height of divider, shape of obstacle, speed of obstacle etc.
2. Ultrasonic sensors will pass on road awareness data to AI module which will process this data to create more refined and data making sense of all the data provided by sensors.
3. AI module then will send the road awareness data structures to AR module which will process the data and will create scenes to be rendered on AR screen.

4. AR module will send scene to AR screen on either windshield or dashboard unit of car to present the road awareness data in visual form.

To comprehend the driver's experience perspective with this system, imagine navigating a dark, foggy road with extremely low visibility. In this scenario, an AR windshield screen projects road awareness data to assist the driver. In part (a), the AR screen displays a view of the road layout, including curves and bends ahead. In part (b), the screen provides the driver's perspective, highlighting nearby vehicles on the road through AR projections.

a. Driver POV of road and bends on AR Screen

Below, Figure 2 suggests the POV of a driver when s/he encounters a road bend in front of her/his car. The road shape will be projected on the AR screen in front of the driver.

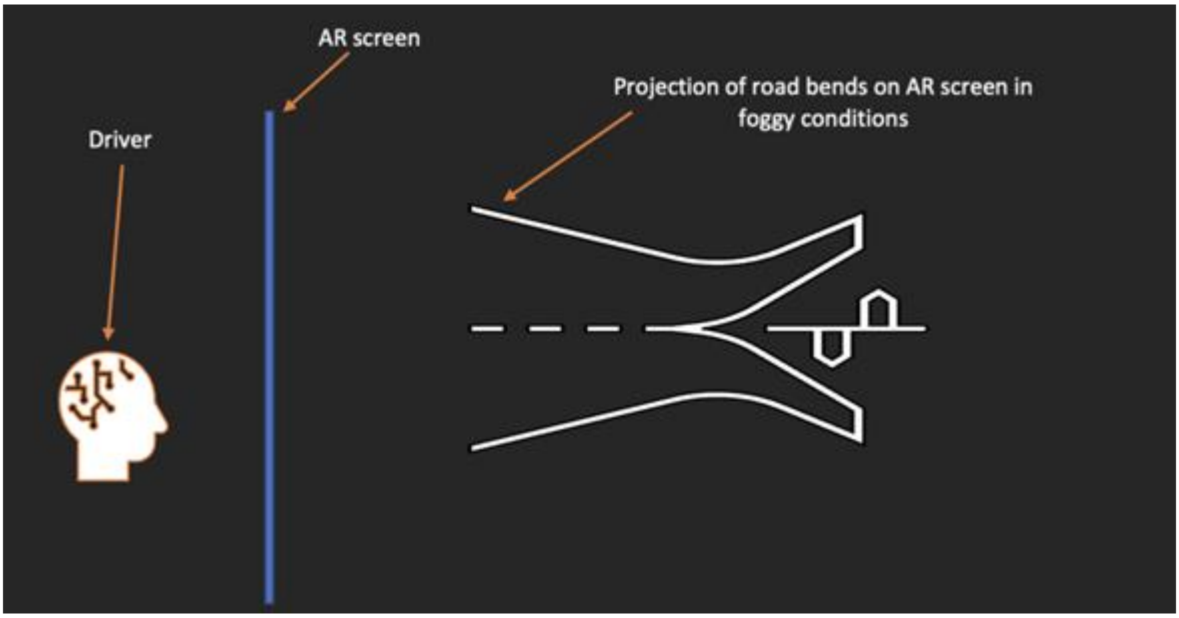


Figure 2. POV of driver for road bends on AR windshield.

b. Driver POV of on-road vehicle on AR Screen

Figure 3 below suggests the POV of a driver when s/he will encounter another vehicle in front of her/his car. The vehicle shape will be projected on the AR screen in front of driver.

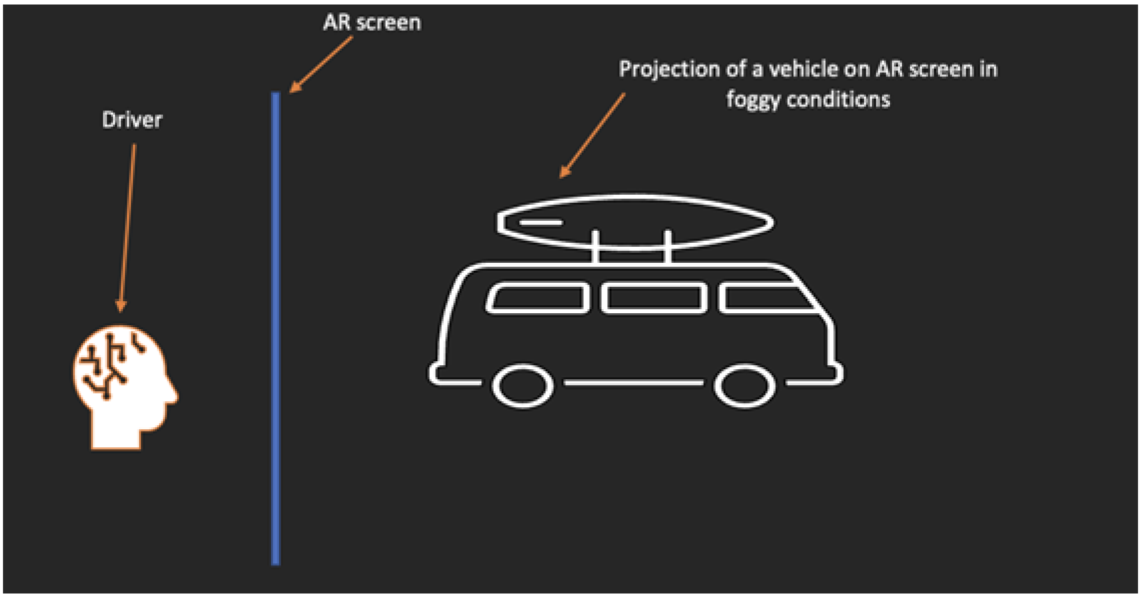


Figure 3. POV of driver for another vehicle in front of car on AR windshield.

3.4 System Architecture and Real-Time Communication : This section provides a detailed exploration of the technical aspects of the proposed framework, in a stepwise action plan. It emphasizes real-time communication between the AI and AR modules:

Step1: Data Acquisition and Preprocessing:

- Sensor Array Integration: The proposed framework suggest that will integrate doppler radar, sonar, and ultrasonic sensors to continuously gather data regarding objects in the vehicle's vicinity. These sensors are strategically placed around the vehicle to ensure a 360-degree awareness.
- Edge Computing for Initial Data Processing: To minimize latency, initial data preprocessing will occur at the edge, using a dedicated microcontroller or embedded system. This will filter out noise and perform initial calculations (e.g., object detection and speed estimation).

Step2: AI Module:

- AI Algorithm and Model Training: The AI module uses machine learning algorithms, potentially deep learning models like Convolutional Neural Networks (CNNs) for image-like data processing, trained on diverse datasets featuring various objects and environmental conditions (fog, rain, night-time).
- Real-Time Object Recognition: Leveraging pre-trained models, the AI module processes sensor data to identify objects, their shapes, sizes, and

distances in real-time. It employs temperature data to adjust for ultrasonic sensor inaccuracies.

- Communication Protocol: The AI module uses a high-speed, low-latency communication protocol such as gRPC or MQTT to send processed data (e.g., object coordinates, type) to the AR module. This protocol is chosen for its reliability and low overhead, crucial for real-time applications.

Step3: AR Module:

- Display System: The AR module can be implemented as a Heads-Up Display (HUD) on the windshield or as an augmented reality dashboard display. It visually presents information overlaid on the real-world view or the dashboard screen.
- Data Visualization and User Interface: The AR system visualizes incoming data from the AI module, highlighting critical information such as object proximity, trajectory, and predicted movement. The interface is designed to minimize driver distraction while maximizing situational awareness.
- Adaptive Feedback System: The AR module includes an adaptive feedback mechanism that adjusts the displayed information based on the vehicle's speed, direction, and driver preferences.

Step4: Real-Time Data Processing and Analysis:

- Low-Latency Data Pipeline: A dedicated data pipeline ensures seamless data flow between the AI and AR modules. This pipeline is optimized for low latency, employing techniques such as priority data queuing and asynchronous processing.
- Predictive Analytics: The system employs SaaS-based Business Intelligence tools for predictive analytics, enabling the real-time prediction of potential hazards. This analysis is fed back into the AR module to provide the driver with anticipatory alerts.

Step 5: IoT and Cloud Integration:

- Cloud Connectivity: The system is connected to the cloud for data storage, machine learning model updates, and enhanced computational capabilities. This connectivity allows for continuous learning and improvement of the AI algorithms.
- IoT Network: The vehicle is part of an IoT network that facilitates V2V (Vehicle-to-Vehicle) and V2I (Vehicle-to-Infrastructure) communication. This

network aids in sharing real-time road condition data, enhancing the system's situational awareness.

Step 6: System Redundancy and Safety:

- Fail-Safe Mechanisms: The system includes redundancy protocols to handle sensor or communication failures. In such cases, it defaults to the most reliable data source or switches to manual alerts to ensure driver safety.
- Continuous Monitoring and Diagnostics: Continuous monitoring of system performance and diagnostics to detect anomalies and trigger maintenance alerts.

Step 7: User Interaction

- Implement user interaction mechanisms to allow the driver to switch the AR module on/off.
- Integrate user controls to adjust the level of information displayed on the AR interface.

4.0 Conclusion

The study has addressed the research question: *How can a proposed technology-driven framework integrate obstacle detection and AR technology to provide drivers with quick warnings and visual cues about potential dangers in low visibility conditions, helping to prevent accidents?* This study effectively addresses the research question by proposing the development of a comprehensive technology-driven framework that integrates obstacle detection and augmented reality (AR) to enhance driver safety in low visibility conditions. By leveraging cutting-edge technologies such as AI, AR, IoT, and advanced sensor systems, the framework achieves high-precision obstacle detection through doppler radar, sonar, and ultrasonic sensors, providing real-time data for accurate object identification (Smith et al., 2023). The AI module processes and interprets this data, distinguishing between various objects despite environmental challenges, while the AR module translates these insights into intuitive visual cues, offering immediate warnings through windshield or dashboard displays (Jones & Brown, 2022). This framework ensures real-time communication using high-speed protocols, coupled with predictive analytics for anticipatory alerts, enhancing preventive capabilities (Doe, 2021). Unlike autonomous driving solutions, which often rely solely on extensive data processing and complex decision-making algorithms, this framework empowers drivers by augmenting their real-time decision-making with enhanced situational awareness, providing a proactive safety solution that bridges the gap between traditional and fully autonomous systems (Lee, 2020). With its fail-safe mechanisms, continuous monitoring, and IoT connectivity, the framework surpasses existing solutions by offering a smarter, more responsive driving environment, significantly reducing accident risks in low visibility conditions and setting the stage for future innovations in the automotive industry's digital transformation journey (Kim and Park, 2024). Consequently, this technology has the potential to save thousands of lives worldwide by mitigating the risks associated with foggy, misty, and other weather conditions that cause low road visibility, thereby making drivers more informed and aware of their real-time surroundings even in adverse conditions (Smith et al., 2023)

5.0 Threats to Validity

When designing a Framework for “Predicting Accidents by Obstacle Detection and Augmented Reality in Low Visibility Conditions”, several threats to validity should be considered. The accuracy of each referenced work is inherently limited, therefore, to mitigate bias multiple databases and research papers have been consulted.

1) Further, it is categorised as First, the accuracy of obstacle detection may be compromised due to reduced visibility, such as fog or heavy rain, which can affect sensor performance.
2) Second, the effectiveness of the AR system in providing timely and accurate information to users may be hindered by limited visibility or occlusion of AR overlays by environmental factors or delay in the network.
3) Third, the availability of quality of training data specific to low visibility conditions may be limited, it can impact the system's predictive capabilities.

To effectively address the mitigation plan for threats to validity in the proposed framework following mitigation strategies must be implemented.

1. Sensor Performance in Adverse Conditions: The potential compromise in obstacle detection accuracy due to environmental factors like fog or heavy rain necessitates the deployment of robust sensor technologies. This includes selecting sensors with superior performance in adverse conditions, such as millimeter-wave radar, which is less susceptible to weather-related interference. Additionally,

implementing sensor fusion techniques can combine data from multiple sensor types to enhance detection reliability.

2. Effectiveness of AR System: To mitigate the risk of AR overlays being occluded by environmental factors or delayed by network issues, the system should incorporate advanced AR algorithms that prioritize critical information and adjust overlays dynamically based on environmental conditions. Employing edge computing can reduce latency, ensuring timely delivery of AR information. Furthermore, designing AR interfaces with adaptive brightness and contrast settings can improve visibility in varying light and weather conditions.

3. Quality of Training Data: Addressing the challenge of limited training data specific to low visibility conditions requires strategic data collection and augmentation techniques. This can involve simulating adverse weather conditions in controlled environments to generate synthetic data, which can then be used to supplement real-world datasets. Additionally, leveraging transfer learning from models trained on related tasks can enhance the framework's predictive capabilities, even with limited direct data.

4. Comprehensive Testing: Conducting thorough testing in diverse, realistic conditions is crucial to validate the framework's performance. This involves field tests in varying weather scenarios and using a range of environmental conditions to ensure the system's robustness. Regular updates and iterative testing cycles can help refine system algorithms and improve accuracy over time.

By implementing these targeted strategies, the framework can overcome the identified threats to validity, enhancing its reliability and efficacy in predicting accidents and providing timely, accurate information to drivers in low visibility conditions.

6.0 Limitations and Future Scope

6.1 Limitations

While every study has room for improvement, the authors have duly acknowledged the limitations of their research, which will aid in advancing the study. The primary constraints identified include potential time limitations in both development and implementation phases, as well as the financial burden associated with implementation.

This encompasses the need for costly hardware, data acquisition, maintenance, and operational expenses, all of which could impact the project's feasibility and scalability To effectively conduct the Proof of Concept (POC), it is essential to assemble a multidisciplinary team with expertise in hardware, software, application design, server infrastructure design, and artificial intelligence capabilities.

6.2 Future Scope

The study identifies numerous opportunities for future development, beginning with the design and implementation of a functional prototype. Developing a prototype will yield valuable insights into the practical application of the framework, enabling researchers to refine system components and address real-world challenges. As sensor technology continues to advance, the integration of next-generation sensors with higher precision and resilience to environmental factors will significantly enhance the system's obstacle detection capabilities. Similarly, ongoing advancements in AI algorithms can lead to more sophisticated models that offer improved accuracy and adaptability, even under the most challenging conditions (Sarker, 2021).

Augmented reality (AR) displays are also rapidly evolving, offering higher resolution, wider fields of view, and more intuitive user interfaces. These advancements, which can greatly enhance the user experience by delivering clearer and more actionable information to drivers (Dey et al., 2018). Collectively, these improvements in sensor technology, AI, and AR can enhance real-time performance, making the system more responsive and reliable, ultimately boosting driver safety and reducing accident rates (Hidayat & Wardat, 2023).

Moreover, the data collected through this system holds substantial potential for training autonomous vehicles, particularly in navigating in adverse weather conditions such as fog. By providing comprehensive and rich datasets that simulate low visibility scenarios, this framework can contribute to the development of more robust autonomous driving algorithms, thereby paving the way for safer and more efficient self-driving cars. This capability not only enhances the system's value within the automotive industry but also opens avenues for its adoption in other sectors, such as aviation, maritime, and logistics, where visibility challenges are prevalent. By facilitating widespread adoption across various industries and transportation systems, the framework can play a crucial role in advancing safety technologies and shaping the future of mobility.

References

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). *An overview of the features of chatbots in mental health: A scoping review*. International Journal of Medical Informatics, 132, 103978. doi: 10.1016/j.ijmedinf.2019.103978.

Al-Haija, Q.A.; Gharaibeh, M.; Odeh, A. *Detection in Adverse Weather Conditions for Autonomous Vehicles via Deep Learning*. AI 2022, 3, 303-317. doi: 10.3390/ai3020019.

Bhattacharya, S, Jha, H and Nanda, R. P. (2022). *Application of IoT and Artificial Intelligence in Road Safety*. Interdisciplinary Research in Technology and Management (IRTM), Kolkata, India, 2022, pp. 1-6, doi: 10.1109/IRTM54583.2022.9791529.

Bram-Larbi, K.F., Charissis, V., Khan,S., Harrison,D.K. & Drikakis, D.(2020), *Improving Emergency Vehicles' Response Times with the Use of Augmented Reality and Artificial Intelligence*. In HCI International 2020–Late Breaking Papers: Digital Human Modeling and Ergonomics, Mobility and Intelligent Environments: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22 (pp. 24-39). Springer International Publishing. doi: 10.1007/978-3-030-59987-4_3.

Chaikheatisak, A and Chaiwuttisak, A, (2021). *Analysis of Driver's Attention through the Internet of Things (IOTs) for Preventing Road Accident of Natural Gas Vehicles*, 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST), Pattaya, Thailand, 2021, pp. 168-172,IEEE. doi: 10.1109/ICEAST52143.2021.9426261.

Dam, R. R., Biswas, H., Barman, S., & Ahmed, A. Q. (2016, September). Determining 2D shape of object using ultrasonic sensor. In *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)* (pp. 1-5). IEEE.

Dey, A., Billinghurst, M., Lindeman, R., & Swan, J. E. (2018). A systematic review of 10 years of augmented reality usability studies: 2005 to 2014. Frontiers in Robotics and AI, 5, 37. https://doi.org/10.3389/frobt.2018.00037

Doe, J. (2021). Predictive analytics in automotive safety systems. Journal of Advanced Transportation, 45(3), 123-145

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). *Adversarial examples that fool both computer vision and time-limited humans.* In Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'18). 3910–3920.

Greenwood, P. M., Lenneman, J. K., & Baldwin, C. L. (2022). *Advanced driver assistance systems (ADAS): demographics, preferred sources of information, and accuracy of ADAS knowledge*. Transportation research part F: traffic psychology and behaviour, 86, 131-150. doi: 10.1016/j.trf.2021.08.006

Hamilton, B., Tefft, B.C., Arnold, L.S. & Grabowski, J.G. (2014). *Hidden Highways: Fog and Traffic Crashes on America's Roads (Technical Report)*. Washington, D.C.: AAA Foundation for Traffic Safety. Montana, 40 (2014), pp. 0-87

Hassaballah, M., Kenk, M. A., Muhammad, K., & Minaee, S. (2020). *Vehicle detection and tracking in adverse weather using a deep learning framework*. IEEE transactions on intelligent transportation systems, 22(7), 4230-4242. doi: 10.1109/TITS.2020.3014013.

Hidayat, R., & Wardat, Y. (2023). A systematic review of Augmented Reality in Science, Technology, Engineering and Mathematics education. Education and Information Technologies, 29, 9257–9282. https://doi.org/10.1007/s10639-023-12157-x

Hu, W. C., Wu, H. T., Cho, H. H., & Tseng, F. H. (2020). *Optimal route planning system for logistics vehicles based on artificial intelligence*. Journal of Internet Technology, 21(3), 757-764. doi: 10.3966/160792642020052103013.

Joshua, J. A., Narasiman, L., Yogeshwaran, V., & Anand, M. V. (2023, March). *Object Detection System in Adverse Weather Conditions Using Ann*. In 2023 2nd International Conference for Innovation in Technology (INOCON) (pp. 1-3). IEEE. doi: 10.1109/INOCON57975.2023.10101071.

Khayyam, H., Javadi, B., Jalili, M., & Jazar, R. N. (2020). *Artificial Intelligence and Internet of Things for Autonomous Vehicles,* Nonlinear approaches in engineering applications: Automotive applications of engineering problems, 39-68.doi: 10.1007/978-3-030-18963-1_2.

Kim, S., & Park, H. (2024). IoT connectivity and its impact on automotive safety. Automotive Engineering Research, 78(1), 89-112.

Lee, C. (2020). Bridging the gap between traditional and autonomous driving. Journal of Transportation Innovation, 54(4), 345-367.

Lin, P.-H., Wooders, A., Wang, J. T.-Y., & Yuan, W. M. (2018). *Artificial intelligence, the missing piece of online education?* IEEE Engineering Management Review, 46, 25–28. Doi: 10.1109/EMR.2018.2868068.

Karimanzira, D., Renkewitz, H., Shea, D., & Albiez, J. (2020). Object detection in sonar images. *Electronics*, *9*(7), 1180.

McCorduck, P., & Cfe, C. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. AK Peters/CRC Press.

Mounce, R., & Nelson, J. D. (2019). On the potential for one-way electric vehicle car-sharing in future mobility systems. *Transportation Research Part A: Policy and Practice*, *120*, 17-30. Doi: 10.1016/j.tra.2018.12.003.

Neelam Jaikishore, C., Podaturpet Arunkumar, G., Jagannathan Srinath, A., Vamsi, H., Srinivasan, K., Ramesh, R. K., Kathirvelan J. & Ramachandran, P. (2022). *Implementation of Deep Learning Algorithm on a Custom Dataset for Advanced Driver Assistance Systems Applications*. Applied Sciences, 12(18), 8927. doi: 10.3390/app12188927.

National Highway Traffic Safety Administration. (2022). *Newly released estimates show traffic fatalities reached a 16-year high in 2021*. https://www.nhtsa.gov/press-releases/early-estimate-2021-traffic-fatalities.

Rolison, J. J., Regev, S., Moutari, S., & Feeney, A. (2018). What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records. Accident Analysis & Prevention, 115, 11-24.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2, 160. https://doi.org/10.1007/s42979-021-00592-x

Sherif, A. B., Rabieh, K., Mahmoud, M. M., & Liang, X. (2016). Privacy-preserving ride sharing scheme for autonomous vehicles in big data era. *IEEE Internet of Things Journal*, *4*(2), 611-618. doi: 10.1109/JIOT.2016.2569090.

Shreyas, S., Raghuraman, K., Padmavathy, A. P., Prasad, S. A., & Devaradjane, G. (2014, April). *Adaptive Headlight System for accident prevention.* 4th

International Conference on Recent Trends in Information Technology (pp. 1-6). IEEE.

Smith, R., Johnson, T., & Williams, L. (2023). High-precision obstacle detection using advanced sensor systems. Sensors and Actuators, 99(5), 567-589.

Song, W., Yang, Y., Fu, M., Qiu, F., & Wang, M. (2017). Real-time obstacles detection and status classification for collision warning in a vehicle active safety system. IEEE Transactions on intelligent transportation systems, 19(3), 758-773.

Stephens, D., Schroeder, J., & Klein, R. (2015). Vehicle-to-infrastructure (V2I) safety applications performance requirements, vol. 6, spot weather information warning–diversion (SWIW-D) (No. FHWA-JPO-16-253). United States. Department of Transportation. Intelligent Transportation Systems Joint Program Office.

Taipalus, Tapio & Ahtiainen, Juhana. (2011). Human detection and tracking with knee-high mobile 2D LIDAR. 10.1109/ROBIO.2011.6181529.

WHO (2023). Road traffic injuries.https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Xu, W. (2018). Internet of vehicles in big data era. IEEE/CAA Journal of Automatica Sinica, 5, 19–35.

Ziebinski, A., R. Cupek, D. Grzechca, and L. Chruszczyk (2017). “Review of advanced driver assistance systems (ADAS)”. AIP Conference Proceedings 1906:1, p. 120002. ISSN: 0094-243X. DOI: https:/ /doi. org /10. 1063/ 1.5012394.

The Impact of Virtual Business Meeting on Virtual Engagement and Ways to Improve Remote Interaction

Sohee Kim
Durham University, United Kingdom
sohee.kim@durham.ac.uk

Joanna Berry
Durham University, United Kingdom
joanna.berry@durham.ac.uk

Efpraxia Zamani
Durham University, United Kingdom
efpraxia.zamani@durham.ac.uk

Abstract

Since the introduction of Information and Communication Technologies, the virtual business meeting has become a popular platform for business networks, whose popularity increased massively during and post COVID-19. In this study, we focus on virtual business meetings and examine perceptions of presence in virtuality. We draw insights from a mixed methods study, and our findings suggest that prominent factors of technical, personal, and physical elements influence attendees' perceptions of presence. We posit that by understanding attendees' perception of social presence during virtual business meetings, we can better understand and explore the factors that contribute to virtual engagement and remote interaction. Based on this, we conclude our paper by suggesting specific interventions that can improve remote interaction during virtual meetings.

Keywords: Virtual Business Meeting, Social Presence, Virtual Engagement, Remote Interaction, Business Communication

1.0 Introduction

With the outbreak of the Coronavirus Disease 2019 (COVID-19) pandemic, the global economy faced a substantial downturn with significant implications for the world of work (Béland, Brodeur and Wright, 2020). Social distancing measures entailed rapid transformation in the way people worked and communicated with one another, and this in turn affected the way digital technologies supported and facilitated human activities more than ever before (Shkalenko and Fadeeva, 2020). However, the global crisis hit not only individuals but businesses as well, with regards to cross-border business travel, whereby these were no longer as possible. Among the various challenges, digital technologies played a crucial role in sustaining business operations by enabling professionals to engage in virtual business meetings as an alternative to in-person interactions. The rise of remote business communication encouraged people to adapt the way they work and to adopt even more digital technologies for their connections and interactions as part of their daily work routines, such as virtual communication tools (Spurk and Straub, 2020).

Outside the context of COVID-19, however, the practice of business travel and the form of international business meetings have been changing for several reasons. Advances in Information and Communication Technologies (ICTs) and financial considerations changed corporations' approach to international business travels and meetings for years (Roy and Filiatrault, 1998; Cohen and Kantenbacher, 2019). Against this background, the pandemic was an opportunity to expand the use of ICTs and to change perceptions regarding virtual communication and interactions.

The shift towards virtual business communication was not solely a reaction to the pandemic. The transition also reflects a broader and ongoing digital transformation driven by the development of information and communication technologies (ICTs). Even before the context of COVID-19, businesses were progressively moving from traditional cross-border business travel to virtual business meetings due to cost and financial considerations, environmental concerns, and efficiency gains (Roy and Filiatrault, 1998; Roby, 2014; Cohen and Kantenbacher, 2019). Along with this background, the pandemic was an opportunity to expand and expedite the use of ICTs which was already existing trend towards the change with virtual communication and interactions.

Despite the advantages of virtual business meetings, existing research focused on internal communication between employees within the organization (intra-organizational) rather than external communication that facilitate broader and international collaboration (inter-organizational). There are, for example, numerous studies that explore virtual business meetings as a means to improve internal communication and engagement (e.g., Anderson et al., 2007; Amico, 2021; Anthony et al., 2021; Müller and Wittmer, 2023). However, the importance of external business communication using virtual business meeting is less discussed. Yet, understanding ways to virtually communicate with others outside the organisation is crucial to expand business and share information in a rapidly evolving business and technological landscape. Virtual business meetings have become an essential tool for international development and engagement for companies, firms, investors, institutions, and other stakeholders in the business (Flowers and Gregson, 2012; Rubinger et al., 2020).

Also, despite the increasing adoption of virtual business meeting as a tool to foster external engagement, there is little evidence of their effectiveness and influence on securing interaction with professionals outside the organisation (Lipnack and Stamps, 1999; Pauleen and Yoong, 2001; Shaik and Makhecha, 2019). While virtual meetings allow business continuity, questions remain regarding their impact on participant's perception of feeling others presence and engagement compared to traditional face-to-face meetings (Spurk and Straub, 2020). Being able to understand the impact of virtual business meetings as a tool to strengthen virtual engagement and remote interaction can provide significant insights about virtuality within the context of business networking and efficiency.

The aim of this study is therefore to explore the impact of virtual business meetings on attendees' perception of social presence and investigate the factors that influence virtual engagement and remote interaction. Our objective is to analyze ways to improve remote interaction and suggest business interventions in developing virtual engagement. Drawing from Social Presence Theory (Short, Williams and Christie, 1976), this research examines how the sense of being 'present' in a virtual setting affect interaction, engagement, and overall meeting satisfaction. Examining the technology, virtual business meeting, as the medium that brings people together, we ask:

- What are the impacts of virtual business meetings on attendees' perception of social presence?
- What factors contribute to the attendees' virtual engagement during the virtual business meeting?

2.0 Theoretical Background

2.1 Business Trips and Meetings

Business trips and in-person meetings are certainly not a new phenomenon, and people have been traveling for work purposes for centuries. Encompassing meetings, conferences, networking events, traditional business trips have long been a fundamental way for corporates to communicate and facilitate relationship, negotiation, and knowledge exchange (Denstadli, Julsrud and Hjorthol, 2011). During the past few decades, traveling for business has increased, especially among managers and executives (Swarbrooke and Horner, 2012). Specifically, before COVID-19, there were approximately 445 million business trips annually, with business travellers making up around 12% of all airline passengers (Finances Online, 2019; Rosen, 2020; Becken and Hughey, 2021). Some of the factors related to globalisation and the growing need to communicate and engage with long-distance partners (Aguiléra, 2008), whereby the geographical expansion into new markets, multinational companies, and improved infrastructures influenced and necessitated a substantial increase in business travel (Beaverstock et al., 2010; Jones, 2013).

Prior to COVID-19, business trips were considered necessary for conducting meetings with partners outside of the company, even abroad (Oddou, Mendenhall and Ritchie, 2000; Coscia, Neffke and Hausmann, 2020). However, as the pandemic brought travel to a standstill, the transition to a new normal with virtual business meetings replacing business travel was inevitable. At the same time, these inevitable changes resurfaced previous concerns with regard to productivity, efficiency, and dilemmas of business travel and face-to-face meetings (Gustafson, 2012; Mason and Gray, 1999; Ivancevich, Konopaske and Defrank, 2003).

2.2 The Virtual Business Meeting as Business Communication Tool

Considering the availability of advanced ICTs, many question the need for business travel for the purpose of in-person business meetings. The rise of virtual business meetings including virtual conferences, forums, seminars, and networking events, redefined corporate communication and engagement (Waizenegger et al., 2020).

Increased demands in engagement and interaction between people working across different geographical locations and companies have resulted in the widespread use of virtual platforms. Communication through ICTs, such as Zoom and Microsoft Teams, became the norm globally during the pandemic, and these new forms of virtual business meetings replaced face-to-face meetings, where ICTs allowed remote engagement and interaction (Denstadli et al., 2013).

Virtuality in the business environment has influenced perceptions regarding communication. The move from the real environment to a virtual environment, which was originally proposed by Milgram and Kishino (1995) with the concept of 'virtuality continuum', has transformed how communication is perceived and conducted in business. This continuum expands from physical environments to virtual spaces including various levels of technologies such as augmented reality (AR), virtual reality (VR), mixed reality (MR), and immersive experiences. As computer-mediated realities and ICTs developed, the virtual business environment became more prevalent. Some of the early technologies for virtual meetings ranged from simple audioconferencing systems to electronically mediated meeting systems, allowing participants to connect to each other through technological devices and to meet within virtual spaces (Nunamaker et al., 1991). Over the years, companies recreated their business environments with the help of virtual meeting experiences, as firms such as Cisco and Polycom featured advanced telepresence technologies (Strengers, 2014). Continuous developments in ICTs allowed people to engage in mediated forms of interaction using live audio, video, and the integration of diverse collaborative tools (Panteli and Dawson, 2001; Laitinen and Valo, 2018). Although there are differences in the access and use of such ICTs, the flexibility entailed by the different forms of virtual business meetings encouraged distant partners, clients, and members to adopt and make use of these (Cisco WebEx, 2014). This further resulted in the evolution of contemporary platforms, whereby most of these now include a number of collaboration features, such as document sharing, break-out rooms, and Q&A pop-ups.

The COVID-19 outbreak further accelerated these advances (Jacobs, 2020; Frith, 2020). With the virtual business environment emerging as the new normal, the corporate environment itself changed to accommodate new work modalities, i.e., work-from-home, remote work, and hybrid work (Hassan, 2021; Da et al., 2022). During this process, the virtual business meeting became an effective complement for corporations and several businesses approached these changes as new opportunities to introduce virtuality in the business context (Garsten and Wulff, 2003; Chudoba et al., 2005; Dixon and Panteli, 2010). In doing so, they also began looking at ways to improve business productivity and efficiency on the basis of virtual communications (Chebly, Schiano and Mehra, 2020; Loredana, Irimias and Brendea, 2021; Farooq and Sultana, 2021).

2.3 Social Presence Theory

Within the context of virtuality, virtual communications, and thus virtual business meetings, the concept of presence and presence perceptions becomes important. Schuemie et al. (2001) have found that factors such as immersion, social interaction, naturalness, social reality, and interpersonal communication cues are important precursors to feelings and perceptions of presence in a virtual environment.

According to Short, Williams, and Christie (1976), Social Presence Theory claims that the effectiveness of mediated communication depends on the degree to which individuals perceive others as real and present in the interaction. The concept of social presence emerged as a result of telecommunications and has since developed further, due to the popularity of computer-mediated communications, such as e-mail and video calls. The concept can be defined as a “degree of salience of the other person in the interaction and the consequent salience of the interpersonal relationships...” and as the sense that another person is “real” and “there” when using a communication medium (Short, Williams and Christie, 1976, p. 65).

The Social Presence Theory is rooted in the social psychological theories that draw from interpersonal communication and symbolic interactionism (Biocca et al., 2001; Tu, 2001). Originally, it was focused on the non-mediated interaction associated with two different concepts of ‘intimacy (Argyle and Dean, 1965) and ‘immediacy’ (Wiener and Mehrabian, 1986), whereby the degree of intimacy is expressed by verbal and non-verbal behaviour such as physical distance, eye contact, smiling, and personal topics of conversation, and the degree of immediacy can be measured by psychological distance. Clearly, social presence, intimacy, and immediacy are interrelated as immediacy behaviours are used to create and maintain intimacy while immediacy behaviours enhance social presence (Gunawardena, 1995).

Social presence is essential in influencing online interaction and improving the effectiveness of virtual meetings (Terry and Garrison, 2001; Schrire, 2004). It affects online interaction and user satisfaction as well as the depth of online discussions (Tu and McIsaac, 2002; Schroeder, 2002). The presence of others and co-existence can provide multi-sensory experiences and perspectives (Tu and McIsaac, 2002; Schroeder, 2002), whereby virtuality and social presence originate within the human experience view, rather than that technology and computer hardware (Steuer, 1992). Along these lines, with regard to virtual meetings, social presence can be perceived as the degree to which attendees feel they are participating in communicative interactions (Biocca, Harms and Burgoon, 2003). As such, there is no limit to the degree of feeling that the other person is there and to the degree of feeling as if one is having a direct conversation with another person. Therefore, the scope of communication, engagement, and interaction broadens (Weidlich, Göksün and Kreijns, 2022).

2.3.3 Telepresence

As technology advanced, the characteristics of social presence and its quality evolved with the ICTs and expanded to how individuals perceive and experience the sense of 'being there' in a mediated environment (Steuer, 1992; Cui, Lockee and Meng, 2012). The development of computer-mediated communication and online technologies led to the further development of the social presence concept and to the introduction of the term 'telepresence' (Biocca, et al., 2001). Telepresence is particularly relevant to virtuality and virtual environment whereby broadband connectivity and technologies (i.e., high-quality audio-visual elements, interactive features) allow people to feel and be 'fully present' through the ICT-mediated environment (Steuer, 1992; Denstadli et al., 2013). Today, with increased requirements for virtual rather than face-to-face meetings, the concept of telepresence plays an important role in expanding the scope of human activity and the sense of human presence at a global level and outside the confines of the physical space. Communication technology enables the development of impressions of vivid experiences with and of others who are not physically together (Muhlbach, Bocker and Prussog, 1995; Erickson, at el., 2010).

2.4 Virtual Engagement and Remote Interaction

An important question that arises is, how can virtual engagement and interaction be examined? According to Kearsley and Shneiderman (1998), engagement involves an active cognitive process, problem-solving, reasoning, decision-making, and evaluation, and reflects the attention level of meeting attendees (Shaw and Linnecar, 2007; Schwarz et al., 2014). It is often described as a psychological or affective state such as commitment, involvement, attachment, and others, and it is considered a performance construct (role performance, effort, or observable behaviours) (Macey and Schneider, 2008; Robertson-Smith and Markwick, 2009). Irrespective of the precise definition, engagement is typically among the core objectives of a business meeting (virtual or not), because it is engagement that can lead to and encourage more collaborative activities and opportunities between attendees. In the long term, these can then benefit organisational performance and success (Arnfalk and Kogg, 2003; Yoerger, Crowe and Allen, 2015; Moreira et al., 2023).

Despite the fact that online presence and virtuality have evolved into aspects of everyday life, there is no clear definition for virtual engagement specifically. Chewning (2018, p. 441) defines it as “the social enactment of ICT (Information and Communication Technology) as part of a larger relational context in which one connects with social, information, and resource networks in order to affect change, cocreation, and commitment toward a particular engagement object.” In other words, virtual engagement is positioned as a unique type of engagement that is influenced and shaped by technological affordances.

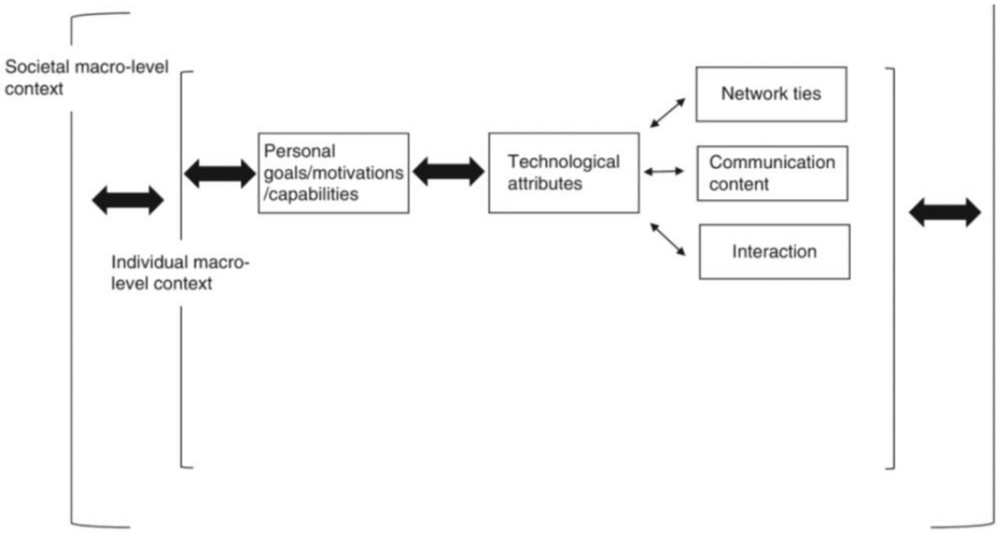

Societal macro-level context
Individual macro-level context
Personal goals/motivations /capabilities
Technological attributes
Network ties
Communication content
Interaction

Figure 1. **Virtual Engagement Framework by Chewning (2018)**

According to Chewning’s Virtual Engagement Framework (2018) as depicted in Figure 1, virtual engagement in digital communication is shaped by three core elements: Affordances, the features of digital platforms that foster interaction by using technical tools such as chat functions and break-out rooms; Networks, the social and professional connections which influence participation of attendees; and

Communication, the strategies used to continue interaction between attendees and promote engagement. Based on this framework, virtual engagement has resulted from the interaction that occurs between users and technologies. Technology becomes the medium through which attendees come together remotely and interact and thus engagement takes a different form from when interacting face-to-face.

As businesses continue adopting and leveraging remote work modalities and virtual environments, the virtual business meeting has become an opportunity to increase communication, that, with the help of technology, can create and facilitate more effective engagement and interactions (Woodruff et al., 2021; Al-Sharafi et al., 2022; Chang et al., 2022). The transition allows businesses to engage not only internal members from different locations but also external stakeholders as well. To make the virtual business meeting a place for successful relationship formation and a place to engage and interact with other attendees, what is needed is for this to be strategically introduced (Porter et al., 2011).

Earlier studies have examined intra-organisational and team-based virtual engagement and discussed strategies for designing and using effective virtual engagement systems and creating remote interaction between team members (Porter et al., 2011). These studies highlighted the importance of creating structured and engaging environments to encourage active interaction between team members by using clear communication processes or well-designed virtual systems that can help replace face-to-face communication. Not much is known however with regards to inter-organisational virtual meetings, i.e., meetings that bring attendees together from different organisations. In this study, we therefore focus specifically on such inter-organisational meetings.

3.0 Methodology

This study focuses on inter-organisational virtual business meetings (VBM) and adopts a mixed-methods approach to examine the benefits and challenges in relation to attendees' engagement and interaction. The effectiveness and efficiency of virtual business communication depends on user engagement, interaction quality, and the ability to stimulate a sense of connection among participants (Panteli and Dawson, 2001). These factors are more likely to actively contribute when participants perceive a stronger sense of presence. Applying understanding of virtual engagement based on

Chewning's framework, this study delves into the intersection of Social Presence theory and factors that highlight the importance of presence in virtual business meetings.

The study was conducted over a three-month period (June-August 2022) in two phases. Both quantitative and qualitative data were collected sequentially to provide a comprehensive understanding of presence and virtual engagement (Fetters, 2020). The study consisted of two primary methods: in phase 1, a user experience questionnaire was used to capture quantitative insights to investigate the impact of virtual business meeting and ways to improve virtual engagement as well as remote interaction, and in phase 2, semi-structured interviews were conducted to explore participant's experiences in greater depth.

In the first phase, the study recruited participants through purposive sampling by targeting professionals with prior experience in virtual business meetings across diverse industries for organizational and business purposes. They were recruited through online profiles (LinkedIn) or through attendance lists of business events such as business meetings, conferences or forums A total of 104 participants from finance, technology, consulting, education, and customer services completed the questionnaire, however, 8 participants were removed after retrieving the data, which resulted a primary sample of 96 respondents from 20 different countries, all of whom had the experience of inter-organisational VBM. The questionnaire was built based on User Experience Questionnaire form on a typical usability aspects (efficiency, perspicuity, dependability) and user experience aspects (originality, stimulation) (Schrepp, Hinderks and Thomaschewski, 2017; Alberola, Walter and Brau, 2018). The questionnaire included items that measured on a 5-point Likert Scale and included a free text field where participants were invited to describe their own experiences and feelings with regard to virtual business meetings in narrative form. This phase was largely exploratory and aimed at exploring attendees' perception of social presence and at understanding virtual engagement during remote interactions.

The second phase focused on respondents' experiences and perspectives (Kendall, 2010) through conversational in-depth semi-structed one-to-one interviews, where the aim was to delve into their individual perspectives and perceptions of VBM in relation to virtual engagement. Overall, 10 participants were selected for the interview based on their survey responses. The selected participants were identified through the first phase and ensured a diverse representation in terms of gender, nationality, and job roles. This allowed us to pool together data (their narrative responses to the questionnaire and interview material), which was then analysed through content analysis. This approach helps us adopt a more in-depth approach and consider and appreciate the meaning of the data (Potter and Levine-Donnerstein, 1999; Hsieh and Shannon, 2005; Elo and Kyngäs, 2008), as we were able to explore further the links and the patterns between the two phases and explore subjective interpretations (Hsieh and Shannon, 2005).

4.0 Findings

4.1 Virtual Business Meeting and Social Presence

Participants in the study shared a variety of perspectives on their experiences with virtual business meetings in relation to if and how such meetings give them the opportunity and space to feel the presence of fellow attendees (38.45% responded positively; see figure 2) and in ways that feel like genuine connection (47.8% responded positively; see figure 3). Some words and phrases they used included ‘real’, ‘same space’, and ‘others are there with you’. They further provided diverse conditions and situations they experienced during these meetings, still, many agreed that they felt to some extent of others’ presence during VBMs despite such differences and their meetings being hosted with a medium of technology in a virtual space. However, the different conditions, i.e., the size of the business meeting and attendance numbers, weakened or strengthened their perceptions of others’ presence. In other words, social presence is experienced and felt during virtual business meetings, and factors such as the above influence and control the overall experience of sensing others’ presence in virtuality.

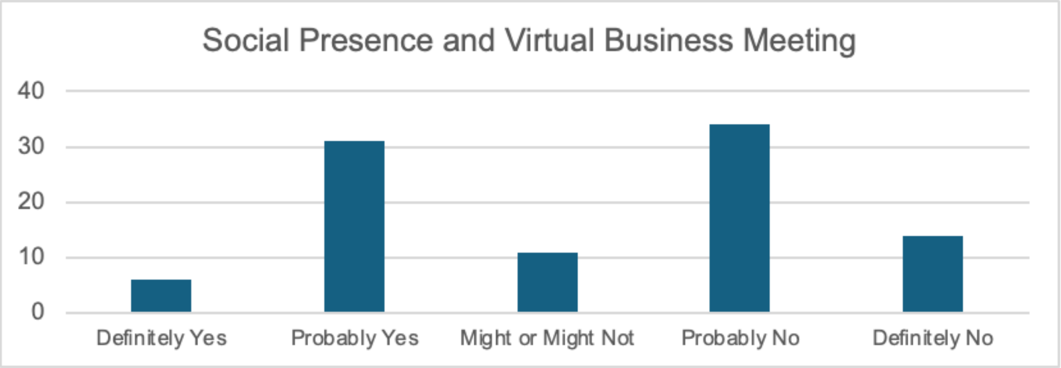


Figure 2. Social Presence and Virtual Business Meeting

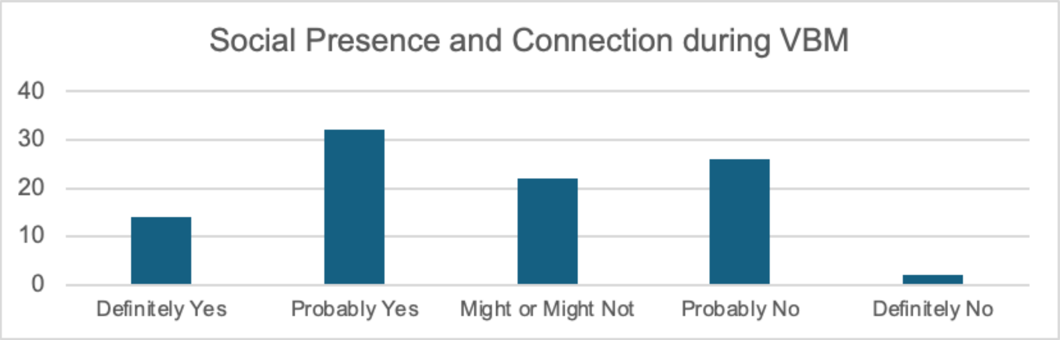


Figure 3. Social Presence and Connection during Virtual Business Meeting

	NOR	Ratio	NOR	Ratio	NOR	Ratio	NOR	Ratio	NOR	Ratio
Figure 2	6	6.25%	31	32.2%	11	11.4%	34	35.4%	14	14.9%
Figure 3	14	14.5%	32	33.3%	22	22.9%	26	27.1%	2	2.1%

** NOR refers to Number of Respondents*

Table 1. Number of Respondents and Ratio of Figure 2 and 3

4.2 Virtual Engagement and Remote Interactions

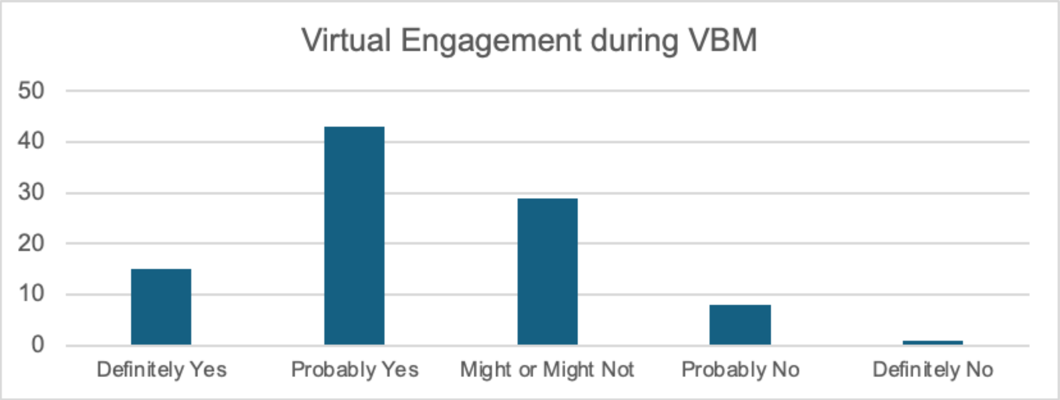


Figure 4. Virtual Engagement during Virtual Business Meeting

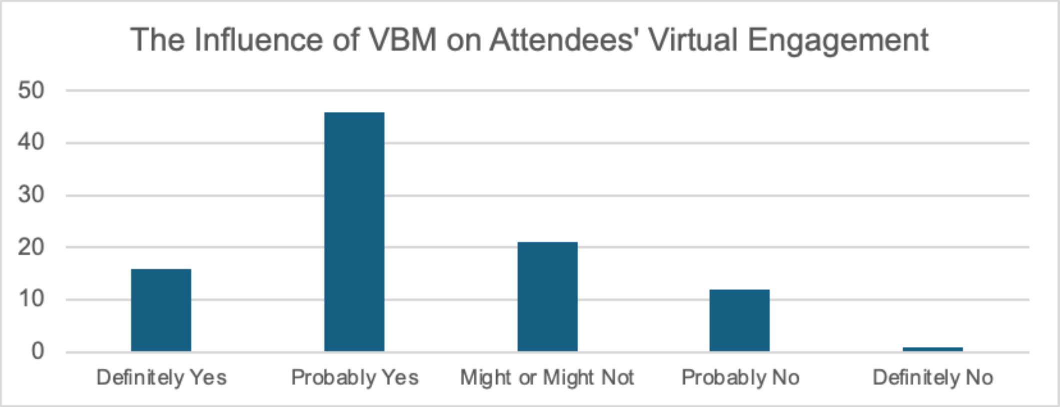


Figure 5. Influence of Virtual Business Meeting on Attendees’ Virtual Engagement

	NOR	Ratio	NOR	Ratio	NOR	Ratio	NOR	Ratio	NOR	Ratio
Figure 4	15	15.6%	43	44.8%	29	30.2%	8	8.3%	1	1.1%
Figure 5	16	16.6%	46	47.9%	21	21.8%	12	12.5%	1	1.1%

** NOR refers to Number of Respondents*

Table 2. Number of Respondents and Ratio of Figure 4 and 5

To understand virtual engagement and remote interaction during virtual business meetings, we asked participants to talk about their experiences and feelings of engagement in a virtual space with others. In their majority, participants expressed positive feelings and confirmed that they experienced engagement (60.4% responded being definitely or mostly engaged during VBMs; see figure 4). Still, an interesting question to be asked is what are the factors that facilitate virtual engagement and remote interaction.

Drawing from the interview materials and free text field from the questionnaire, different factors facilitate and also influence virtual engagement and remote interactions during virtual business meetings. Some of those that were repeatedly mentioned (explicitly or not) reflected factors that can be grouped as ‘technical’, ‘personal’, or ‘physical’. These are discussed next.

4.2.1 Technical Factors

Technical factors refer to elements such as ‘screen sharing’, ‘Q&A’, and ‘break-out room’, which were most commonly brought up by participants (across questionnaire respondents and interviewees). They also emphasized the importance of camera and microphone features and settings. These features have a direct influence on attendees’ engagement, and features that are related to visual components have an impact on their perception of social presence. This is precisely because they allow them to see each other’s reactions through their screens. With regards to audio in particular, audio was highlighted as a common inconvenience whereby it can create disruption during collaborative activities and participation, because, often, only one attendee at a time can speak while the rest of the attendees are able only to listen. This reveals one of the biggest differences and challenges compared to in-person business meetings: virtual business meetings are typically governed by a different etiquette and tend to be much more regimented, allowing in many cases only one-way communication, such as presentations or speeches.

4.2.2 Personal Factors

Personal factors exhibit great variety and refer primarily to attendees' personal abilities, skills, and personalities. Some interviewees used descriptives such as '*introvert'* to describe other attendees and highlighted *individual preparation* and *communication skills* as important elements.

Interestingly, participants offered a range of experiences of virtual business meetings, whereby *individual personality* seemed to be of importance. Some argued that virtuality encourages attendees to be more confident regardless of their personality in real life seeing due to the perceived distance between attendees. Others also highlighted that these meetings helped them to be more talkative and active, both of which contributed toward more interaction and engagement.

Individual communication skills, preparation, and *formal manners* were brought up by many. In particular, they explained through various examples how these elements had positive effects by improving their virtual engagement. *Formal manners* were described by referring to their' and others' *'dress', 'outfit',* and *'attitude'* and that these made much of a difference in their virtual engagement during virtual business meetings. A few of them also described that attendees' dressing choices led to having more/less formal interactions and attitudes, as such choices influenced their attitudes and responses.

4.2.3 Physical Factors

One common theme that was mentioned was the behaviour of the attendees during virtual business meetings. Words describing body movements such as *'hand gesture', 'nodding', 'eye contact',* and *'facial expression'* were commonly used to express their experiences. Still, different approaches to these body languages were examined as some described that these gestures made them feel as if they were talking to other attendees' face to face, i.e., bridging the distance between them, where these helped convey their level of excitement and interest. Some were particularly specific in the way they described these gestures and facial expressions, by referring to *visual cues,* such as *eyebrow movement, eye contact, smiling, frowning,* and *shaking heads.* These expressions helped attendees feel that others remained to engage in the conversations. Others also mentioned that the technological medium itself afforded certain physical movements to be hidden away from the camera. For instance, one of the interviewees mentioned that attendees can hide their nervousness or anxiety movements such as touching their hands or clenching their legs as the angle of the camera limits/protects them from exposing feelings and unintentional body movement that attendees wish to conceal. In other words, the medium of technology possibly masks away attendees' feelings and actions when they do not wish to share or reveal these to others.

4.3 Advantages of Virtual Business Meeting and Recommendation Rate

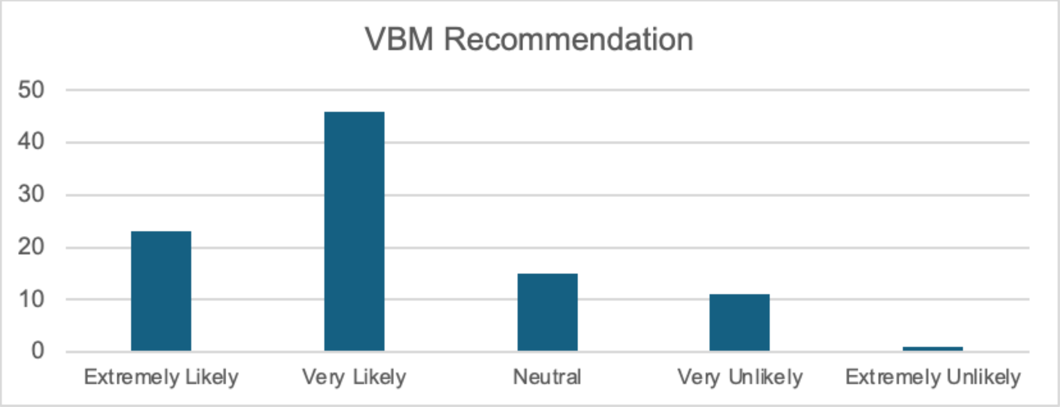


Figure 6. Recommendation

	NOR	Ratio	NOR	Ratio	NOR	Ratio	NOR	Ratio	NOR	Ratio
Figure 6	23	23.9%	46	47.9%	15	15.6%	11	11.5%	1	1.1%

** NOR refers to Number of Respondents*

Table 3. Number of Respondents and Ratio of Figure 6

As part of our study, we also asked participants how likely they were to recommend to others and to prefer virtual business meetings (over face-to-face meetings) and the reasons why. A total of 71.8% (23.9% extremely likely; 47.9% with very likely; see figure 6) expressed a positive attitude and preference for virtual meetings because they considered these to be more advantageous. Table 4 includes some consolidated comments, grouped per thematic areas. As shown, participants indicated that among the main advantages are those of being able to reach a wider range of collaborators in terms of their geographical location, as expected. However, other reasons included the reduction in unnecessary traveling, and in many cases, participants referred to environmental sustainability as well. In addition, participants explained that virtual business meetings support interactions and engagement in a fashion that does not sacrifice or jeopardise their effectiveness and that such meetings tend to be more convenient as well, and supportive of work-life balance and overall efficiency.

Category	Respondents' Comments
Geographic Range	The geographic range that can be covered (i.e., anywhere in the world)
	Organising a meeting that would not have been possible otherwise (people attending from different countries)
	No limitation on physical location

	…can have meeting with people from different countries, especially who are located in different countries
	…can talk with people who are at different countries or far away
Reduction in Unnecessary Business Travel(s)	Lack of travel and ease of connection globally
	Minimize travel time, more timing flexible
	No more unnecessary business travel or long international travel
	No far business travels/ don't have to drive hours for a meeting
	Reduction in unnecessary, carbon-intensive travel. More time to meet, not using time to travel. Loved not having to purchase gas
Virtual Interaction and Remote Engagement	Remote access, there is no need for everyone to be in the same physical location and it makes a meeting much more accessible.
	Maintaining connection, Business as usual
	Can interact with others remotely
	virtually getting to know each other and communicate
	Remote engagement. It was simpler to actually get people to commit to meeting.
	Virtual interaction with others
	I could talk to people I wouldn't have been able to face to face
Convenience	The usefulness of virtual business meeting was comfortableness as it is easily accessible through the computer. Also, some people will be able to express yourself further.
	The speed and efficiency of starting and ending a meeting
	Don't have to go to office and don't have to prepare proper setting like clothing
	Shared documents / Chat options / Raised hand icon / Calendar invitation
	Easily accessible, can meet anywhere with Wi-Fi access
Others	Gives employees freedom to care for self. Don't have to be IN the office. offices are an antiquated environment for most operations.
	I don't need to run to the other meeting room if I have two meetings.
	Being able to work from home
	Access to more webinars and seminars
	Health and safety / flexibility

Table 4. Advantages and Usefulness of Virtual Business Meeting

5.0 Discussion

5.1 The Impact of Virtual Business meeting on Attendees' Perception of Social Presence

Our findings indicate that attendees' perception of social presence during virtual business meeting influence their engagement and interactions. Also, these perceptions lead to positive experiences and outcomes regarding the effectiveness of attendees and virtual business meetings. This aligns with Social Presence Theory which highlights the degree of individuals' perception of feeling other's existence as 'real' in a mediated communication environment and affects their satisfaction (Short, Williams and Christie, 1976).

However, across the two phases, participants noted that their perception of social presence depend on specific conditions under which the virtual business meeting is held, shared specific experiences and circumstances, and described the factors that influence their perceptions (e.g., camera on/off, degree of formality). These insights suggest that while virtual environment facilitate the sense of presence, different factors can enhance or diminish this experience. As such, and based on our findings, we thus propose that the diversity of factors we identified (technical, personal, physical) as situational variables that impact perceived presence in virtual settings and influence attendees' virtual engagement and remote interaction during such meetings (see figure 7). These findings also reflect with Chewning's (2018) Virtual Engagement Framework which emphasizes the interconnection and interplay between technological affordance, individual user characteristics and personalities as network, and the communication environment in shaping virtual engagement.

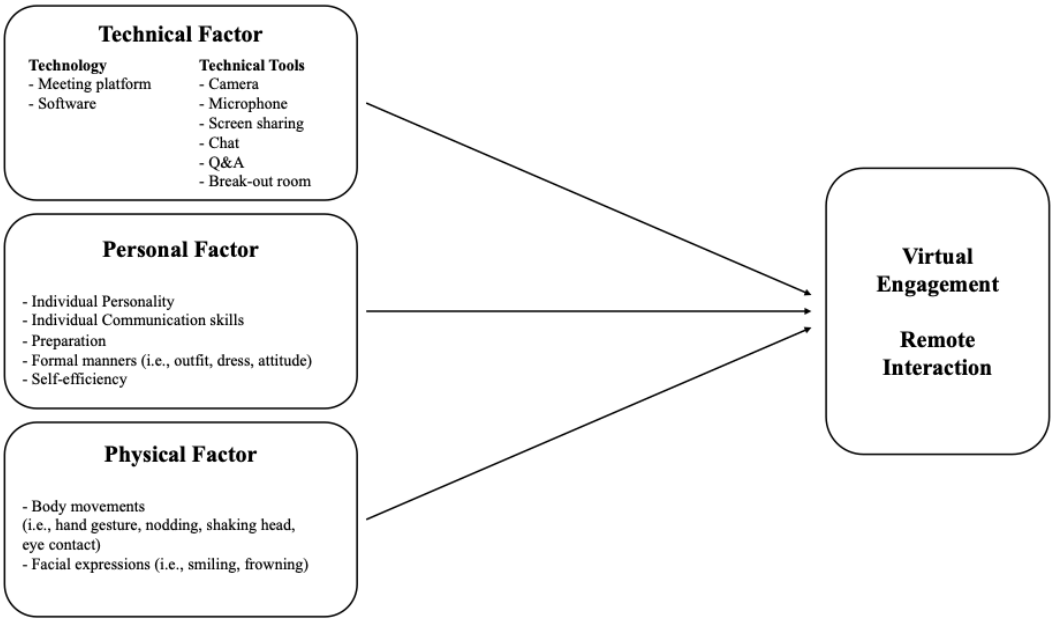


Figure 7. Factors contributing to virtual engagement and remote interaction during virtual business meeting

In more detail, technical factors refer to *technology choices*, including the meeting platform and software, and *technical tools* including camera/microphone, screen sharing, chat, Q&A, and break-out rooms. These elements function together as the essentials that enable social presence and can actively support virtual engagement and remote interaction, as it is inevitable to conduct a virtual business meeting without using these. Features such as Q&A options or survey pop-ups, break-out rooms, or chat boxes encourage attendees to interact effectively. These features encourage attendees to communicate and interact more actively, and in turn support greater engagement in the virtual space. Correspondingly, if there are technical failures (in equipment or software) or omissions (e.g., the chatbox is disabled), then virtual business meetings are not as engaging and effective as the disruptions have negative effects on virtual engagement. To put it in another way, it is vital to highlight the digitized workforce using technologies and improve advanced technical strategies to develop a smooth virtual business meeting.

Next, the personal factors, as discussed earlier, correspond to aspects such as *individual personality, individual communication skills and preparation, formal manners,* and *self-efficiency.* Existing evidence suggests that these aspects are important for interpersonal communications; yet our study highlights that, when such communications and meetings take place in the virtual space, technology itself can act as the fence or veil that can help, e.g., introverts to overcome nervousness, and thus participate and engage with other attendees in a more constructive manner. Understandably, the medium of technology will have different impacts depending on one's personality (i.e., introvert vs extrovert, more/less talkative), but it can be leveraged regardless to support engagement and enable them to feel more at ease compared to an in-person meeting. The above can be beneficial for the rest of the attendees as well: the way a speaker talks, reacts and the way they use body language during a meeting can significantly influence the behaviours of attendees, and encourage or discourage engagement and participation. In turn, this will have an impact on the flow and conduct of the virtual business meeting as a whole, and thus support inter-organisational collaboration, knowledge, and insights exchange and cross-fertilisation of ideas, or not.

Finally, physical factors reflect elements such as body movements (e.g., hand gestures) and facial expressions (e.g., nodding, shaking the head, smiling or frowning, eye contact). Our findings confirm that such physical factors are crucial for capturing the attention and interest of attendees, as well as for sharing and communicate each other's attitudes. The individual movement provides attendees with a sense of others' understanding, participation, and following the flow of the virtual business meeting. In other words, the nonverbal communication tools, i.e., body movements and facial expressions, play a role in the construction and conveyance of others' status. As technology functions as the medium between parties, how technology functions and what it allows attendees to do is important. For example, turning off cameras that do not allow the communication of these physical factors can limit the way attendees perceive each other and their degree of engagement, because non-verbal cues (e.g., body movements and facial expressions) are equally important for building relationships and communicating with others.

5.2 Relationships between factors

While our study is exploratory and thus our findings cannot provide exact correlations, we note that our analysis indicates that the above-discussed factors are not independent of each other, but rather influence and shape each other (see figure 8). For example, even if attendees come together within a virtual space for the meeting, technology mediates their images, voices, and the overall ambiance of the room. (Technical factors - Physical factors) In other words, the camera feature and audio settings affect the way attendees can or cannot observe and experience others' body movements, facial expressions, and voice (Ekman and Friesen, 2003). (Technical factors - Personal factors) Similarly, attendees can manipulate (on most occasions) the positioning of their camera and use noise reduction features. In such cases, attendees may wish to point the camera to an area of their space where they feel more comfortable, or which frames them better, and equally hide away other areas that they do not wish to share with others. In other words, technology can be leveraged for impression management (Blunden and Brodsky, 2024), which directly influences how others perceive them, but further supports the relationship between technical and personal factors. Finally, because verbal and non-verbal cues influence one's personality and the way one communicates with others (de Vries et al., 2011; Hargie, 2019), we can also expect a relationship between personal and physical factors.

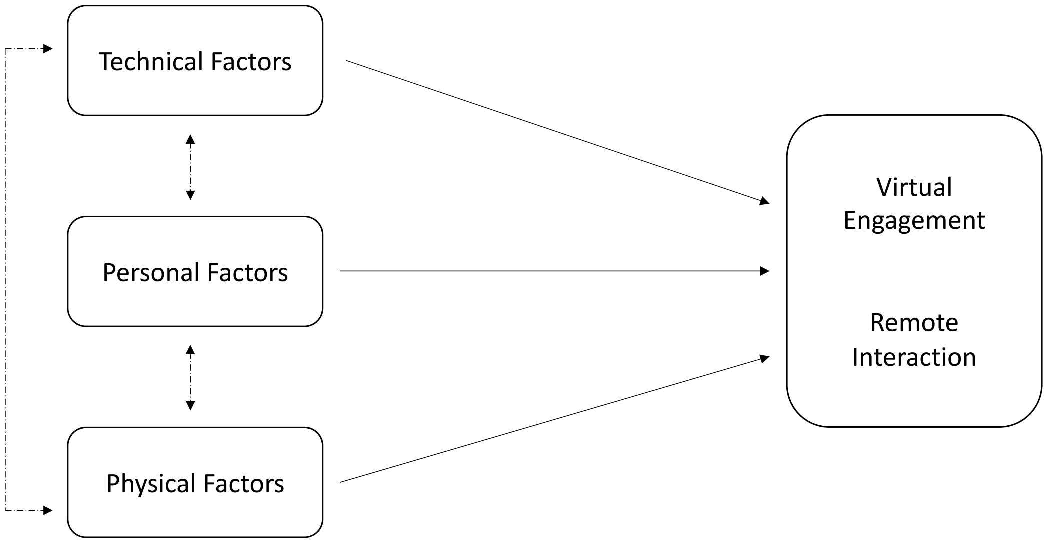


Figure 8. Correlations of factors

5.3 Recommendations

Based on our preliminary findings, we propose a set of recommendations that can support improvements in the way virtual business meetings are typically conducted. Our findings indicate that, while virtual business meetings are beneficial for several reasons, including the reduction of business travel that can be disruptive for attendees, participants reported having primarily neutral perceptions and experiences of others' social presence (more neutral rather than positive/negative). Poor social presence has a direct implication for the degree of interaction and engagement, thereby it can have a negative effect on capturing others' attention and stay connected.

Still, we can also deduce from our findings a few ways in which technological features can be used to enhance the overall experience and improve interaction virtually and remotely. *Pre and During Meeting Rules* (e.g., order of points in the agenda, follow-up actions, task allocation) have been frequently invoked as supportive of increased engagement. These help in establishing and communicating the purpose and significance of the meeting by helping attendees prepare in advance, organise their thoughts, and prepare their IT equipment. Also, by reducing to an extent possible distractions that may occur during the meetings, the rules will allow the virtual business meetings to progress smoothly and flawlessly. Similarly, and due to the nature of a virtual meeting with technology in the medium, *Camera and Audio Settings* have also been invoked as something that can be negotiated between attendees but as something that also needs to be agreed upon clearly. Our findings suggest that some may selectively turn on/off the camera as well as their audio. However, the camera has a strong impact on each other's perception of social presence because it influences virtual engagement and helps increase remote interaction (Reidsma et al., 2007; Molinillo et al., 2018) through visualisation, one of the five basic senses. Similarly, the audio feature may support a better flow (e.g., whereby an attendee appreciates when is the right time to speak and thus be better integrated within the meeting flow) but also create disruption and negative experiences (e.g., echo and noise). Whether and how these features should be used will largely depend on the nature of the meeting and the participants themselves, but still, as mentioned earlier, this can be negotiated. However, the meeting etiquette would need to be mutually agreed upon in terms of the use of such technological features in advance to support interactions and engagement that can lead to positive experiences. Even with the medium of technology and virtuality provides a different environment from a in-person meeting, it is crucial to attend with *formal manners* as attendees are participating in a business meeting. Virtual business meetings are a space where people from different countries, regions, and companies are gathered to conduct their formal activities on business-related issues (Bilbow, 1995; Rogerson-Revell, 2008; A. Allen et al., 2014). In other words, professionalism from individuals will be judged through formal attitudes, i.e., appropriate interaction skills and presentable outfits, for the virtual business meeting.

6.0 Conclusion

The COVID-19 pandemic has led to an increased number of people and organisations adopting new ways of working, such as hybrid work and remote work, whereby technology and systems have played a clear role in supporting business continuity through information sharing and networking building. We consider that the importance of virtual business meetings will continuously increase in the future. Also, due to digital transformations taking place across sectors, virtual business communication tools will continue to be a necessary tool to facilitate distance communication not only between internal employees but also external stakeholders. In light of this, our study provides some preliminary insights into how technology and communication tools can support engagement and interactions by strengthening social presence perceptions.

Like all studies, our research comes with certain limitations. First, in terms of sampling, the first phase of the study was conducted through 96 respondents with substantial experience in virtual business meetings within the context of inter-organisational collaborations. However, these meetings take place intra-organisationally between cross-functional and project teams. It would be interesting to explore whether and to what extent the three identified groups of factors influence differently inter- and intra- organisational collaborations, productivity, and engagement.

Further, during the second phase, we sampled 10 participants from the first phase for a more in-depth investigation through semi-structured interviews. Yet, a different research design could have provided even richer insights. For example, considering that often there is a difference between what technology users think, say, and what we actually do, it would be interesting to explore the above through observational methods (e.g., ethnographic and ethnomethodological techniques), which can be coupled with digital trace data for corroboration and triangulation purposes.

Third, we conducted our study during the COVID-19 pandemic which reflects a point in time when several had to transition to new ways of working abruptly from the physical workplace to the digital environment. Our sample consists of participants who had experienced virtual business meetings prior to the pandemic. We consider, however, that future research could revisit these questions in the near future and examine how business professionals interpret virtual business communication post-pandemic and have shifted the perceptions in terms of social presence and virtual engagement.

References

Aguilera, A. (2008). Business travel and mobile workers. *Transportation Research Part A: Policy and Practice*, 42(8), pp.1109–1116. doi:10.1016/j.tra.2008.03.005.

Al-Sharafi, M.A., Al-Emran, M., Arpaci, I., Marques, G., Namoun, A. and Iahad, N.A. (2022). Examining the Impact of Psychological, Social, and Quality Factors on the Continuous Intention to Use Virtual Meeting Platforms During and beyond COVID-19 Pandemic: A Hybrid SEM-ANN Approach. *International Journal of Human–Computer Interaction*, pp.1–13. doi:10.1080/10447318.2022.2084036.

Alberola, C., Walter, G. and Brau, H. (2018). Creation of a Short Version of the User Experience Questionnaire UEQ. *i-com*, 17(1), pp.57–64. doi:10.1515/icom-2017-0032.

Allen, J., Beck, T., W. Scott, C. and G. Rogelberg, S. (2014). Understanding workplace meetings; A qualitative taxonomy of meeting purposes. *Management Research Review*, 37(9), pp.791–814. doi:10.1108/mrr-03-2013-0067.

Amico, L. (2021). *A Guide to the Virtual Meeting*. [online] Harvard Business Review. Available at: https://hbr.org/2021/10/a-guide-to-the-virtual-meeting.

Anderson, A.H., McEwan, R., Bal, J. and Carletta, J. (2007). Virtual team meetings: An analysis of communication and context. *Computers in Human Behavior*, 23(5), pp.2558–2580. doi:https://doi.org/10.1016/j.chb.2007.01.001.

Anthony, S.D., Cobban, P., Painchaud, N. and Parker, A. (2021). 3 Steps to Better Virtual Meetings. *Harvard Business Review*. [online] 19 Feb. Available at: https://hbr.org/2021/02/3-steps-to-better-virtual-meetings.

Argyle, M. and Dean, J. (1965). Eye-Contact, Distance and Affiliation. *Sociometry*, 28(3), p.289. doi:10.2307/2786027.

Arnfalk, P. and Kogg, B. (2003). Service transformation—managing a shift from business travel to virtual meetings. *Journal of Cleaner Production*, 11(8), pp.859–872. doi:10.1016/s0959-6526(02)00158-0.

Beaverstock, J.V., Derudder, B., Faulconbridge, J. and Witlox, F. (2010). *International business travel in the global economy*. Farnham: Ashgate, Cop, pp.217–238.

Becken, S. and Hughey, K.F. (2021). Impacts of changes to business travel practices in response to the COVID-19 lockdown in New Zealand. *Journal of Sustainable Tourism*, 30(1), pp.108–127. doi:10.1080/09669582.2021.1894160.

Béland, L.-P., Brodeur, A. and Wright, T. (2020). *The Short-Term Economic Consequences of COVID-19: Exposure to Disease, Remote Work and Government Response*. www.iza.org. Available at: https://www.iza.org/publications/dp/13159/the-short-term-economic-consequences-of-covid-19-exposure-to-disease-remote-work-and-government-response.

Bilbow, G.T. (1995). Requesting strategies in the cross-cultural business meeting. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 5(1), pp.45–55. doi:10.1075/prag.5.1.02bil.

Biocca, F., Harms, C. and Burgoon, J.K. (2003). Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence: Teleoperators and Virtual Environments*, 12(5), pp.456–480. doi:10.1162/105474603322761270.

Biocca, F., Harms, C., & Gregg, J. (2001). The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. Paper presented at the 4th International Workshop on Presence, Philadelphia, PA.

Blunden, H. and Brodsky, A. (2024). A Review of Virtual Impression Management Behaviors and Outcomes. *Journal of Management*. doi:https://doi.org/10.1177/01492063231225160.

Chang, H., Varvello, M., Hao, F. and Mukherjee, S. (2022). A Tale of Three Videoconferencing Applications: Zoom, Webex, and Meet. *IEEE/ACM Transactions on Networking*, pp.1–16. doi:10.1109/tnet.2022.3171467.

Chebly, J., Schiano, A. and Mehra, D. (2020). The Value of Work: Rethinking Labor Productivity in Times of COVID-19 and Automation. *American Journal of Economics and Sociology*, 79(4), pp.1345–1365. doi:10.1111/ajes.12357.

Chewning, L.V. (2018). Chapter 7. Virtual engagement : A theoretical framework of affordances, networks and communication. In: *The handbook of communication engagement*. Hoboken: Wiley-Blackwell, pp.439–452.

Chudoba, K.M., Wynn, E., Lu, M. and Watson-Manheim, M.B. (2005). How virtual are we? Measuring virtuality and understanding its impact in a global organization. *Information Systems Journal*, 15(4), pp.279–306. doi:10.1111/j.1365-2575.2005.00200.x.

Cisco WebEx. (2014). *Why WebEx? Connect with anyone, anywhere, any time*. Retrieved from http://www.webex.com/why-webex/overview.html

Cohen, S.A. and Kantenbacher, J. (2019). Flying less: personal health and environmental co-benefits. *Journal of Sustainable Tourism*, 28(2), pp.361–376. doi:10.1080/09669582.2019.1585442.

Coscia, M., Neffke, F.M.H. and Hausmann, R. (2020). Knowledge diffusion in the network of international business travel. *Nature Human Behaviour*, 4(10), pp.1011–1020. doi:https://doi.org/10.1038/s41562-020-0922-x.

Cui, G., Lockee, B. and Meng, C. (2012). Building modern online social presence: A review of social presence theory and its instructional design implications for future trends. *Education and Information Technologies*, 18(4), pp.661–685. doi:10.1007/s10639-012-9192-1.

Da, S., Fladmark, S.F., Wara, I., Christensen, M. and Innstrand, S.T. (2022). To Change or Not to Change: A Study of Workplace Change during the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, 19(4), p.1982. doi:10.3390/ijerph19041982.

de Vries, R.E., Bakker-Pieper, A., Konings, F.E. and Schouten, B. (2011). The Communication Styles Inventory (CSI). *Communication Research*, 40(4), pp.506–532. doi:10.1177/0093650211413571.

Denstadli, J.M., Gripsrud, M., Hjorthol, R. and Julsrud, T.E. (2013). Videoconferencing and business air travel: Do new technologies produce new interaction patterns? *Transportation Research Part C: Emerging Technologies*, 29, pp.1–13. doi:10.1016/j.trc.2012.12.009.

Denstadli, J.M., Julsrud, T.E. and Hjorthol, R.J. (2011). Videoconferencing as a mode of communication: A comparative study of the use of videoconferencing and face-to-face meetings. *Journal of Business and Technical Communication*, 26(1), pp.65–91. doi:https://doi.org/10.1177/1050651911421125.

Dixon, K.R. and Panteli, N. (2010). From virtual teams to virtuality in teams. *Human Relations*, 63(8), pp.1177–1197. doi:10.1177/0018726709354784.

Ekman, P. and Friesen, W.V. (2003). *Unmasking the face : a guide to recognizing emotions from facial clues*. Cambridge (Mass.): Prentice-Hall, Cop.

Elo, S. and Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1), pp.107–115. doi:10.1111/j.1365-2648.2007.04569.x.

Erickson, T., Kellogg, W. A., Shami, N. S., & Levine, D. (2010). Telepresence in virtual conferences: An empirical comparison of distance collaboration technologies. In *Proceedings of CSCW 2010*.

Farooq, R. and Sultana, A. (2021). The potential impact of the COVID-19 pandemic on work from home and employee productivity. *Measuring Business Excellence*, 26(3), pp.308–325. doi:10.1108/mbe-12-2020-0173.

Fetters, M.D. (2020). *The Mixed Methods Research Workbook: Activities for Designing, Implementing, and Publishing Projects*. [online] SAGE Publications, Inc. Available at: https://doi.org/10.4135/9781071909713.

Finances Online. (2019). 105 Critical Business Travel Statistics: 2021/2022 Spending & Concerns Analysis. [online] Available at: https://financesonline.com/business-travel-statistics/#link1.

Flowers, A.A. and Gregson, K. (2012). Theoretical and Practical Aspects of Conducting Meetings and Events in Virtual Worlds. *International Journal of Strategic Information Technology and Applications*, 3(4), pp.48–64. doi:https://doi.org/10.4018/jsita.2012100104.

Frith, J. (2020). Introduction to Business and Technical Communication and COVID-19: Communicating in Times of Crisis. *Journal of Business and Technical Communication*, 35(1), pp.1–6. doi:10.1177/1050651920959208.

Garsten, C. and Wulff, H. (2003). *New technologies at work: people, screens, and social virtuality*. *New technologies at work: people, screens, and social virtuality*, Oxford: Berg, pp.91–164.

Gunawardena, C. (1995). Social Presence Theory and Implications for Interacion and Collaborative Learning in Computer Conferences. *Intl. J. of Educational Telecommunications*, 1(2/3), pp.147–166.

Gustafson, P. (2012). Managing business travel: Developments and dilemmas in corporate travel management. *Tourism Management*, [online] 33(2), pp.276–284. doi:10.1016/j.tourman.2011.03.006.

Hargie, O. (2019). *The handbook of communication skills*. 4th ed. New York: Routledge.

Hassan, Dr.S. (2021). Impact of Covid-19 on people and Work from Home. *Pakistan BioMedical Journal*, 3(2). doi:10.52229/pbmj.v3i2.10.

Hsieh, H.F. and Shannon, S.E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, 15(9), pp.1277–1288. Available at: https://journals.sagepub.com/doi/abs/10.1177/1049732305276687?journalCode=qhra.

Ivancevich, J.M., Konopaske, R. and Defrank, R.S. (2003). Business travel stress: A model, propositions and managerial implications. *Work & Stress*, 17(2), pp.138–157. doi:10.1080/0267837031000153572.

Jacobs, G. (2020). Business Communication and COVID-19. *Business Communication Research and Practice*, 3(2), pp.73–75. doi:10.22682/bcrp.2020.3.2.73.

Kearsley, G. and Shneiderman, B. (1998). Engagement Theory: A Framework for Technology-Based Teaching and Learning. *Educational Technology*, [online] 38(5), pp.20–23. Available at: https://www.jstor.org/stable/44428478.

Kendall, L. (2010). Chapter 5. The Conduct of Qualitative Interviews: Research Questions, Methodological Issues, and Researching Online. In: J. Coiro, M.

Knobel, C. Lankshear and D.J. Leu, eds., *Handbook of research on new literacies*. New York ; London: Routledge, pp.133–150.

Laitinen, K. and Valo, M. (2018). Meanings of communication technology in virtual team meetings: Framing technology-related interaction. *International Journal of Human-Computer Studies*, 111, pp.12–22. doi:10.1016/j.ijhcs.2017.10.012.

Lipnack, J. and Stamps, J. (1999). Virtual teams: The new way to work. *Strategy & Leadership*, 27(1), pp.14–19. doi:https://doi.org/10.1108/eb054625.

Loredana, M., Irimias, T. and Brendea, G. (2021). Teleworking During the COVID-19 Pandemic: Determining Factors of Perceived Work Productivity, Job Performance, and Satisfaction. *www.amfiteatrueconomic.ro*, 23(58), p.620. doi:10.24818/ea/2021/58/620.

Macey, W.H. and Schneider, B. (2008). The Meaning of Employee Engagement. *Industrial and Organizational Psychology*, 1(01), pp.3–30. doi:10.1111/j.1754-9434.2007.0002.x.

Mason, K.J. and Gray, R. (1999). Stakeholders in a hybrid market: the example of air business passenger travel. *European Journal of Marketing*, 33(9/10), pp.844–858. doi:10.1108/03090569910285779.

Milgram, P., Takemura, H., Utsumi, A. and Kishino, F. (1995). Augmented reality: a class of displays on the reality-virtuality continuum. *Telemanipulator and Telepresence Technologies*, 2351, pp.282–292. doi:10.1117/12.197321.

Molinillo, S., Aguilar-Illescas, R., Anaya-Sánchez, R. and Vallespín-Arán, M. (2018). Exploring the impacts of interactions, social presence and emotional engagement on active collaborative learning in a social web-based environment. *Computers & Education*, 123, pp.41–52. doi:10.1016/j.compedu.2018.04.012.

Moreira, C., Simoes, F.P.M., Lee, M.J.W., Zorzal, E.R., Lindeman, R.W., Pereira, J.M., Johnsen, K. and Jorge, J. (2023). Toward VR in VR: Assessing Engagement and Social Interaction in a Virtual Conference. *IEEE Access*, 11, pp.1906–1922. doi:https://doi.org/10.1109/access.2022.3233312.

Muhlbach, L., Bocker, M. and Prussog, A. (1995). Telepresence in Videocommunications: A Study on Stereoscopy and Individual Eye Contact. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(2), pp.290–305. doi:10.1518/001872095779064582.

Müller, A. and Wittmer, A. (2023). The choice between business travel and video conferencing after COVID-19 – Insights from a choice experiment among frequent travelers. *Tourism Management*, [online] 96(96), p.104688. doi:https://doi.org/10.1016/j.tourman.2022.104688.

Nunamaker, J.F., Dennis, A.R., Valacich, J.S., Vogel, D. and George, J.F. (1991). Electronic meeting systems. *Communications of the ACM*, 34(7), pp.40–61. doi:10.1145/105783.105793.

Oddou, G., Mendenhall, M.E. and Ritchie, J.B. (2000). Leveraging travel as a tool for global leadership development. *Human Resource Management*, 39(2-3), pp.159–172. doi:https://doi.org/10.1002/1099-050x(200022/23)39:2/3%3C159::aid-hrm6%3E3.0.co;2-j.

Panteli, N. and Dawson, P. (2001). Video conferencing meetings: Changing patterns of business communication. *New Technology, Work and Employment*, 16(2), pp.88–99. doi:10.1111/1468-005x.00079.

Panteli, N. and Dawson, P. (2001). Video conferencing meetings: Changing patterns of business communication. *New Technology, Work and Employment*, 16(2), pp.88–99. doi:https://doi.org/10.1111/1468-005x.00079.

Pauleen, D.J. and Yoong, P. (2001). Facilitating virtual team relationships via Internet and conventional communication channels. *Internet Research*, 11(3), pp.190–202. doi:https://doi.org/10.1108/10662240110396450.

Porter, C.E., Donthu, N., MacElroy, W.H. and Wydra, D. (2011). How to Foster and Sustain Engagement in Virtual Communities. *California Management Review*, 53(4), pp.80–110. doi:10.1525/cmr.2011.53.4.80.

Potter, W.J. and Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, 27(3), pp.258–284. doi:10.1080/00909889909365539.

Reidsma, D., op den Akker, R., Rienks, R., Poppe, R., Nijholt, A., Heylen, D. and Zwiers, J. (2007). Virtual meeting rooms: from observation to simulation. *AI & SOCIETY*, 22(2), pp.133–144. doi:10.1007/s00146-007-0129-y.

Robertson-Smith, G. and Markwick, C. (2009). *Employee engagement : a review of current thinking.* Institute For Employment Studies, pp.5–21.

Roby, H. (2014). Understanding the development of business travel policies: Reducing business travel, motivations and barriers. *Transportation Research Part A: Policy and Practice*, 69, pp.20–35. doi:https://doi.org/10.1016/j.tra.2014.08.022.

Rogerson-Revell, P. (2008). Participation and performance in international business meetings. *English for Specific Purposes*, 27(3), pp.338–360. doi:10.1016/j.esp.2008.02.003.

Rosen, E. (2020). *Will Business Travel Rebound Post-COVID-19?* [online] Condé Nast Traveler. Available at: https://www.cntraveler.com/story/how-covid-19-will-change-business-travel.

Roy, J. and Filiatrault, P. (1998). The impact of new business practices and information technologies on business air travel demand. *Journal of Air Transport Management*, 4(2), pp.77–86. doi:10.1016/s0969-6997(98)00009-x.

Rubinger, L., Gazendam, A., Ekhtiari, S., Nucci, N., Payne, A., Johal, H., Khanduja, V. and Bhandari, M. (2020). Maximizing virtual meetings and conferences: a review of best practices. *International Orthopaedics*, [online] 44(8), pp.1461–1466. doi:https://doi.org/10.1007/s00264-020-04615-9.

Schrepp, M., Hinderks, A. and Thomaschewski, J. (2017). Construction of a Benchmark for the User Experience Questionnaire (UEQ). *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(4), pp.40.

Schrire, S. (2004). Interaction and cognition in asynchronous computer conferencing. *Instructional Science*, 32(6), pp.475–502. doi:10.1007/s11251-004-2518-7.

Schroeder, R. (2002). *The social life of avatars : presence and interaction in shared virtual environments*. London: Springer.

Schuemie, M.J., van der Straaten, P., Krijn, M. and van der Mast, C.A.P.G. (2001). Research on Presence in Virtual Reality: A Survey. *CyberPsychology & Behavior*, [online] 4(2), pp.183–201. doi:10.1089/109493101300117884.

Schwarz, J., Marais, C.C., Leyvand, T., Hudson, S.E. and Mankoff, J. (2014). Combining body pose, gaze, and gesture to determine intention to interact in vision-based interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. doi:10.1145/2556288.2556989.

Shaik, F.F. and Makhecha, U.P. (2019). Drivers of Employee Engagement in Global Virtual Teams. *Australasian Journal of Information Systems*, 23(23). doi:https://doi.org/10.3127/ajis.v23i0.1770.

Shaw, P. and Linnecar, R. (2007). Chapter 1. Effective Engagement. In: *Business coaching: achieving practical results through effective engagement*. Southern Gate: Capstone.

Shkalenko, A.V. and Fadeeva, E.A. (2020). *Analysis of the Impact of Digitalization on the Development of Foreign Economic Activity During COVID-19 Pandemic*. [online] www.atlantis-press.com. doi:10.2991/aebmr.k.200502.197.

Short, J., Williams, E. and Christie, B. (1976). *The social psychology of telecommunications.* London; Toronto: J. Wiley.

Spurk, D. and Straub, C. (2020). Flexible employment relationships and careers in times of the COVID-19 pandemic. *Journal of Vocational Behavior,* 119(119), p.103435. doi:10.1016/j.jvb.2020.103435.

Steuer, J. (1992). Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication*, 42(4), pp.73–93. doi:10.1111/j.1460-2466.1992.tb00812.x.

Steuer, J. (1992). Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication*, 42(4), pp.73–93.

Strengers, Y. (2014). Meeting in the Global Workplace: Air Travel, Telepresence and the Body. *Mobilities*, 10(4), pp.592–608. doi:10.1080/17450101.2014.902655.

Swarbrooke, J. and Horner, S. (2012). *Business Travel and Tourism*. 1st ed. London: Routledge, pp.3–4. doi:10.4324/9780080490601.

Terry, L. and Garrison, A. (2001). *Journal of Distance Education/Revue de l'enseignement à distance (2001) Assessing Social Presence In Asynchronous Text-based Computer Conferencing*. [online] *Assessing Social Presence In Asynchronous Text-based Computer Conferencing*. Available at: https://core.ac.uk/download/pdf/58774853.pdf.

Tu, C.-H. (2001). How Chinese Perceive Social Presence: An Examination of Interaction in Online Learning Environment. *Educational Media International*, 38(1), pp.45–60. doi:10.1080/09523980010021235.

Tu, C.-H. and McIsaac, M. (2002). The Relationship of Social Presence and Interaction in Online Classes. *American Journal of Distance Education*, 16(3), pp.131–150. doi:10.1207/s15389286ajde1603_2.

Weidlich, J., Göksün, D.O. and Kreijns, K. (2022). Extending social presence theory: social presence divergence and interaction integration in online distance learning. *Journal of Computing in Higher Education*. doi:10.1007/s12528-022-09325-2.

Wiener, M. and Mehrabian, A. (1968). *Language within language : immediacy, a channel in verbal communication*. New York: Appleton-Century-Crofts.

Woodruff, P., Wallis, C.J.D., Albers, P. and Klaassen, Z. (2021). Virtual Conferences and the COVID-19 Pandemic: Are We Missing Out with an Online Only Platform? *European Urology,* 80(2), pp.127–128. doi:10.1016/j.eururo.2021.03.019.

Yoerger, M., Crowe, J. and Allen, J.A. (2015). Participate or else!: The effect of participation in decision-making in meetings on employee engagement. *Consulting Psychology Journal: Practice and Research*, 67(1), pp.65–80. doi:https://doi.org/10.1037/cpb0000029.

Appendix

Participant No.	Age	Gender	Nationality	Employment	Interview Medium
1	26-30	F	Thailand	Former CRM and Marketing Assistant Current postgraduate student (Management)	Face-to-face
2	26-30	F	Thailand	Former Project Sales Consultant Current postgraduate student (Marketing)	Face-to-face
3	21-25	M	India	Former Engineer Current postgraduate student (Management)	Video call
4	26-30	M	Singapore	Former Sales Staff Current postgraduate student (Cognitive Neuroscience)	Video call
5	21-25	F	United Kingdom	Current postgraduate student (Marketing)	Video call
6	21-25	F	Taiwan	Former International Sales Specialist Current postgraduate student (Marketing)	Video call
7	26-30	F	Japan	Former Administrative Officer/Press Secretary Current postgraduate student (Politics and IR)	Video call
8	21-25	F	South Korea	Current Front-end Developer / Manager	Audio call
9	26-30	F	South Korea	Current MICE company Manager	Audio call
10	26-30	F	South Korea	Current Insurance Manager	Audio call

Table 1. Biographical Information of Interview Participants

Researcher-Driven AI-Enabled Adaptive Student Research Experience Platform with Contextual Recommendations and Real-Time Analytics

Dr Mahira Mohamed Mowjoon, Nisha Hirani
University of Technology Sydney, Australia

Research in Progress

Abstract

Despite numerous advancements in technology and its development, research students face various challenges including knowledge updating within their respective fields, unavailability of information sources and finding topic relevance. A student-centred platform making using of AI and personalised support is a long-term solution and hence the "Adaptive Student Research Experience Platform with Contextual Recommendations and Real-Time Analytics" will be able to reduce these challenges. More common problems that research students come across are finding it hard to collaborate with others having similar interests, lacking teamwork opportunities, feeling isolated and having to spare a long time waiting for feedback on ideas. Hence this platform will be extremely helpful in creating a sense of belonging among the students. It provides a collaborative environment for group work, peer feedback, open forums and an AI assistant to answer question quickly which will reduce delays and motivate students to pay more attention to the demands of their research. This study has been conducted by using the survey-based research methodology to gather information from a group of research students. Appropriateness of features was determined through this, and it emphasised that there exists a need for features like real-time chat, peer feedback, and AI-driven suggestions. This project in progress discusses how the platform will address problems with the aim of enabling students to realise more ambitious projects, hereby enhancing productivity in research by means of easily accessible, interactive, and customized support.

Keywords: Collaboration, Sense of Belonging, Adaptivity, Analytics, Personalisation, Recommendations, Feedback.

1.0 Introduction and Theoretical Background

The research field has grown with new implemented platforms, but for the research student, the problem still remains the same which is going through vast information to finding relevant topics and the most accurate material. Students need guidance to explore new areas of study and ideas while exploring current trends. With the rapid rise in data generation, there is a need for smart and reliable platforms which can manage large data and offer students and researchers more personalised assistance in their research journey. (Ali et al., 2024).

Early-stage researchers lack the confidence to explore their ideas and feel limited by lack of resources and support available. Having trending insights, collaborative support within the platform brings down the barricades of pursuing research. With guidance and resources, students feel empowered to take on more ambitious projects, more experiments with new ideas, and further academic boundaries.

Nowadays, there is a shift in educational institutes with an increase in focus towards integrated digital platforms to help students across all different academic domains. These platforms often fall behind and do not address the basic needs of students who need insights and knowledge around latest developments in their field. Traditional systems in research areas do not support the need of personalisation. Research can be an isolating activity, especially for students who are not part of a lab group or team. With increase in complexity of challenges, interdisciplinary research is becoming more critical which can lead to missing opportunities to many innovative ideas.

A study by El-Sabagh, 2021 gave very promising results among students in the adaptive e-learning group, their engagement score was notably higher compared to students in conventional e-learning settings, as confirmed by very large effect sizes. It has been proved that in an adaptive e-learning environment, students are more engaged in developing skills, participating, performing, and becoming emotionally involved. The backbone on which this paper rests, therefore, is the premise that adaptive e-learning environments address challenges like information overload and disengagement issues found in traditional systems by taking a more customized approach to student learning. (El-Sabagh, 2021)

The adaptive learning system comes into the picture with specialised solutions catering to such help for the students in solving challenges. Such platforms make use of Artificial Intelligence and analytics for the creation of personalized experiences for its users, offering them unique content and recommendations based on interest and areas. To the research students, this kind of system can be something very valuable and streamline the entire journey of research. This will translate to increase the efficiency of the platform, since the students will have an improved experience with the service delivery by AI-driven agents.

“The Adaptive Student Research Experience Platform” addresses these needs to create a student centric platform where not only research papers but also latest innovations and topics are easily accessible. The platform will provide real time insights and information about technological advancements and trends in the research sector. One main component of the platform is AI-Driven agent, which will enable students to get clarification on queries by minimising time gaps and allowing them to focus more on the intellectual part of their research. Also, the platform is designed to grow a community where everyone can join to share knowledge, exchange ideas and build networks. This will transform research from isolated tasks into more interactive and co-operative experiences. (Khosravi et al., 2019).

2.0 Methodology

Engineering research methodologies are systematic approaches for problem-solving and knowledge discovery. Some of the key methodologies include:

2.1 Quantitative Analysis

In this research method, measurable data is used in assessing patterns and theories in any engineering studies. It typically makes use of statistical, mathematical and computational tools and data which is collected by experimentation and numerical analysis (Schoenfeld, 2020).

2.2 Qualitative Research

In Qualitative Research, it involves non-numerical data collection such as conducting interviews, observations and textual analysis. It provides a deeper understanding of complex issues which are not included in quantitative research, this allows the exploration of the use experience and human factors in engineering to ensure better design and project management (Borrego et al., 2009).

2.3 Case Studies

Making use of different data sources such as reports and interviews and carrying out an in-depth investigation of any particular engineering project refers to case studies. They are used to identify and document best practices and understanding challenges to avoid future project performance (Stake, 1995).

2.4 Action Research

This approach involves researchers collaborating with stakeholders to solve practical problems and implement changes. It's especially useful for continuous improvement in areas like engineering management, allowing real-time evaluation and adaptability (Guertler & Sick, 2024).

2.5 Survey-based study

In a survey-based methodology, data collections from a group of people using structured questionnaires to identify the need, trends, preferences and behaviours are conducted. In the field of engineering, surveys are widely used to study the views of stakeholders in satisfaction. This approach is particularly suitable for constructing a broad perspective on an issue where human input is sought, such as in technology adoption, need identification or understanding the risks in engineering projects (Hassine & Amyot, 2016).

In this project, we have used Survey-based study methodology. We have identified a survey-based approach because, with it, we can draw direct insights/feedback/suggestions from the research student participants, which is helpful in understanding their opinions, behaviours, and experiences. It is a very efficient method that is used to collect responses from a large group in a relatively short period of time. This gave us the flexibility to gather both quantitative data for statistical analysis and qualitative insights into deeper understanding. The structured nature of a survey would mean that even the responses are standardized, hence making analysis easier and more reliable. Surveys are affordable, especially online, which enables me to reach a greater and more diverse audience-a reason this will be an ideal choice to capture a broad range of perspectives relevant to my research (Raj & Renumol, 2024).

We shared the surveys among research students in Sydney through the creation of a Google Form that would help in organized data collection. We, thereafter, forwarded the link via WhatsApp to get to them directly for easy access. Connecting through college groups and mutual friends helped us in collecting their contact details to identify a relevant group for the purpose of our research. This approach allowed us to engage a diverse set of students and thus collect valuable insights from the targeted audience.

Data Collection

For this project in progress, the most important basis of data collection is a research student survey form. The survey is designed in such a way that it can capture the current issues, platforms used and then the future features required for them. It also collects information from the students regarding their experiences while doing research projects. Additionally, open-ended questions will provide participants with opportunities to give more suggestions on any additional features to be included in the platform.

Objectives of the Survey

- To gather first-hand feedback from various research participants about their preferences and past experiences while performing research.
- Understand the pain points faced by the students while research and how some features included in the platform might help make it easier for them.
- Recognize patterns and preferences that can lead to enhancing the adaptive student research experience.

Data Collection Method

This survey was done online to ensure that maximum number of students have reach and convenience for filling. The questionnaires were forwarded to a wide range of research students across different departments at UTS. By analysing data gathered in this survey, we will have a complete picture of how the collaborative features on the platform, like peer feedback are influencing the students both in their research experience and in creating community. The human participants in this study will largely involve undergraduate and postgraduate research students hailing from diverse disciplines at UTS.

Nature of the Questions

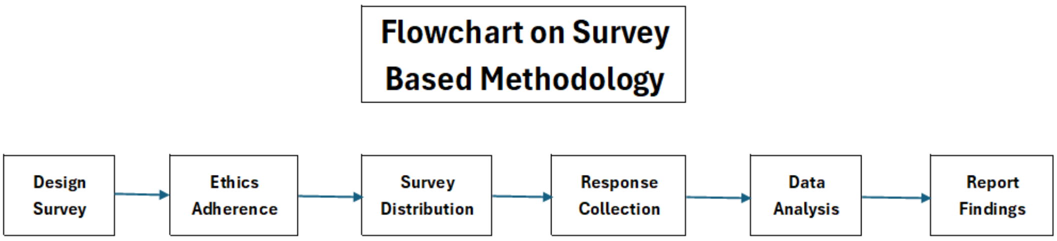
- Quantitative Questions: Likert scale or multiple-choice questions about engagement, ease of use, and effectiveness of the platform, for example, "How helpful would a research collaboration platform be to you?
- Qualitative Questions: Open-ended questions to get at detailed feedback as regards students' research behaviors, needs, and proposed areas of change, such as " What features would make collaboration easier for you?

Ethical Clearance and Considerations

Participation in this survey is on a voluntary basis, and university students were informed about the purpose of the study conducted and the use of the data to be collected. Furthermore, this research will be carried out in accordance with the rules that protect the privacy of respondents and their data.

Data Analysis and Interpretation

A flowchart on the methodology of the survey:



Flowchart – Survey Based Methodology

This flowchart is a step-by-step process of how research can be conducted based on a survey for the proposed project work. The sequence would start with the designing of the survey that includes preparation of questions and their formats, followed by gaining Ethics Clearance that validates the students' ethical stands, furthering to the Distribution of the survey to the researchers, Responses Collected, Data Analysis to bring out insight, and findings Reported effectively. It works systematically, considers ethics, and thus assures reliable outcomes.

The responses to this survey will be analysed. Each question in the survey will record and summarize responses to identify an overall trend. Data from the questions will be represented by descriptive statistics, showing, with pie charts, the distribution of responses as a means of outlining visually the overview of students' perceptions.

3.0 Preliminary Implementation and Findings

3.1 Findings

Based on the conducted surveys, the following are some of the key findings identified for this project:

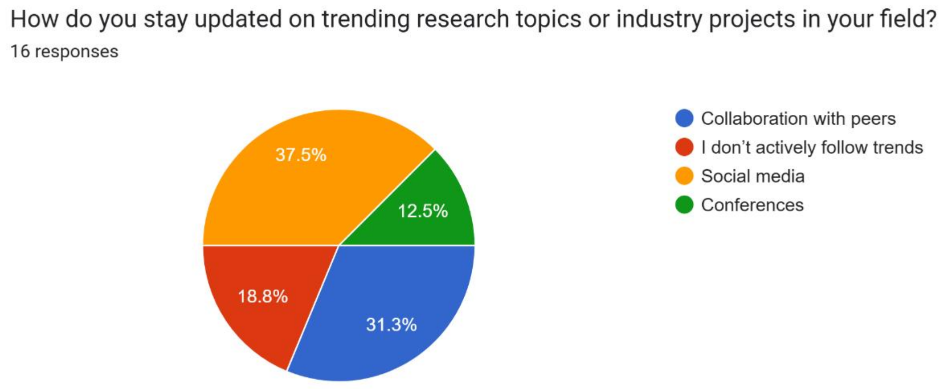


Figure 1. Methods for Staying Updated on Trending Research Topics or Industry Projects

The pie chart shows how research students keep up with their research projects or industrial projects. The majority of students accounting for almost 37.5% use social media while 31.3% get their information from active collaborations with peer networks. 18.8% of the students do not actively follow trends in the industry and just about 12.5% of them stay updated via conferences. This gives the conclusion that social media and peer collaboration is the current need of researchers. Hence a proposed feature allows fellow research students to collaborate with other researchers.

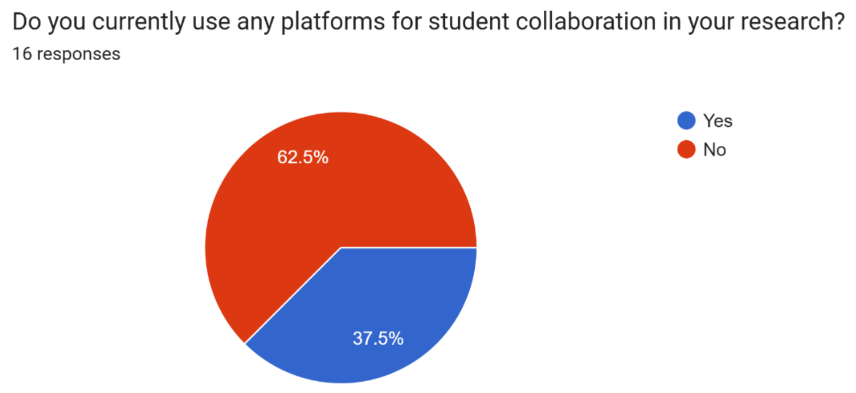
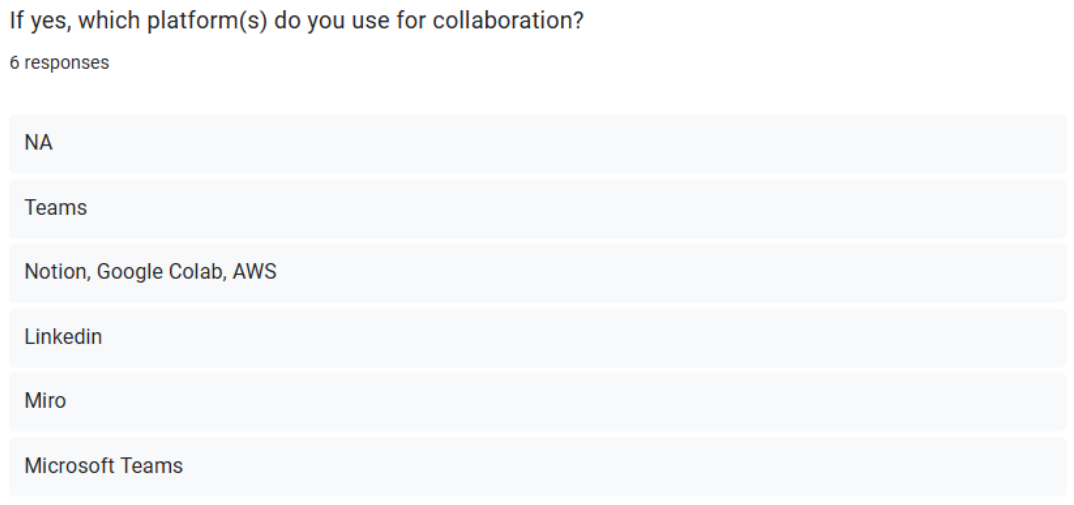


Figure 2. Use of Platforms for Student Collaboration in Research

This pie chart represents the responses related to whether the research students use any platforms for student collaboration in research. Out of the total responses, 62.5% had replied to not using any platform for collaboration (represented by the red segment), which shows students are lacking digital tools for teamwork and coordination in research. This would mean that they either prefer other ways of collaboration or are not involved in digital collaboration tools at all. On the contrary, 37.5% represented by the blue segment do use some platform. This can be indicative of an increase in the importance of online platforms for academic research collaboration.



If yes, which platform(s) do you use for collaboration?

6 responses

NA

Teams

Notion, Google Colab, AWS

Linkedin

Miro

Microsoft Teams

Figure 3. Platforms Used for Student Research and Collaboration

This data presents which platforms are currently being used by the students in their research journey. In the five responses recorded, there is a mention of Microsoft Teams, Notion, Google Collab, AWS, LinkedIn and Miro. This ranges from very popular team collaboration tools to specialized technical ones, including AWS and Google Collab.

This results tells us that a platform where all the things are available at single place is really necessary.

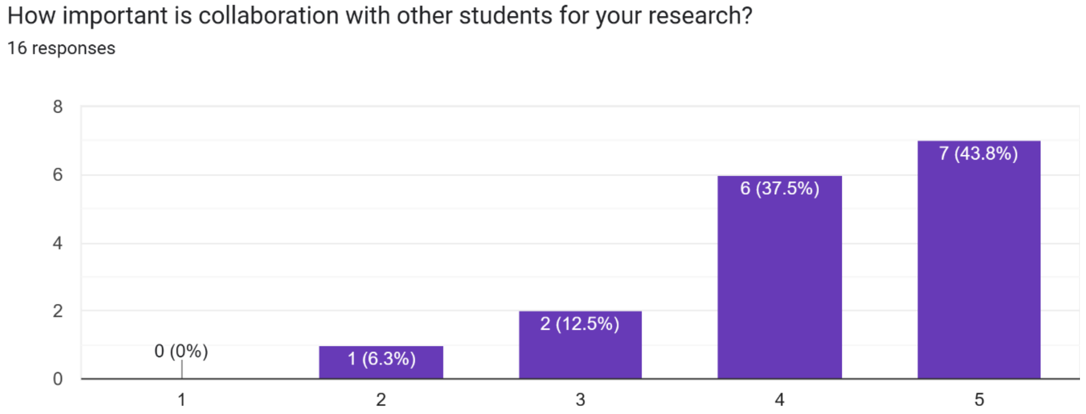


Figure 4. Importance of Peer Collaboration for Research Work

These responses are gathered to answer the question about the importance of peer collaboration while doing research projects. With 43.8% research students feel that it is extremely important to have a collaborative environment with other students to create a sense of belonging, while 67.5% of them also find it relatively essential. Less than 20% of students are more inclined towards lesser need. This question clearly answered that more than half of the population finds it really important to have teamwork given attention to.

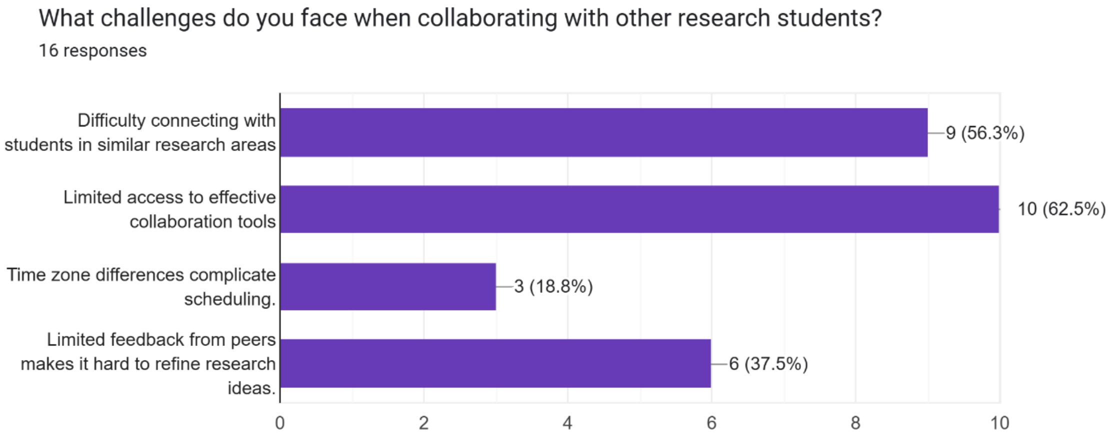


Figure 5. Collaboration challenges faced by researchers

This is a bar chart representing the key problems research students face when trying to collaborate with peers. The most frequently mentioned challenges for majority of respondents were insufficient access to collaboration tools by 62.5% and difficulty reaching others who are studying similar areas by 56.3%. Apart from this, 37.5% students found it challenging to get feedback from peers and hence making it hard to improve their work, while only 18.8% found time zone differences to complicate scheduling as a challenge.

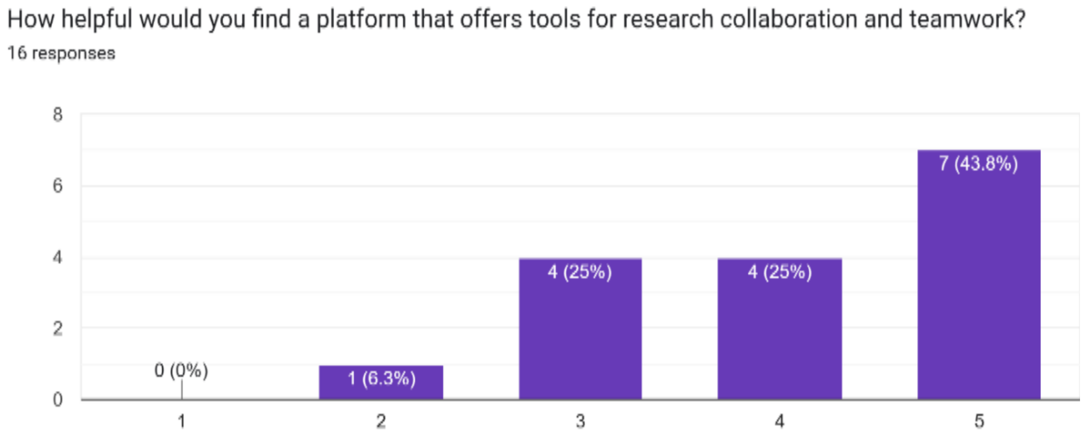


Figure 6. Helpfulness of a Research Collaboration Platform

This bar chart shows the usefulness of a platform for research collaboration. The majority of the respondents indicated that such a platform would be helpful to them. Of these, 43.8% gave it a rating of 5 out of 5, while 25% rated the helpfulness as 4 and 3 each. Only one person had chosen 2, and no person chose 1 rating. These results show that most of the respondents support the need to create a platform for research collaboration and teamwork, highlighting the access to tools needed in this area.

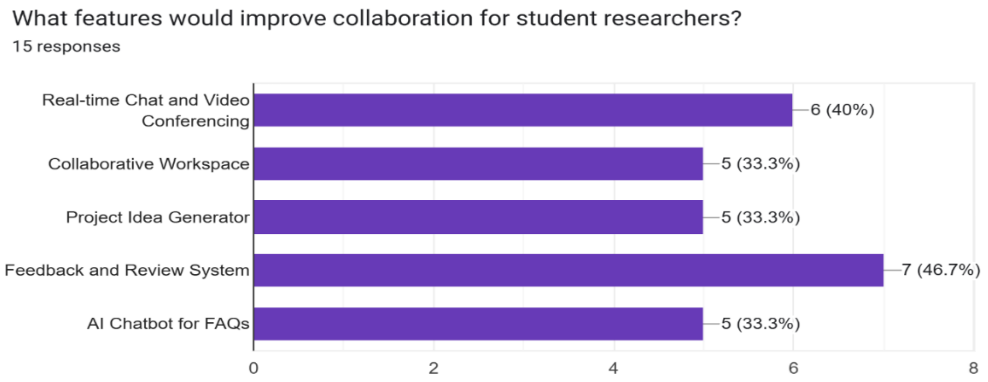
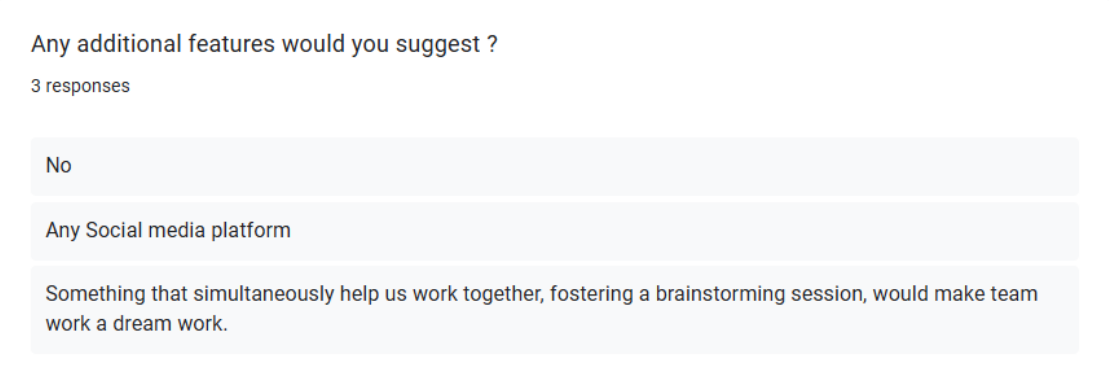


Figure 7. Features to Enhance Student Collaboration

This is a bar chart showing the features that students believe will make collaboration easier. The highest demanded feature is the feedback and review system, selected by 46.7%, hence depicting a strong need for receiving peer feedback, second is real time chat and video conferencing with 40%, showing the importance of live communication. Other features include collaborative workspace, project idea generator and AI chatbot for FAQs, each chosen by 33.3% of students. In conclusion, students value tools supporting real-time discussion and idea sharing.



Any additional features would you suggest ?

3 responses

No

Any Social media platform

Something that simultaneously help us work together, fostering a brainstorming session, would make team work a dream work.

Figure 8. Additional Features Suggested

In the last question of the survey, it asks researchers to suggest any feature that will enhance the use case of the platform. Suggestions included creation of something that simultaneously helps students to work together, have a brainstorming session and would make team work a dream work and this definitely aligned with the project in progress.

3.2 Preliminary Implementation

3.2.1 Project Lifecycle

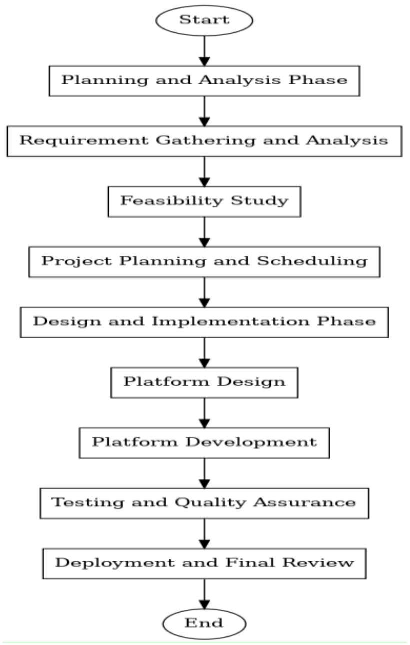

Start
Planning and Analysis Phase
Requirement Gathering and Analysis
Feasibility Study
Project Planning and Scheduling
Design and Implementation Phase
Platform Design
Platform Development
Testing and Quality Assurance
Deployment and Final Review
End

Figure 9. Project Lifecyle Flowchart

The flowchart takes the project lifecycle from "Start" to "End." Some of the major phases involved are Planning and Analysis, Requirement Gathering, Feasibility Study, Project Planning, and Design and Implementation. Each phase comprises specific activities like Platform Design, Development, Testing, and Quality Assurance, leading to Deployment and Final Review. Rectangles are employed to depict process steps, ovals for both start and end points, so that one would clearly and structurally get an overview of the project right from its very beginning up to its completion.

3.2.1.1. Planning and Analysis Phase

The identification of objectives and in-depth analysis to understand the project requirements come under the Planning and Analysis phase for this project. This phase involves all the work necessary to understand the project scope, challenges, and key deliverables to set a good foundation for effective execution.

3.2.1.2. Requirements Gathering

Identifying stakeholders like students, educators and researchers, conducting user surveys, technology adaptations, all these activities come under the requirements gathering phase of the project. In gathering the functional requirements, our focus would be essentially on how the platform shall adapt to every user's needs, what type of content it shall offer, how it can recommend personalized resources, how users can search and discover information, and what tools will be available for collaboration.

3.2.1.3. Feasibility Study

The Feasibility Study helps to check the survival of the platform by studying the technical, operational, schedule and economic factors. It checks whether the required technology is available, whether the platform can integrate with existing systems, what the costs involved are, whether it is in compliance with law, and if it's possible to make the project within the time available.

3.2.1.4. Project Planning and Scheduling Phase

The objective during the Project Planning and Scheduling of the adaptive student research platform is to logically break down the work into specific tasks, including but not limited to the development of personalised learning features, addition of content, and collaboration tools. Allocate resources, define roles, and set deadlines that would ensure the project is on track and well within the estimated completion time, all comes under this phase.

3.2.1.5. Design and Implementation Phase

The Design and Implementation stage is all about developing the detailed design based on the insights from the planning and subsequently implementing the designs. This structured approach enables us to proactively deal with various issues and provide a solution that meets functional and quality standards. The backend design of the code and how the implementation plan will roll out has to be fixed. In the progressing project, the functionality of every feature needs to be designed and implemented to meet the end goal overall.

3.2.1.6. Platform Design and Development

During this phase, the frontend of the platform should be designed and developed. All the functionalities of the platform should be in a working state and combined with the frontend developed.

3.2.1.7. Testing and Quality Assurance

Once all the above phases are successfully complete, the entirely working platform must go through numerous test cases and only when the platform passes all the cases it can be prepared to deploy. Along with tests, the quality and reusability of the code should also be confirmed.

3.2.1.8. Deployment and Final Review

This is the final phase of the project life cycle, and this stage will only be implemented when all the previous phases are finalised. After this, the deployment activities are carried out like hosting the platform, integrating with existing systems, etc. Once deployed, the research students finally can provide their reviews, feedback and suggestions post using the platform developed.

3.3 Prototype

This section includes a blueprint of how the platform might look in the next phase i.e. development phase when actual implementation of the project idea will take place.

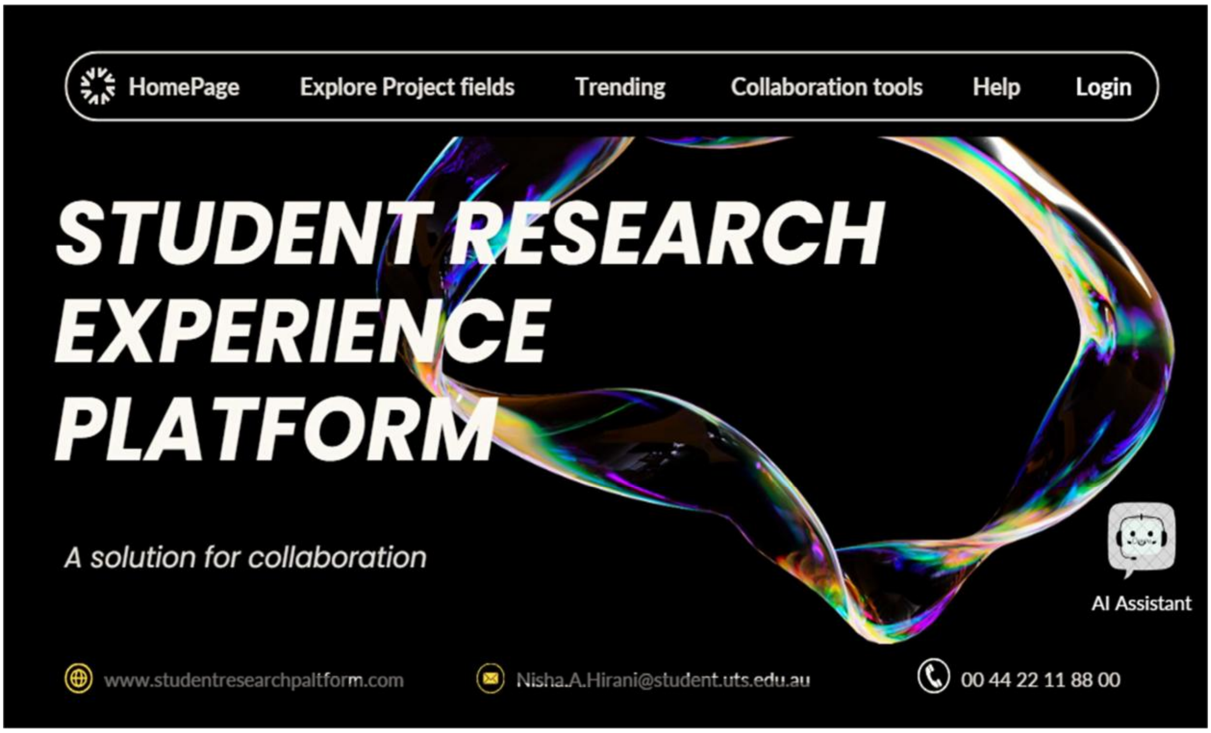
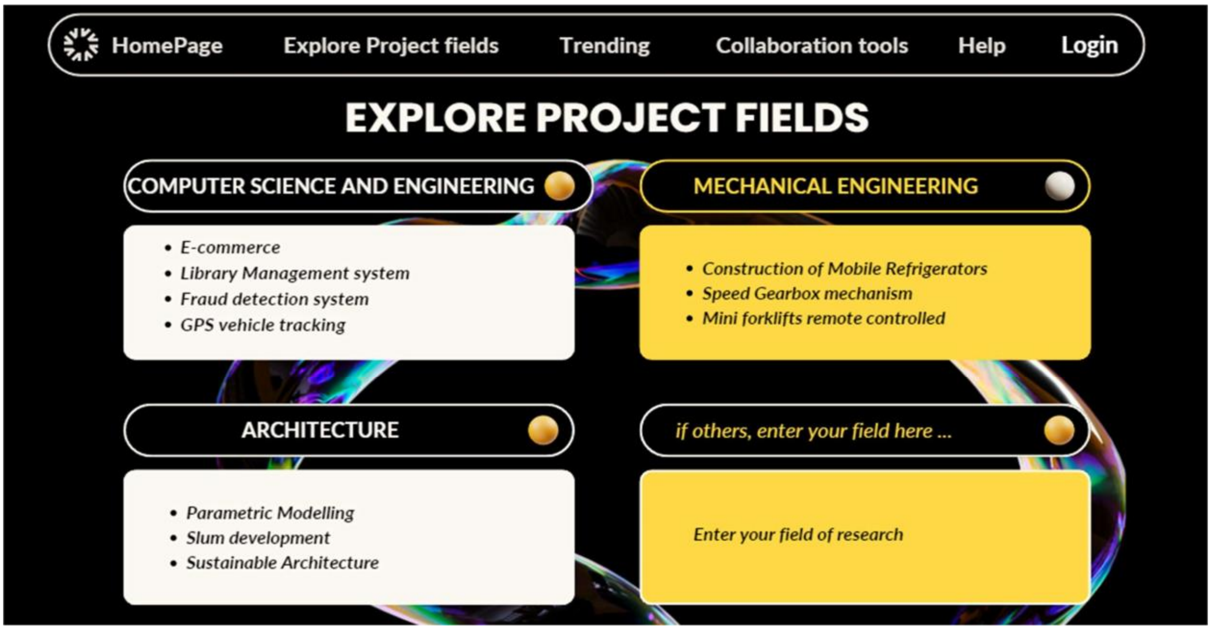

HomePage
Explore Project fields
Trending
Collaboration tools
Help
Login
STUDENT RESEARCH EXPERIENCE PLATFORM
A solution for collaboration
AI Assistant
www.studentresearchpaltform.com
Nisha.A.Hirani@student.uts.edu.au
00 44 22 11 88 00

Figure 10. 3.3.1 Landing Page of the Student Research Platform

- **3.3.1** This will be the landing page of the platform whenever a student will click on the URL of it or google it out. It would be ideal if the adaptive student research platform incorporated most of the functionalities of its landing page like project ideas based on respective fields, an AI-powered digital assistant for queries, collaboration and teamwork tools. Users will stay updated about projects that are trending in industry and search out information on support resources available at UTS. This page also includes a home page, login options, URL of the platform, and the contact customer support details like email and phone numbers. It ensures all aspects for students to access resources easily, communicate with support easily, and navigate around the platform without any issues.



EXPLORE PROJECT FIELDS

COMPUTER SCIENCE AND ENGINEERING

- *E-commerce*
- *Library Management system*
- *Fraud detection system*
- *GPS vehicle tracking*

MECHANICAL ENGINEERING

- *Construction of Mobile Refrigerators*
- *Speed Gearbox mechanism*
- *Mini forklifts remote controlled*

ARCHITECTURE

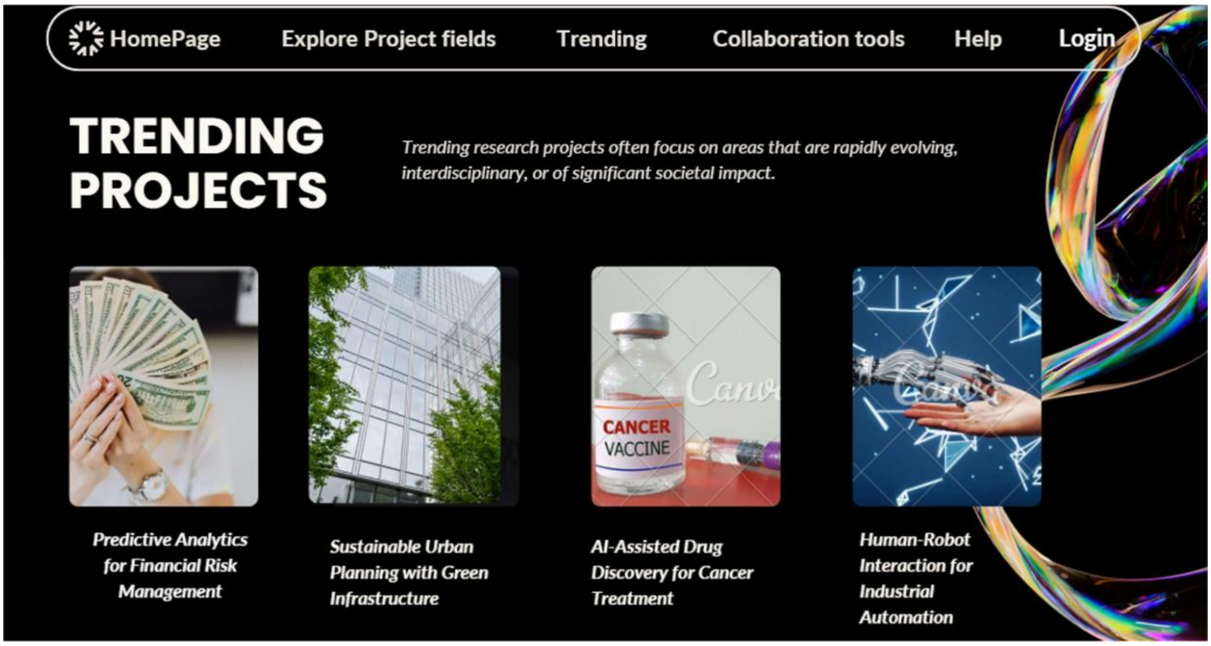
- *Parametric Modelling*
- *Slum development*
- *Sustainable Architecture*

if others, enter your field here ...

Enter your field of research

Figure 11. 3.3.2 Explore Project fields selection

3.3.2 If a student clicks on "Explore project fields," he/she will be taken to another page where they can choose a field for the interested research project. A list of the project ideas regarding the selected field then shows. The system will already include some of the most used fields. If the field that a student is interested in is not listed, they will have an option to manually search it in the system. That means, in short, this feature would provide flexibility and, at the same time, ease of use for the students to search for relevant ideas for projects or find new research fields according to interest.



TRENDING PROJECTS

Trending research projects often focus on areas that are rapidly evolving, interdisciplinary, or of significant societal impact.

Predictive Analytics for Financial Risk Management

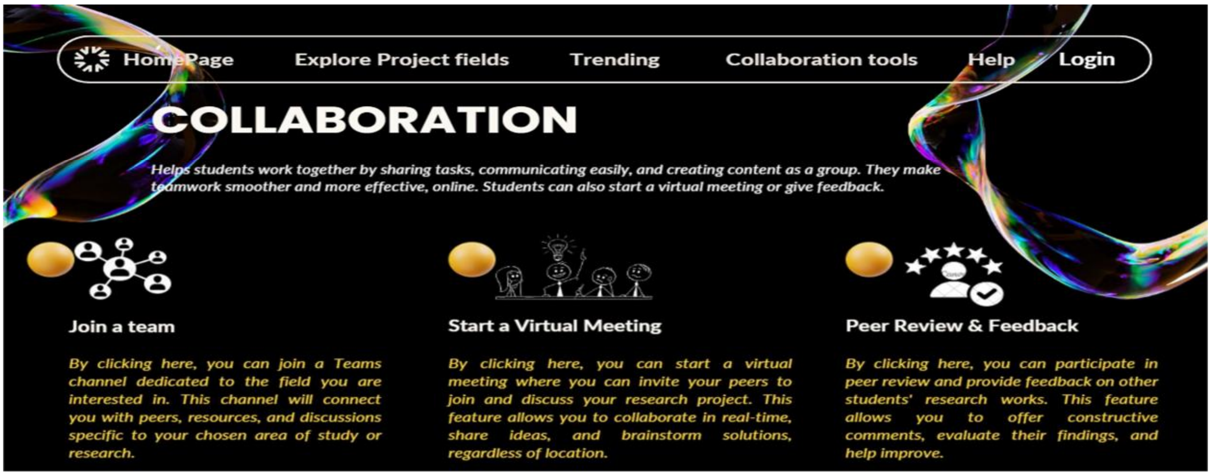
Sustainable Urban Planning with Green Infrastructure

AI-Assisted Drug Discovery for Cancer Treatment

Human-Robot Interaction for Industrial Automation

Figure 12. 3.3.3 Trending Projects Selection

- **3.3.3** The Trending Projects section summarises the latest happenings within the research world, covering the most popular and up-and-coming topics. Each student involved in research can click on any of these projects to get more detailed information about the projects, associated research papers used, technologies, publications, links to resources, among other aspects. In this case, it gives students an overview of what is new and advanced in their field and provides an in-depth idea of ongoing research. It helps them to access materials related to their work, keeps them updated on the developments happening in their profession, and provides resources useful for their own research work.



COLLABORATION

Helps students work together by sharing tasks, communicating easily, and creating content as a group. They make teamwork smoother and more effective, online. Students can also start a virtual meeting or give feedback.

Join a team

By clicking here, you can join a Teams channel dedicated to the field you are interested in. This channel will connect you with peers, resources, and discussions specific to your chosen area of study or research.

Start a Virtual Meeting

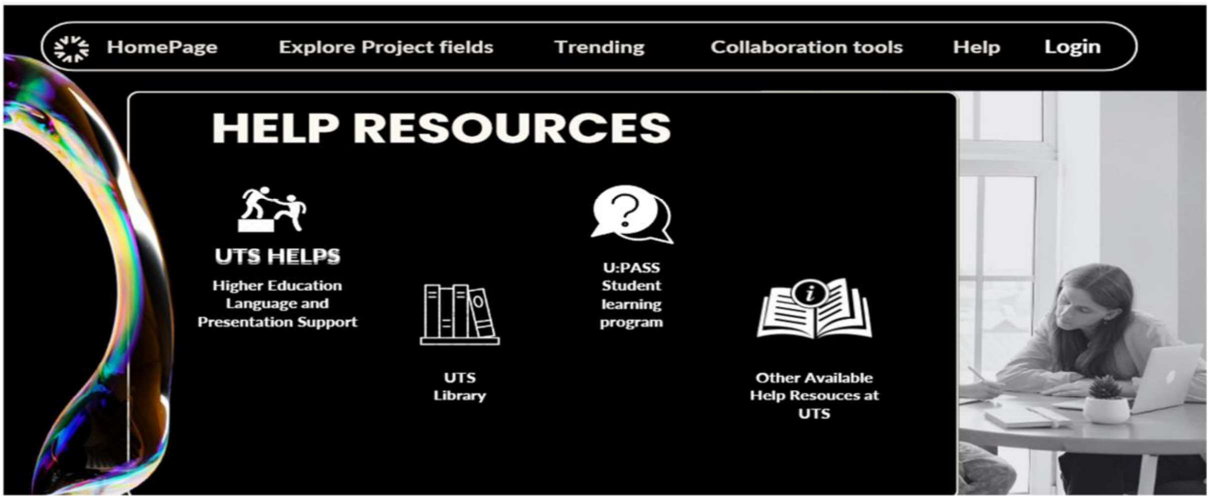
By clicking here, you can start a virtual meeting where you can invite your peers to join and discuss your research project. This feature allows you to collaborate in real-time, share ideas, and brainstorm solutions, regardless of location.

Peer Review & Feedback

By clicking here, you can participate in peer review and provide feedback on other students' research works. This feature allows you to offer constructive comments, evaluate their findings, and help improve.

Figure 13. 3.3.4 Collaboration Tools Selection

- **3.3.4** Collaboration icon allows you to explore 3 different collaboration tools. First is *JOIN A TEAM*. By clicking here students will be able to join a team where other research students from the same field can collaborate with each other. Second is *START A VIRTUAL MEETING* where peers can start a meeting and invite others to join in and provide suggestions, share ideas and open discussions. Third is *PEER REVIEW AND FEEDBACK* where students can review and provide feedback on other student's research work.



HELP RESOURCES

UTS HELPS
Higher Education Language and Presentation Support

UTS Library

U:PASS Student learning program

Other Available Help Resouces at UTS

Figure 14. 3.3.5 Help Resources Selection

- **3.3.5** The help button can be easily accessed by the students. Within this button, it finds links on going to places offering support services like: UTS HELPS, U-PASS and research materials within the UTS Library. It also gives links to other important UTS resources, which the students should know about so that they will be in good standing on how to handle university life and academics. In this regard, it is easy for students to identify and access UTS support services.

4.0 Future Research and Conclusion

This research project highlighted the blueprint for the “Adaptive Student Research Experience Platform” which is of particular interest to research students. In the future, the AI algorithms are be used in the implementation of this platform must be more advanced and refined to increase the precision and particularity of recommendations. The platform will enhance its use of AI chatbot and train the model to learn from the varied interactions performed by the students which will result in the development of better suggestions and delivery to each student’s needs.

Further research will take into consideration the ethical dimensions of AI use in the platform which focuses mainly on data privacy and security. How interdisciplinary projects can best be facilitated through the platform would underpin how flexible the platform might be across different academic fields. The relevance and usefulness of the platform will be more helpful with future versions which will adapt to the evolving needs of the researchers as they move forward in their research journeys.

Making the informed use of machine learning techniques will help to create a more responsive environment for tailored suggestions based on academic focuses. Additionally, to integrate with larger academic databases, data processing capabilities should be improved. Optimizing processing speeds and enabling the system to handle more complex research queries are another factors. Moreover, giving priority to user feedback and insights will guide appropriate improvements to make the platform more reliable. Addressing all these areas will strengthen the platform’s usage for research students and make it more effective and user-friendly.

In conclusion, this platform addresses very important issues and challenges currently faced by students and a missing sense of belonging among them. A platform offering an enabling, interactive, personalised environment can really change the student research experience. The findings of this research project provide a clear image for demand for such a platform and gives the motivation for future improvements. In the future, such platforms will be essential tool in helping students to conduct quality research, develop a sense of belonging by peer collaboration and teamwork, thus accelerating innovation in all research areas.

5.0 References

Ali, O., Murray, P. A., Momin, M., Dwivedi, Y. K., & Malik, T. (2024). The effects of artificial intelligence applications in educational settings: Challenges and strategies. Technological Forecasting and Social Change, 199, 123076. https://doi.org/10.1016/j.techfore.2023.123076

Andrews-Todd, J., & Rapp, D. (2015). Benefits, costs, and challenges of collaboration for learning and memory. Translational Issues in Psychological Science, 1(2), 182-191. https://doi.org/10.1037/tps0000025

Borrego, M., Douglas, E. P., & Amelink, C. T. (2009). Quantitative, qualitative, and mixed research methods in engineering education. Journal of Engineering Education, 98(4), 253-271. https://doi.org/10.1002/j.2168-9830.2009.tb01005.x

Falcon Editing. (n.d.). Research collaboration tools: Enhancing teamwork and productivity. Falcon Editing. https://falconediting.com/en/blog/research-collaboration-tools-enhancing-teamwork-and-productivity

Fernández-Morante, C., Cebreiro-López, B., Rodríguez-Malmierca, M. J., & Casal-Otero, L. (2022). Adaptive learning supported by learning analytics for student teachers' personalized training during in-school practices. Sustainability, 14(1), 124. https://doi.org/10.3390/su14010124

Guertler, M., & Sick, N. (2024). Action research: Combining research and problem solving for socio-technical engineering and innovation management research. CERN IdeaSquare Journal of Experimental Innovation, 8(1), 5–8. https://doi.org/10.23726/cij.2024.1509

Hassine, J., & Amyot, D. (2016). A questionnaire-based survey methodology for systematically validating goal-oriented models. Requirements Engineering, 21(3), 285–308. https://doi.org/10.1007/s00766-015-0221-7

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. Computers and Education: Artificial Intelligence, 2, 100017. https://doi.org/10.1016/j.caeai.2021.100017

Kertész, C.-Z. (2015). Using GitHub in the classroom - A collaborative learning experience. Proceedings of the International Symposium for Design and Technology in Electronic Packaging (SIITME), 381-386. https://doi.org/10.1109/SIITME.2015.7342358

Khosravi, H., Kitto, K., & Williams, J. J. (2019). RiPPLE: A crowdsourced adaptive platform for recommendation of learning activities. Journal of Learning Analytics, 6(3), 91–105. https://doi.org/10.18608/jla.2019.63.12

Raj, N. S., & Renumol, V. G. (2024). An improved adaptive learning path recommendation model driven by real-time learning analytics. Journal of

Computers in Education, 11(1), 121–148. https://doi.org/10.1007/s40692-022-00250-y

Schoenfeld, A. H. (2020). Quantitative methodologies in engineering education research. In M. A. Hjalmarson & E. C. Ferguson (Eds.), International Handbook of Engineering Education Research (pp. 641-652). Routledge.

Stake, R. E. (1995). The Art of Case Study Research. SAGE Publications.

OpenAI. (2024). *ChatGPT* (Version 4) [Large language model]. https://chat.openai.com

Tan, S. C., Lee, A. V. Y., & Lee, M. (2022). A systematic review of artificial intelligence techniques for collaborative learning over the past two decades. Computers and Education: Artificial Intelligence, 3, 100097. https://doi.org/10.1016/j.caeai.2022.100097

Vesin, B., Mangaroska, K., & Giannakos, M. (2018). Learning in smart environments: User-centered design and analytics of an adaptive learning system. Smart Learning Environments, 5(24). https://doi.org/10.1186/s40561-018-0071-0

Yang, X.-S., Koziel, S., & Leifsson, L. Þ. (2013). Computational optimization, modelling and simulation: Recent trends and challenges. Procedia Computer Science, 18, 855-860. https://doi.org/10.1016/j.procs.2013.05.250

Zawacki-Richter, O., Marín, V. I., Bond, M., & et al. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? International Journal of Educational Technology in Higher Education, 16(39). https://doi.org/10.1186/s41239-019-0171-0

El-Sabagh, H. A. (2021). Adaptive e-learning environment based on learning styles and its impact on development students' engagement. International Journal of Educational Technology in Higher Education, 18(1), 53. https://doi.org/10.1186/s41239-021-00289-4

Unlocking the potential of blockchain technology in supply chain: an exploration of critical success factors

Xiaotian Xie (Newcastle University Business school), Glenn Parry (University of Surrey), Ying Yang (Newcastle University Business school), Jiayao Hu (Newcastle University Business school) and Yan Jiang (Newcastle University Business School)

Research in progress

Abstract

Blockchain technology (BCT) has a disruptive impact on supply chain operations, offering significant potential to enhance efficiency and transparency. However, there is relatively limited research focused on exploring the critical success factors (CSFs) for BCT implementation in supply chains. This study addresses this gap by conducting a Delphi survey. Through a systematic literature review (SLR), 33 potential success factors were initially identified, of which 28 CSFs were selected in the first round of the Delphi study. In the second and third rounds, panel members ranked these CSFs and reached a consensus on their relative importance. This study serves as a valuable resource for supply chain stakeholders, providing managers with insights into the significance of each CSF. By ranking these factors, the study offers guidance for managers to optimize resource allocation in BCT implementation.

Keywords: Blockchain, Supply chain management, Success factors, Digital transformation

1. Introduction

In recent years, technologies related to Industry 4.0 have provoked a disruptive impact on supply chain. Blockchain, one of the key technologies, offers transformative solutions for supply chain models and strategies (Bhatia et al., 2024). Blockchain is a decentralised digital shared ledger that distributes data over a network (Dutta et al., 2020). Each node in this peer-to-peer network retains an updated version of a database including the complete transaction history (Scott et al., 2017). After the records are incorporated, they cannot be modified by an individual entity but are authenticated and overseen using automated and collaborative governance mechanisms (Christidis and Devetsikiotis, 2016). BCT has attracted considerable attention from supply chain managers due to its potential for disintermediation, transparency, confidentiality, security and automation in data management (Wang et al., 2021). Many organisations have started to investigate the utilisation of BCT in the supply chain, encompassing product provenance and traceability, smart contacts, supplier trust, cross-border trade and supply chain sustainability (Han and Fang, 2024, Ahmed and MacCarthy, 2023,
Mukherjee et al., 2022).

According to Gather's analysis, despite initial enthusiasm for the technology, a dearth of viable use cases resulted in 90% of blockchain-based supply chain efforts experiencing 'blockchain fatigue' by 2023 (Omale, 2019). The recent failures of large-scale blockchain projects by both Walmart Inc's (Bousquette, 2022) and a joint IBM/Maersk shipping platform (Kjærgaard-Winther, 2022) highlighted the difficulties in achieving the promised blockchain transformation. The influence of BCT on supply chains extends beyond the technological aspect since BCT represents not merely a technology but a multifaceted socio-technological phenomenon (Hastig and Sodhi, 2020). Moreover, with blockchain deployments still in the early stages, scholarly research integrating blockchain with supply chain management is still in its preliminary phase (Queiroz et al., 2019, Wang et al., 2021, Oriekhoe et al., 2024). Although several studies investigate the impact of blockchain on the supply chain, most researchers focus on conceptualising and examining blockchain possibilities and challenges for supply chains (Rejeb et al., 2021, Duan et al., 2024), this way, many enterprises do not understand how to build BCT capabilities in practice and overcome the related organizational challenges. Thus, it is important for supply chain participants to understand the critical success factors (CSF) in the adoption of BCT. The technology, organization and environment (TOE) framework has been recognised as the model that provides a comprehensive evaluation of the aspects that determine adoption (Chittipaka et al., 2023). Moreover, Blockchain as a novel technology, refers to the resources utilised by enterprises. The primary objective of enterprises is to convert these resources into capabilities (Yin and Ran, 2021). Therefore, this study uses the resources-based view and TOE framework to explore the following questions:

RQ1. What are the CSFs for implementing blockchain technology in supply chain?

RQ2. How do supply chain management practitioners perceive the CSFs in supply chains?

2. Methodology

This study used a three-phase structure to address the research issues, as illustrated in Figure 1. The initial step commences with comprehensive literature research to ascertain the preliminary success factors for the adoption of BCT in supply chain management. In the second phase, the Delphi technique is employed to identify and rank the CSF. Following the Delphi method, thematic analysis is employed to examine how supply chain managers and scholars interpret CSF.

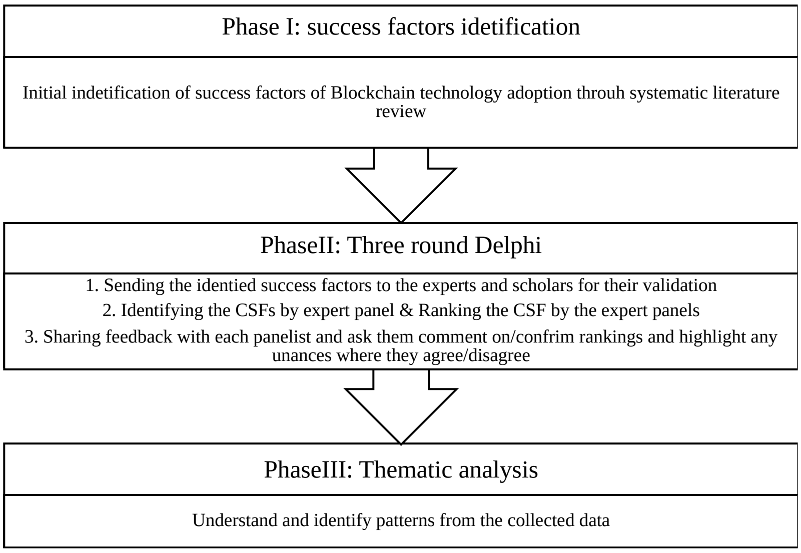

Phase I: success factors idetification
Initial indetification of success factors of Blockchain technology adoption throuh systematic literature review
PhaseII: Three round Delphi
1. Sending the identied success factors to the experts and scholars for their validation
2. Identifying the CSFs by expert panel & Ranking the CSF by the expert panels
3. Sharing feedback with each panelist and ask them comment on/confrim rankings and highlight any unances where they agree/disagree
PhaseIII: Thematic analysis
Understand and identify patterns from the collected data

Figure 1. Research design

2.1 Systematic literature review

A systematic literature review (SLR) was conducted to identify CSFs for blockchain system implementation in supply chain contexts. This uses the SLR process (Figure 2) described by Tranfield et al. (2003) and Ghadge et al. (2012). After the screening process of title /abstract monitoring, diagonal reading, and full paper reading 68 papers were selected for inclusion. A total of 33 success factors have been identified by SLR.

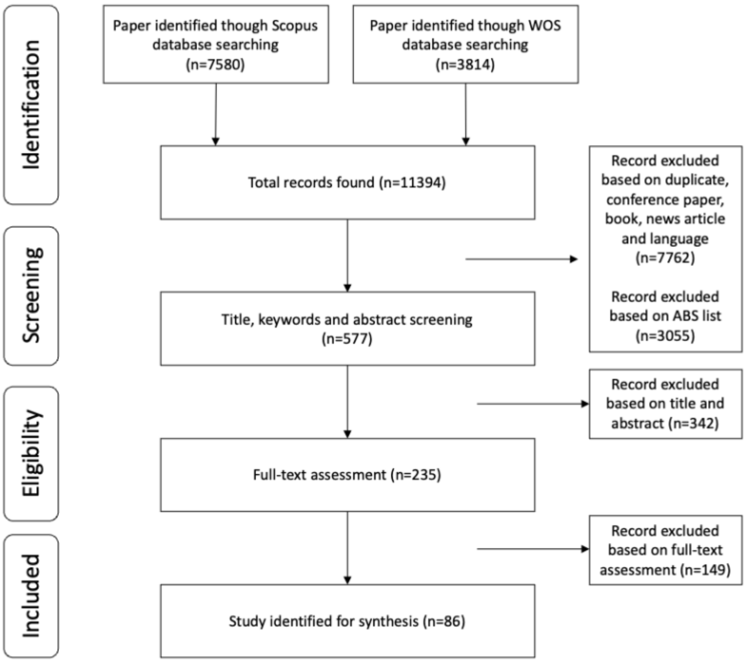

Identification
Screening
Eligibility
Included
Paper identified though Scopus database searching (n=7580)
Paper identified though WOS database searching (n=3814)
Total records found (n=11394)
Record excluded based on duplicate, conference paper, book, news article and language (n=7762)
Record excluded based on ABS list (n=3055)
Title, keywords and abstract screening (n=577)
Record excluded based on title and abstract (n=342)
Full-text assessment (n=235)
Record excluded based on full-text assessment (n=149)
Study identified for synthesis (n=86)

Figure 2. PRISMA flowchart

2.2 Delphi study

The Delphi process is a well-established tool for research, decision-making and forecasting that aims to obtain consensus on the opinions of experts, termed panel members, through a series of structured questionnaires (Durugbo et al., 2020). This approach can gather feedback on particular subjects and determine a group consensus on disagreements or unidentified problems (Saheb and Mamaghani, 2021). To ensure a representative and unbiased sample, expert selection in this study was based on clearly defined criteria. Each panel member was required to have a minimum of two years of professional experience and demonstrable expertise in both supply chain management and BCT. Additionally, efforts were made to include a diverse group of experts from various industries, academic institutions, and regulatory bodies to minimize potential bias and enhance the reliability of the consensus. This study employs three rounds of a Delphi survey with professionals in supply chain management and BCT due to the insufficient understanding of literature concerning critical success factors for the deployment of BCT in supply chains.

The goal of the first round was to generate an initial list of all factors deemed important in blockchain adoption for subsequent Delphi stages. At this stage, 15 experts are convened to examine the 33 success factors identified from the SLR. Experts were asked to confirm success factors, remove the duplicated factors, combine similar factors and modify the descriptors. The initial round resulted in 28 identified factors (Table 1) for use in the development of the questionnaire applied during the subsequent rounds of the Delphi study.

During the second round, all panel experts are requested to select the factors they consider essential for blockchain adoption in the supply chain from those recognised in the first-round poll. They are requested to rank the importance of selected factors and explain the reasons for selecting and ranking these as critical. During the third round, we compile findings and solicit feedback from panel experts to validate ranks and highlight any nuances where they agree/disagree. To further validate the findings, divergent views among panel members were carefully examined, with any persistent disagreements analysed to identify underlying reasons. By addressing contrasting perspectives, this study enhances the robustness of its results and provides a more nuanced understanding of the factors influencing blockchain implementation in supply chains.

2.3 Thematic analysis

After the Delphi process, a thematic analysis was employed to group the identified CSFs into themes. The analysis involves deciphering and interpreting common themes and then placing them in rank order of importance as defined by the experts. The analysis leads to a theoretical framework of capabilities needed for building Blockchain systems in the context of supply chains.

Group	Sub-themes
Intra-organisational Success factors	1.Senior management support 2. Basic resources (sufficient investment and time) 3. Clear vision, objective and strategy. 4. Overcome organization’s resistance and hesitation 5. Digital infrastructures 6. Understanding knowledge and potential of blockchain 7. Business process re-engineering 8. Business model aligned with blockchain capability
Inter-organizational success factors	1. Collaboration and engagement of supply chain partners and stakeholders 2. Trust 3. Information sharing 4. Cost and benefit sharing 5. System gains customers supports 6. Rivalry pressure
Factors related to Governance and Regulation	1. Rich ecosystem 2. Government policy-maker support 3. Legal and regulatory support 4. Rewards and Subsidy programs
Factors related to technology	1. Security, Privacy, and integrity 2. Scalability 3. Technology standardisation 4. Compatibility/Interoperability 5. Usability 6. Data capture and storage, access 7. Technological feasibility 8. Smart contract design
Success factors related to the people	1. BCT skilled people (technical skills) 2. BCT skilled people (managerial skills)

Table 1. List of success factors

3. Preliminary results

Preliminary findings reveal a divergence in perspectives between scholars and industry experts on the critical success factors (CSFs) of blockchain in supply chain management. Industry experts prioritize legal and regulatory support, likely due to their need to satisfy diverse stakeholder requirements, including those of customers, supply chain participants, and regulatory bodies. In globally connected supply chains, reaching consensus among participants on information access and blockchain content remains challenging (Wang et al., 2019). Furthermore, the complexity of national and international regulations, along with varying economic systems and cultural factors, leads to inconsistencies in blockchain regulations (Kopyto et al., 2020). In contrast, scholars tend to prioritize technical feasibility, focusing on exploring blockchain's capabilities and potential for innovation (Lim et al., 2021). One possible explanation for this difference is that, unlike industry practitioners, scholars often have limited direct exposure to real-world supply chain operations and regulatory constraints. Their research typically relies on theoretical models, simulations, and controlled case studies, leading them to emphasise technical feasibility as the most critical success factor.

As for the potential contributions to managerial practices, first, this research informs supply chain stakeholders and managers about the CSFs of blockchain implementation in the supply chain. Second, this study has implications for strategy in terms of the necessary capabilities required to implement BCT in the supply chain. This is provided in terms of high and low-priority CSFs, and thus the respective capabilities required. Third, this study can assist decision-makers and policymakers in understanding each CSF's significance and devising the appropriate strategies or policies to overcome the same.

4. Conclusions

The purpose of this article was to examine the CSFs for BCT implementation in the supply chain. The perspectives of professionals and scholars were employed to corroborate the critical success factors gathered from academic literature and to rank these factors. This study reveals the potential CSF for adopting blockchain in the supply chain. This study can assist decision-makers and policymakers in understanding each CSF's significance and devising the appropriate strategies or policies to overcome the same.

5. Acknowledgements

The authors gratefully acknowledge the funding contribution of the Engineering and Physical Sciences Research Council (UK) to the Next Stage Digital Economy Centre in the Decentralized Digital Economy (DECaDE) [Grant Reference EP/T022485/1].

References

AHMED, W. A. & MACCARTHY, B. L. 2023. Blockchain-enabled supply chain traceability–How wide? How deep? *International Journal of Production Economics,* 263**,** 108963.

BHATIA, M. S., CHAUDHURI, A., KAYIKCI, Y. & TREIBLMAIER, H. 2024. Implementation of blockchain-enabled supply chain finance solutions in the agricultural commodity supply chain: a transaction cost economics perspective. *Production Planning & Control,* 35**,** 1353-1367.

BOUSQUETTE, I. 2022. Blockchain fails to gain traction in the enterprise. *The Wall Street Journal*.

CHITTIPAKA, V., KUMAR, S., SIVARAJAH, U., BOWDEN, J. L.-H. & BARAL, M. M. 2023. Blockchain Technology for Supply Chains operating in emerging markets: an empirical examination of technology-organization-environment (TOE) framework. *Annals of Operations Research,* 327**,** 465-492.

CHRISTIDIS, K. & DEVETSIKIOTIS, M. 2016. Blockchains and smart contracts for the internet of things. *Ieee Access,* 4**,** 2292-2303.

DUAN, K., PANG, G. & LIN, Y. 2024. Exploring the current status and future opportunities of blockchain technology adoption and application in supply chain management. *Journal of Digital Economy*.

DURUGBO, C. M., AL-BALUSHI, Z., ANOUZE, A. & AMOUDI, O. 2020. Critical indices and model of uncertainty perception for regional supply chains: insights from a Delphi-based study. *Supply Chain Management: An International Journal,* 25**,** 549-564.

DUTTA, P., CHOI, T.-M., SOMANI, S. & BUTALA, R. 2020. Blockchain technology in supply chain operations: Applications, challenges and research opportunities. *Transportation research part e: Logistics and transportation review,* 142.

GHADGE, A., DANI, S. & KALAWSKY, R. 2012. Supply chain risk management: present and future scope. *The international journal of logistics management*.

HAN, Y. & FANG, X. 2024. Systematic review of adopting blockchain in supply chain management: bibliometric analysis and theme discussion. *International Journal of Production Research,* 62**,** 991-1016.

HASTIG, G. M. & SODHI, M. S. 2020. Blockchain for supply chain traceability: Business requirements and critical success factors. *Production and Operations Management,* 29**,** 935-954.

KJæRGAARD-WINTHER, C. 2022. AP Moller-Maersk and IBM to discontinue TradeLens, a blockchainenabled global trade platform. Maersk.

KOPYTO, M., LECHLER, S., VON DER GRACHT, H. A. & HARTMANN, E. 2020. Potentials of blockchain technology in supply chain management: Long-term judgments of an international expert panel. *Technological Forecasting and Social Change,* 161**,** 120330.

LIM, M. K., LI, Y., WANG, C. & TSENG, M.-L. 2021. A literature review of blockchain technology applications in supply chains: A comprehensive analysis of themes, methodologies and industries. *Computers & industrial engineering,* 154**,** 107133.

MUKHERJEE, A. A., SINGH, R. K., MISHRA, R. & BAG, S. 2022. Application of blockchain technology for sustainability development in agricultural supply chain: Justification framework. *Operations Management Research,* 15**,** 46-61.

OMALE, G. 2019. Gartner Predicts 90% of Blockchain-Based Supply Chain Initiatives Will Suffer 'Blockchain Fatigue' by 2023. EGHAM: Gartner.

ORIEKHOE, O. I., OMOTOYE, G. B., OYEYEMI, O. P., TULA, S. T., DARAOJIMBA, A. I. & ADEFEMI, A. 2024. Blockchain in supply chain management: a systematic review: evaluating the implementation, challenges, and future prospects of blockchain technology in supply chains. *Engineering Science & Technology Journal,* 5**,** 128-151.

QUEIROZ, M. M., TELLES, R. & BONILLA, S. H. 2019. Blockchain and supply chain management integration: a systematic review of the literature. *Supply Chain Management: An International Journal*.

REJEB, A., REJEB, K., SIMSKE, S. & TREIBLMAIER, H. 2021. Blockchain technologies in logistics and supply chain management: a bibliometric review. *Logistics,* 5**,** 72.

SAHEB, T. & MAMAGHANI, F. H. 2021. Exploring the barriers and organizational values of blockchain adoption in the banking industry. *The Journal of High Technology Management Research,* 32**,** 100417.

SCOTT, B., LOONAM, J. & KUMAR, V. 2017. Exploring the rise of blockchain technology: Towards distributed collaborative organizations. *Strategic Change,* 26**,** 423-428.

TRANFIELD, D., DENYER, D. & SMART, P. 2003. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management,* 14**,** 207-222.

WANG, Y., CHEN, C. H. & ZGHARI-SALES, A. 2021. Designing a blockchain enabled supply chain. *International Journal of Production Research,* 59**,** 1450-1475.

WANG, Y., SINGGIH, M., WANG, J. & RIT, M. 2019. Making sense of blockchain technology: How will it transform supply chains? *International Journal of Production Economics,* 211**,** 221-236.

YIN, W. & RAN, W. 2021. Theoretical exploration of supply chain viability utilizing blockchain technology. *Sustainability,* 13**,** 8231.

Experiential Learning in the Metaverse: Implications for Workplace Training

Myrto Dimitriou, Efpraxia Zamani, Mariann Hardey, Sofoklis Efraimidis
Durham University Business School (UK)- Maggioli S.p.A. (Italy), Durham University Business School (UK), Durham University Business School (UK), Maggioli S.p.A. (Greece)

Completed Research

Abstract

The Metaverse is a virtual, immersive digital world that has opened new spaces for experiential learning and workplace training. This paper examines the effectiveness of training within the Metaverse environment and its potential impact on corporate training practices, through a review of current literature and case studies. We explore the shift that the Metaverse can initiate, not only for creating immersive, interactive learning experiences, but for placing the employee at the centre of them. Our findings suggest that these environments can significantly enhance learning retention, engagement, and practical skills development positioning organizations to take a responsible and constructive approach in leading this shift. However, challenges such as technological barriers, data privacy, accessibility and content development complexity must be addressed for successful integration.

Keywords: Metaverse, experiential learning, workplace training, virtual reality, immersive technology

1.0 Introduction

Learning and development (L&D) is a key function within human resource management that is designed to enhance employees' skills, knowledge, and competencies. Organizations invest in employee development as an essential measure for creating productive and effective teams while improving workplace performance. L&D initiatives are regarded as critical to ensuring the long-term success of a company by equipping the workforce with the necessary tools to adapt and excel in dynamic environments.

Organizations allocate substantial resources to L&D, with an average expenditure of approximately $1,300 per employee (Statista, 2023). However, the return on investment (ROI) for many training methods remains limited, and in some cases, difficult to quantify. This highlights a significant challenge in evaluating the true effectiveness of current training practices, necessitating a re-evaluation of how companies can optimize their L&D strategies to maximize both employee development and organizational outcomes.

The Experiential Learning Theory (ELT) emphasizes the significance of direct experience in the learning process, proposing that learning is a cyclical process involving concrete experience (CE), reflective observation (RO), abstract conceptualization (AC), and active experimentation (AE) (Kolb, 1984).

As organizations increasingly explore virtual and hybrid work models, the potential of the Metaverse as a platform for workplace training and development has come to the forefront of academic and industry discussions. (Riches et al., 2023; Wigert et al., 2023; Bloom, 2021)

The goal of this paper is to demonstrate that the Metaverse can serve as a valuable tool for enhancing learning and training processes. It has the potential to increase the productivity of training programs while simultaneously making them more engaging for employees. Additionally, the Metaverse allows for the customization of training experiences to meet the specific needs of individual learners, placing the employee at the centre of the training process. This approach fosters a more personalized and employee-centric learning environment, ultimately leading to improved outcomes.

In a constantly evolving world, it becomes crucial to understand how this technology can be effectively leveraged to enhance learning outcomes and address the evolving needs of the modern workplace. This paper aims to address the potential of experiential learning within the Metaverse and its implications for workplace training. The research questions guiding the research are as follows:

1. How can the affordances of the Metaverse be leveraged to create immersive, experience-based learning opportunities that enhance skill development, employee engagement, and learning effectiveness in corporate training programs?
2. What are the implications of Metaverse-based experiential learning on corporate training effectiveness, considering both its potential to enhance learning outcomes and the challenges organizations must address for successful implementation?

By addressing these questions, we seek to provide insights into the potential of the Metaverse as a transformative tool for workplace learning and development.

The remainder of this paper is structured as follows: Section 2 provides a detailed definition and context for understanding the Metaverse. Section 3 describes the theoretical background, analysing the concept of experiential learning and its relevance within virtual environments. This is followed by a presentation and analysis of existing case studies that have focused on experiential learning within virtual environments, in Section 4, and a discussion about the benefits of Metaverse training programs. In Section 5 we propose the necessary steps for integration, while in Section 6 we consider the potential challenges and risks associated with implementing training programs in the Metaverse. Section 7 summarizes the key findings and Section 8 the limitations of the research, closing with Section 9 that addresses potential future research.

2.0 Metaverse, a new virtual, immersive world

The concept of the Metaverse, first coined by Neal Stephenson in his science fiction novel "Snow Crash" (Stephenson, 1992) has evolved from a fictional idea to a rapidly developing technological reality. When referring to the Metaverse we should note that is not confined to a singular definition; rather, its interpretation varies across different industries, purposes, and applications. This diversity reflects the Metaverse's broad adaptability and evolving role within different technological, social, and economic contexts.

The following description is used in the paper as a starting point of the research: The Metaverse creates a shared, immersive digital world, designed to be experienced synchronously and persistently by an effectively unlimited number of users. This environment offers new levels of immersion, control, and ownership, where physical presence is simulated through avatars. The goal of the Metaverse is to create a seamless, interoperable and immersive digital universe where interactions and transactions mirror the complexity and continuity of the real world (Ball, 2022).

The development of the Metaverse has been facilitated by the advancements in virtual reality (VR), augmented reality (AR), 3d rendering, haptic technology, 5G and edge computing. These innovations enable high degrees of interaction and immersion, allowing users to transition between real and simulated environments. This seamless integration fosters more engaging and lifelike experiences, positioning the Metaverse as a transformative platform for various applications (Dwivedi et al., 2022; Mystakidis, 2022).

The progression of these technologies is not only enhancing user experience but also encourages organizations to explore the potential of the Metaverse within their existing business models. The Metaverse is increasingly recognized as both a marketplace and a workplace, with various industries leveraging its capabilities. Sectors such as gaming and fashion were among the first to adopt it, but a growing number of sectors, including - but not limited to - education, hospitality, and tourism, are beginning to assess how they can utilise the potential of Metaverse to innovate, improve service delivery and enhance customer engagement (Buhalis et al., 2022; Mystakidis & Lympouridis, 2024).

3.0 Theoretical background and framework

The Metaverse, with its immersive and interactive nature, presents a unique environment for implementing experience-based learning strategies, enhancing employee training and development processes. This is particularly relevant in the context of the ongoing evolution of work environments, where traditional training methods may not suffice. Acquiring knowledge is inadequate for training to be considered effective (Grossman & Salas, 2011) and traditional methods often foster a passive learning environment where learners are primarily recipients of information rather than active participants in their education. Training programs should enable trainees to apply their learning effectively while at the same time strengthen their confidence in transferring skills and ensuring long-term retention of training content (Velada et al., 2007). The rigid structure with no personalised material or experience, overlooking the individual learning style or pace of learning (Hayes & Allinson, 1996) and most importantly the gap between theory and practice are some of the reasons that we are seeing a shift to training. The need for continuous learning in addition with the new skills and abilities needed, are complementing this transition, driving organizations to adopt more personalized and tailor-made training methods.

Research indicates that the Metaverse can facilitate transformative learning experiences by providing interactive and engaging training environments. Organizations should adapt their training strategies to effectively utilize collaboration tools within virtual environments, which can improve team cooperation and boost overall organizational performance (Polyviou & Pappas, 2022). Immersive technologies have significant implications for organizational learning and development, because by offering realistic simulations and interactive experiences, can greatly enhance the learning process and increase employee engagement in skill development (Dastane, 2024).

The integration of the Metaverse into training programs is not merely a theoretical proposition; practical applications have already been observed. Real-world case studies reveal how organizations have successfully leveraged the Metaverse for training, leading to enhanced individual and organizational performance (Hajjami & Park, 2023). Similarly, the potential for tacit knowledge transfer in remote workplaces via the Metaverse highlights its capability to transform traditional knowledge management practices (Lau, 2022).

Moreover, the ability to create immersive training environments in the Metaverse, aligns with the growing need for organizations to adapt to hybrid work models. (Buffer, 2023; CIPD, 2023)

This technology offers distinct advantages over traditional training methods by reshaping organizational culture and enhancing employee performance through innovative and interactive training approaches (Lim et al., 2024). VR-based simulated training within the Metaverse offers a flexible, safe, and scalable platform, particularly effective for developing both technical and interpersonal skills (Akdere et al., 2022).

3.1 Experiential Learning Theory and Virtual Environments

Kolb's Experiential Learning Theory (ELT) provides the theoretical framework for understanding how the Metaverse can facilitate effective learning experiences. Learning is a holistic process of adaptation to the world, that involves thinking, feeling, perceiving and behaving. To conceptualise it; a cyclical process that involves four stages is used: concrete experience, reflective observation, abstract conceptualization, and active experimentation (Kolb, 1984).

The key to learning lies in active involvement, and the Metaverse, with its immersive and interactive nature, presents a unique environment for implementing experience-based learning programs. Under that prism, the four stages of ELT are enhanced providing deeper engagement and extending the capabilities of traditional learning:

- Immersive Concrete Experiences: The Metaverse allows the creation of highly engaging and interactive learning experiences that simulate real-world scenarios. Participants can practice complex procedures and interactions in a safe, controlled environment, allowing them to practice realistic situations that might be too costly, dangerous, or rare to encounter in traditional training settings. The immersiveness of the experience provides a better understanding of the situation that adds to the feeling and perceiving (Bailenson, 2018).
- Enhanced Reflective Observation: The persistent nature of the Metaverse enables learners to revisit and analyse their experiences in greater detail. Participants can record their virtual interactions and review them later, having the opportunity to reflect on their performance and decision-making processes. The experience can also be shaped to the needs and interests of the employees, promoting self-awareness and critical analysis of their actions (Dominguez-Noriega et al., 2011).
- Collaborative Abstract Conceptualization: The social and collaborative aspects of the Metaverse support the AC stage of learning by allowing users to engage in group discussions fostering teamwork and communication skills (Jovanović & Milosavljević,

2022). Employees from different global locations could meet in virtual spaces to share insights and develop strategies for applying their learning to real-world scenarios. Participants can also be given the opportunity to collaborate with people that they might not have the chance to work with in the real world (Hwang & Chien, 2022).

- Active Experimentation in Virtual Environments: The Metaverse provides a safe space for learners to experiment with new ideas and behaviours allowing them to experience or observe things from different perspectives or roles. Safety training programs can allow employees to practice emergency procedures and test different approaches to hazardous situations without physical risk. This ability to "learn by doing" in a consequence-free environment encourages innovation and builds confidence in applying new skills.

4.0 Corporate Training in Virtual Environments

The Metaverse has the potential to enhance learning by offering an immersive and interactive environment that fosters learning motivation and supports the effective acquisition and retention of knowledge (Lee et al., 2021).

To provide practical insights into Metaverse applications in corporate training, we analysed nine case studies of organizations that have implemented immersive learning strategies (Table 1): Vodafone's VR Training for Field Engineers (Vodafone, 2018), Child services departments in the states of Georgia, Indiana and San Diego County (Accenture, 2022 and 2024) as well as The University of Kentucky College of Social Work (Barnes, 2023), Hilton's VR Training for Empathy and Behavioral Change (Hilton, 2020; Kover, 2020), BMW's Use of Digital Twins in Production Planning (Hecht, 2024; BMW Group, 2024), Lowe's Digital Twins for Enhanced Customer Service (Lowes, 2022), Walmart's VR-based Immersive Learning program (Incao, 2018; Goldenberg, 2023) and PWC's VR soft skills training (PWC, 2020 and 2022). The selection was guided by three primary criteria: industry diversity (highlighting patterns of innovation that transcend sector-specific constraints), varied training objectives (creating a mosaic of possibilities that can be tailored to specific needs), and impact and scalability. The following case studies do not serve as direct points of comparison but rather as pieces of a larger puzzle, illustrating a broader shift toward immersive learning strategies and Metaverse applications in workplace training. Each organization faced unique challenges and adopted distinct training methods tailored to its specific needs. By examining the 'why' behind each transformation, these cases provide insights into the driving forces behind the adoption of immersive technologies across diverse sectors.

Organization	Industry	Training Type	Key Performance Indicators (KPIs)	Outcomes	References
BMW	Automotive	Digital Twins for production planning	Efficiency, collaboration	30% increase in efficiency; enhanced team collaboration in production planning	Hecht, 2024; BMW Group, 2024
Child Services (State of Georgia, Indiana & San Diego County)	Social Services	VR for skill development in social services and child welfare	Turnover rate, confidence in decision-making	20% decrease in caseworker turnover; increased confidence and competency in child welfare investigations	Accenture, 2022 and 2024
Hilton	Hospitality	VR for corporate team members and training	Behaviour change, Training efficiency	Reduced training from 4 hours to 20 minutes; 87% behaviour change post-training	Hilton, 2020; Kover, 2020
Lowe’s	Retail	Digital Twin for customer service	Efficiency, customer satisfaction	Enhanced associate capability and customer service through interactive digital twins	Lowes, 2022
PwC	Professional Services	VR Soft Skills Training	Retention rate, engagement, adaptability	Improved soft skills retention; higher engagement; increased adaptability in various client-facing scenarios	PWC, 2020 and 2022

Organization	Industry	Training Type	Key Performance Indicators (KPIs)	Outcomes	References
University of Kentucky (CoSW)	Social Work Education	VR for Child welfare training	Competency, confidence in investigations	Increased competency and confidence in conducting child welfare investigations	Barnes, 2023
Vodafone	Telecommunications	VR for maintenance crew team managers	Training time reduction, knowledge retention	96% reduction in training time; improved skill retention	Vodafone, 2018
Walmart	Retail	VR for training and onboarding	Satisfaction rating, knowledge retention, training time	30% higher satisfaction; 70% higher content remembrance; 10-15% higher knowledge retention; training time reduced from 90 to 20 minutes	Incao, 2018; Goldenberg, 2023

Table 1. Case studies

The common thread between all the cases is the desire for improvement and enhancement, with KPIs revolving around learning, performance, and organizational effectiveness (Table 1). An overarching term that could encapsulate all the KPIs could be "employee development and performance optimization" with the connecting element being the impact of training, decision-making, and efficiency on both employees and organizational outcomes. However, achieving these required a transformation of the methods used.

The initiative for Walmart was to provide with efficient training the new employees and prepare them for difficult scenarios (like managing the customers during Black Friday) while at the same time cultivate their soft skills. These onboarding and training sessions have yielded remarkable results. Employees reported 30% higher satisfaction compared to those in traditional training, score higher on content retention tests 70% of the time and showed a 10-15% retention boost. With the VR training, the traditional 90-minute classroom session, reduced to just 20 minutes. Walmart has decided to transform the training of associates, bringing the tailor-made VR training to more than 1 million employees, in three different areas: new technology, soft skills (empathy, customer-service) and compliance.

Lowe's adopted digital twin technologies (a virtual replica of a physical asset, system, or process that is continuously updated with real-time data) and VR technologies to enhance both customer and associate experiences in their stores. The company envisions that with the use of these technologies, the expertise can be shared among customers and associates in transformative ways, enabling new modes of collaboration regardless of location. Digital twining was used to create virtual replicas of two of the stores giving the opportunity to employees to visualize and interact with store data in innovative ways. This approach allows associates to optimize store layouts, manage inventory more effectively, and provide personalized assistance to customers (Lowes Innovation Lab). The distribution and democratization of knowledge is pushing the boundaries of traditional interactions and learning, creating a new era.

Digital twining is also used by BMW, who wants to set new standards in terms of sustainability, lean processes, efficient and flexible production. The goal of the company is to make planning and simulation of all processes and the entire production system 100% virtual. This will enable collaboration in real time between different locations and across different time zones, giving a unique advantage to employees to make effective and faster decisions. Different professional groups can experiment with product development, factory planning, and the layout of the entire production flow.

In telecommunications, Vodafone wanted to reduce health and safety incidents by raising awareness about the dangers of working at heights to managers of maintenance crews that were spread across the world. The solution was to develop a VR application that allowed employees to simulate the installation and maintenance of cell phone infrastructure. The results were a 96% reduction in training time but more importantly the preparation of employees for real-world challenges without the physical risk.

PwC, understanding the importance of training leadership and soft skills, created a VR pilot to study the impact of using VR to train new managers on inclusive leadership, focusing on diversity and inclusion. Three learning programs were tested (classroom, e-learn and v-learn) with selected participants from a group of new managers in 12 locations across USA. VR training had augmented effects in comparison with the other programs in confidence, focus, speed and empathy. Participants were up to 275% more confident to act on what they learned (a 40% improvement over classroom and 35% improvement over e-learn) and up to four times more focused than e-learners. The completion time of VR training was on average four times faster than classroom and 1.5 times faster than e-learn, providing also better emotional connection.

In another field, hospitality, Hilton, wanted the corporate team members to understand the complexity and physicality of hotel operations. The use of VR training provided a valuable tool for giving them the insight needed to make better decisions. Compared with the traditional training methods, VR reduced in-class training from 4 hours to just 20 minutes, with almost 90% of employees changing their behaviour.

In the sensitive sectors of social services and social work education we have case studies demonstrating that the benefits of VR training can create positive ripple effects for employees, children, and families. The University of Kentucky College of Social Work (CoSW) has developed a new VR simulation to help current and future social workers build competency and confidence conducting child welfare investigations. Child services departments in the states of Georgia, Indiana and San Diego County implemented AVEnueS (Accenture Virtual Experience Solutions), aiming to build experience and confidence for the caseworkers, to address retention issues and error rates in eligibility decisions. The use of this technology provided a unique tool for the caseworkers to practice their skills and make eligibility decisions in a safe, supported and consequence-free environment. In Indiana the department was experiencing a turnover rate of approximately 40 percent among family case managers, costing the agency close to $72 million annually. Most departures occurred within the first two years of employment, negatively impacting open cases whose primary objective is to reunite children with their families in a safe environment. After implementing a VR solution, the caseworker turnover decreased by nearly 20%. (Stackpole, 2020)

Through analysis of the case studies, clear patterns emerge regarding the implementation and impact of virtual training environments. Companies choose between three primary models: risk-free learning environments, digital twin integration, and immersive skill development platforms (Figure 1). Each model exhibits unique characteristics that align with different aspects of Experiential Learning Theory while addressing specific organizational challenges in workforce development. It is important to note that many organizations are incorporating elements of immersive technology without explicitly label them as “Metaverse” solutions. Instead, these technologies are integrated under broader categories such as digital transformation, remote/virtual training. While the overarching term in Figure 1 is “Virtual training environments”, the applications within this field are increasingly adopting Metaverse-related features such as persistent worlds and avatar interaction.

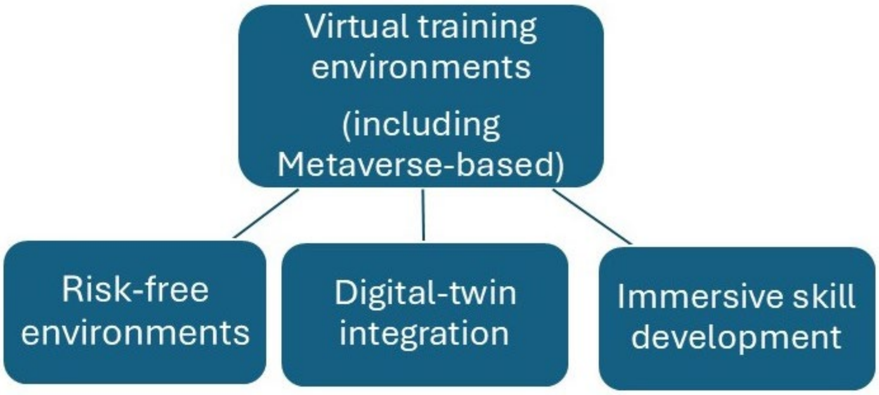

Virtual training environments
(including Metaverse-based)
Risk-free environments
Digital-twin integration
Immersive skill development

Figure 1. Virtual training environments (produced by the authors)

Risk-free learning environments serve as controlled spaces where employees can develop critical competencies without real-world repercussions, particularly valuable in high-risk sectors where practical mistakes could yield significant consequences. These environments foster Experiential Learning giving added value to each of the four stages. The concrete experience, without the agony and fear of wrongdoing, enables employees to engage with multiple and extreme scenarios (active experimentation) that would be impractical or impossible to recreate in traditional training contexts, enhancing their psychological safety. The experimentation and immediate feedback mechanisms with the possible adjustment of behaviours, supports the reflective observation and abstract conceptualization phases of Kolb's Experiential Learning Theory.

Digital twin integration, primarily used for optimizing operational efficiency through production planning and simulation, enables real-time synchronization between physical systems and their virtual counterparts, ensuring data-driven representations that adapt to actual operational conditions (Jones et al., 2020). However, certain implementations have adapted digital twins to support training objectives (Kaarlela et al., 2020). The ability to bridge abstract concepts with concrete applications, enables employees to visualize complex system interactions and understand cause-and-effect relationships at both macro and micro levels, providing another layer to conceptualization. Furthermore, the collaborative capabilities of digital twin environments facilitate social learning processes, allowing geographically dispersed teams to engage in simultaneous problem-solving and decision-making exercises, thereby enhancing both individual and collective learning outcomes (Singh et al., 2021).

Immersive skill development platforms can be used for creating trainings that focuses on both technical and interpersonal competencies. They also provide an excellent tool for understanding other roles, as shown in Hilton case study (Hilton, 2020; Kover, 2020), providing a clearer picture of the team and organization. The immersive nature promotes higher retention rates through enhanced emotional engagement and contextual learning, while enabling customizable training experiences that adapt to individual learning trajectories (Hajjami & Park, 2023). Employees can sharpen their skills, improve their decision-making and enhance their performance through these concrete experiences while at the same time the observation and conceptualization stages are intertwined.

Together, these three models represent a significant advancement in corporate training methodology, offering unprecedented opportunities for experiential learning that combines safety, efficiency, and engagement. The technology enables organizations to create what might be termed "hybrid learning ecosystems" where physical and virtual training complement each other, maximizing the benefits of both approaches while mitigating their respective limitations.

4.1 Benefits of Metaverse-Based Training Programs

The above case studies provide us with an understanding of the change that training programs in the Metaverse can bring. One of the major benefits is time reductions while maintaining or improving learning outcomes. Enhanced learning retention suggests that the immersive experiential training represents a significant advancement over traditional methodologies.

The effectiveness of these implementations is particularly evident in high-stakes environments where real-world training carries significant -physical or psychological- risks. A key finding across the implementations is that we can address both technical and soft skills development. While the importance of technical training applications is undeniable, the development of soft skills and empathy between a wide array of skills, shows the versatility of Metaverse-based learning environments. This dual capability makes this technology particularly valuable for comprehensive workforce development programs.

The immersive nature of the Metaverse significantly enhances learner engagement compared to traditional concepts, fostering deeper interaction and participation in training programs, improving knowledge retention, enhancing muscle memory and decision-making skills, while allowing for repeatable and scalable training experience (Mystakidis & Lympouridis, 2024). In addition to that, the ability to deliver training on a global scale offers substantial benefits in terms of scalability and accessibility, allowing organizations to reduce costs and logistical complexities associated with in-person training. Furthermore, this global reach enables employees from diverse geographic locations to engage in shared virtual spaces, cultivating a stronger sense of community and reinforcing organizational culture. The use of digital twins could lead to more dynamic and responsive learning environments that adapt in real-time to business needs and user performance (Dwivedi et al., 2022).

The digital infrastructure of the Metaverse enables comprehensive tracking and analysis of learner behaviours and performance metrics. Trainers can gather precise data on decision-making processes, task completion times, and error frequencies, facilitating the provision of more tailored feedback and the development of personalized learning pathways.

The flexibility of Metaverse environments allows for easy modification and customization to address varying training objectives and scenarios, ensuring adaptability to specific organizational needs. Additionally, for high-risk or complex training situations, the Metaverse provides a risk-free environment for employees to develop and practice skills through simulations, enhancing both learning outcomes and the overall training experience.

From the employee's perspective, Metaverse can serve as a unique tool that provides a learning environment for active experimentation, reducing the emotional anxiety experienced in difficult and stressful situations. The ability to immersively experience a particular scenario and replay it as needed to explore different outcomes could be invaluable, especially considering the diverse personalities of individuals. Training can be customized to meet varying expectations and target specific areas where an employee may feel less confident. This approach allows employees to tailor the training to their individual needs and pace, and, in some cases, even become co-creators of the process through feedback and engagement.

5. Steps for integration

The successful integration of Metaverse-based training in the workplace is contingent upon three critical factors: a robust technical infrastructure, alignment with organizational objectives, and effective user adoption strategies. Establishing a solid technical foundation is crucial, as it ensures the seamless operation of the virtual environments and minimizes disruptions during training. Moreover, the Metaverse training must be clearly aligned with the company's strategic goals to maximize its impact on performance and development. This alignment fosters greater organizational commitment and positions Metaverse-based training as a key driver of business outcomes.

A progressive implementation is essential (pilot programs in specific departments or for particular skills before rolling out Metaverse-based training across the entire organization), allowing organizations to evaluate the system's effectiveness and address any challenges before expanding the training across the entire workforce. Pilot programs are widely used for implementing organizational changes. Given that one of the main goals of introducing Metaverse training is to enhance effectiveness and productivity, it is crucial to place the employee at the centre of the training process. Pilot programs will help employees build familiarity with virtual tools and will address any potential challenges or knowledge gaps. Additionally, specific needs will become clearer, allowing the organization to develop effective onboarding processes, support mechanisms and instructional materials tailored to different learning paces, so that the training can be more effective and tailored to the workforce (Salas et al., 2012). This approach helps employees feel that they contributed and actively participated in shaping the training program.

A hybrid approach that combines Metaverse-based training with traditional learning methods can further enhance outcomes by leveraging the strengths of both. This blended strategy allows for greater flexibility and caters to diverse learning preferences and contexts.

Finally, continuous assessment and iteration are paramount. Organizations should regularly evaluate the effectiveness of Metaverse training through performance metrics, user feedback, and the practical application of learned skills. This ongoing process can ensure the training remains responsive to both employee needs and evolving organizational goals.

6. Challenges and Considerations

While the potential benefits of Metaverse-based learning are significant, the adoption in corporate training presents several challenges, starting from the infrastructure and resource requirements. The reliance on specialized hardware and the necessity of high-speed internet connections can create significant barriers to accessibility while also elevating implementation costs (Koohang et al., 2023). Furthermore, employees with limited exposure to immersive technologies may face a steep learning curve, experiencing discomfort or resistance when interacting with these platforms.

We should also consider that the creation and development of high-quality, interactive content for the Metaverse necessitates advanced technical skills, a process that is both time-consuming and costly. As a result, organizations will be compelled to collaborate with external developers or invest in upskilling their internal teams to meet these demands. Although different implementations have demonstrated promising outcomes, the heterogeneity of platforms- ranging from Oculus for Business to Nvidia’s Omniverse- reflects the nascent state of the industry. This diversity presents challenges for organizations in selecting and integrating the most suitable solutions for their specific needs.

The implications for employee's health and well-being should also be addressed, since prolonged exposure to immersive environments may pose health risks, such as motion sickness, digital eye strain, visual fatigue, whether the use of VR -depending on different factors- may also lead to acute stress and muscle fatigue (Souchet et al., 2023). Working, learning and interacting in a virtual workplace, where identity and presence is digitally mediated and expressed in the form of avatar, could have different impacts on employee psychology and work-life balance. The excessive use of immersive technology may lead to significant psychological and social challenges, such as dependency, increased sedentary behaviour, withdrawal from social interactions, neglect of physical health (Lim et al., 2024) and has potential for cognitive overload (Breves & Stein, 2023).

The Metaverse presents significant challenges concerning equity, diversity, and inclusion. In order to be genuinely diverse and inclusive -as envisioned-, both the technology and the organizations shaping it must be diverse, bias-free and reflective of a broad range of perspectives. Virtual workplaces should be designed to accommodate employees with disabilities, incorporating accessibility features such as customizable avatars, alternative navigation methods, and adaptive interfaces. At the same time, disparities in access to technology must be addressed, particularly for employees in remote or underserved regions who may lack the necessary resources. The implications for diversity and inclusion extend far beyond accessibility, encompassing the representation of marginalized communities within the technology sector that drives Metaverse development. Achieving a truly diverse, equitable, and inclusive Metaverse remains a complex and ongoing endeavour.

Lastly, researchers have raised critical concerns regarding data privacy and security in the Metaverse (Huang et al., 2023). The use of the technology will allow companies to access an extensive collection of personal data, including biometric, behavioural, and geolocation information, posing significant ethical and regulatory challenges. The risk of unauthorized data access, identity theft, and surveillance raises questions about user autonomy and informed consent. Additionally, issues such as harassment, cyberbullying, and digital violence within virtual environments (Lim et al., 2024) underscore the need for a legal framework to protect users. The question of legal responsibility in virtual spaces further complicates governance, particularly in cases involving avatar misconduct, accountability for harmful actions, and jurisdictional challenges in enforcing regulations across digital and physical boundaries.

While the Metaverse presents transformative opportunities for corporate training, its implementation requires careful consideration of infrastructure demands, health implications, accessibility, content development challenges, rapid technological evolution, interoperability, inclusivity, and cybersecurity. Addressing these challenges will be critical to ensuring the successful integration of Metaverse-based learning in workplace environments.

7. Conclusion

The ability of the Metaverse to offer immersive experiences, realistic simulations and collaborative interactions can revolutionize the way companies approach training and employee development. While the vision of a fully scaled and interoperable virtual world remains a future goal, it is important to note that the Metaverse is already a reality in many respects. Organizations across various industries are actively deploying immersive technologies to enhance learning and training outcomes. Even when companies do not explicitly label their applications as 'Metaverse-based,' the immersive and interactive features they incorporate are integral to its evolving ecosystem. This transitional phase highlights both the current utility and the future transformative potential of the Metaverse in corporate learning. Learning is deeply connected to human nature; our experiences serve as a medium through which we acquire knowledge. With each experience, our memories strengthen, and the knowledge we gain begins to take shape and form. In a world defined by rapid change and technological advancement, where training and upskilling have become essential, (World Economic Forum, 2024), Metaverse training offers organizations compelling reasons to embrace this transformative approach.

However, challenges such as technological barriers, user adaptation, and the complexity of content development must be carefully addressed to ensure successful implementation.

Effective integration of Metaverse learning experiences with traditional training methods and real-world applications requires a thoughtful, phased approach that combines virtual and physical learning modalities.

8. Limitations

While this paper aims to explore the richness of Metaverse possibilities for training and learning, certain limitations must be acknowledged. First, our findings are based on secondary data, which may be subject to biases, particularly in self-reported cases. Additionally, the majority of the cases analysed originate from the USA. This geographical concentration introduces a limitation in terms of cultural representation, as learning styles, epistemological approaches, and workplace training methods vary significantly across different regions. Therefore, while our findings provide a valuable foundation, future research should consider a more diverse sample to account for these variations. Despite these limitations, we believe that the findings remain relevant and significant, as they provide insight into how major companies across different industries are pioneering the adoption of immersive technologies in training. These organizations often serve as early adopters, shaping industry trends and influencing the broader adoption of innovations. While recognizing the potential for bias, we consider their experiences significant in understanding the trajectory of future workplace training transformations.

9. Future research

This paper, synthesizing current literature and case studies, has explored the potential of Experiential Learning in the Metaverse and its implications for workplace training, representing the initial phase of a broader research project. Future research will involve the collection and analyses of primary data (surveys and interviews), to empirically validate the conceptual framework and further explore the practical implications of immersive training solutions. As the Metaverse continues to evolve, further research is needed to explore its long-term impact on learning outcomes, skill retention, and overall organizational performance. Additionally, research into the psychological and cognitive effects of prolonged engagement in immersive virtual environments will be crucial for developing best practices in Metaverse-based corporate training.

Acknowledgement: The research is funded by the MSCA Doctoral Training AGORA project (Grant agreement: 101119937)

References

Accenture (2022). Caseworker training reimagined Turning virtual experiences into expertise. Available: https://www.accenture.com/us-en/case-studies/public-service/caseworker-training-reimagined (Accessed on 19.10.2024)

Accenture Avenues (2024). The Accenture Virtual Experience Solution: Immersive learning to increase empathy and reduce bias. Available: https://www.accenture.com/us-en/services/public-service/caseworker-virtual-reality (Accessed on 19.10.2024)

Akdere, M., Jiang, Y., & Lobo, F. D. (2022). Evaluation and assessment of virtual reality-based simulated training: Exploring the human–technology frontier. European Journal of Training and Development, 46(5/6), 434–449. https://doi.org/10.1108/EJTD-12-2020-0178

Ball, M. (2022). The Metaverse and how it will revolutionize everything, Liveright Publishers, p. 29

Bailenson, J. (2018). Experience on demand: What virtual reality is, how it works, and what it can do. WW Norton & Company

Barnes C. C. (2023) UK Social Work launches virtual reality child welfare investigation simulation, Available: http://uknow.uky.edu/research/uk-social-work-launches-virtual-reality-child-welfare-investigation-simulation (Accessed on 20.10.2024)

Bloom, N. (2021). Hybrid is the future of work. Stanford, CA: Stanford Institute for Economic Policy Research (SIEPR), https://siepr.stanford.edu/publications/policy-brief/hybrid-future-work.

BMW Group. This is how DIGITAL the BMW iFACTORY is, Available: https://www.bmwgroup.com/en/news/general/2022/bmw-ifactory-digital.html (Accessed on 07.11.2024)

Breves, P., & Stein, J.-P. (2023). Cognitive load in immersive media settings: The role of spatial presence and cybersickness. Virtual Reality, 27(2), 1077–1089. https://doi.org/10.1007/s10055-022-00697-5

Buffer (2023). State of remote work 2023, Available: https://buffer.com/state-of-remote-work/2023 (Accessed on 28.10.2024)

Buhalis, D., Lin, M., & Leung, D. (2022). Metaverse as a driver for customer experience and value co-creation: implications for hospitality and tourism management and marketing. International Journal of Contemporary Hospitality Management, 35(2), 701-716. https://doi.org/10.1108/ijchm-05-2022-0631

CIPD, (2023). Flexible and hybrid working pracices in 2023, Available: https://www.cipd.org/en/knowledge/reports/flexible-hybrid-working-2023/ (Accessed on 15.10.2024)

Dastane, O. (2024). Implications of metaverse, virtual reality, and extended reality for development and learning in organizations. Development in Learning Organizations an International Journal, 38(5), 27-32. https://doi.org/10.1108/dlo-09-2023-0196

Dominguez-Noriega, S., Agudo, J. E., Ferreira, P., & Rico, M. (2011). Language learning resources and developments in the Second Life metaverse. International Journal of Technology Enhanced Learning, 3(5), 496–509. https://doi.org/10.1504/IJTEL.2011.042101

Dwivedi, Y. K., Hughes, L., Baabdullah, A. M., Ribeiro-Navarrete, S., Giannakis, M., Al-Debei, M. M., Dennehy, D., Metri, B., Buhalis, D., Cheung, C. M. K., Conboy, K., Doyle, R., Dubey, R., Dutot, V., Felix, R., Goyal, D. P., Gustafsson, A., Hinsch, C., Jebabli, I., … Wamba, S. F. (2022). Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 66, 102542. https://doi.org/10.1016/j.ijinfomgt.2022.102542

Goldenberg, B. (2023) Inside Walmart's VR Training: A Case Study in Innovation, Available: https://ismguide.com/walmart-xr-metaverse-case-study/ (Accessed on 02.10.2024)

Grossman, R., & Salas, E. (2011). The transfer of training: What really matters. International Journal of Training & Development, 15(2), 103–120. https://doi.org/10.1111/j.1468-2419.2011.00373.x

Hajjami, O. and Park, S. (2023). Using the metaverse in training: lessons from real cases. European Journal of Training and Development, 48(5/6), 555-575. https://doi.org/10.1108/ejtd-12-2022-0144

Hayes, J., & Allinson, C. W. (1996). The Implications of Learning Styles for Training and Development: A Discussion of the Matching Hypothesis. British Journal of Management, 7(1), 63–73. https://doi.org/10.1111/j.1467-8551.1996.tb00106.x

Hecht, P., (2024). How BMW leverages the Industrial Metaverse, Available: https://www.t-systems.com/dk/en/insights/newsroom/expert-blogs/how-bmw-leverages-industrial-metaverse-1018168 (Accessed on 03.10.2024)

Hilton case study (2020). Building empathy to enhance hospitality, Available: https://securecdn.oculus.com/sr/oculus-business-hilton-case-study (Accessed on 19.10.2024)

Huang, Y., Li, Y. J., & Cai, Z. (2023). Security and Privacy in Metaverse: A Comprehensive Survey. Big Data Mining and Analytics, 6(2), 234–247. Big Data Mining and Analytics. https://doi.org/10.26599/BDMA.2022.9020047

Hwang, G.-J., & Chien, S.-Y. (2022). Definition, roles, and potential research issues of the metaverse in education: An artificial intelligence perspective. Computers and Education: Artificial Intelligence, 3, 100082. https://doi.org/10.1016/j.caeai.2022.100082

Incao, J. (2018). How VR is Transforming the Way We Train Associates, Available: https://corporate.walmart.com/news/2018/09/20/how-vr-is-transforming-the-way-we-train-associates (Accessed on 05.10.2024)

Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. (2020). Characterising the Digital Twin: A systematic literature review. CIRP Journal of Manufacturing Science and Technology, 29, 36–52. https://doi.org/10.1016/j.cirpj.2020.02.002

Jovanović, A., & Milosavljević, A. (2022). VoRtex Metaverse Platform for Gamified Collaborative Learning. Electronics, 11(3), Article 3. https://doi.org/10.3390/electronics11030317

Kaarlela, T., Pieskä, S., & Pitkaaho, T. (2020). Digital Twin and Virtual Reality for Safety Training. 000115–000120. https://doi.org/10.1109/CogInfoCom50765.2020.9237812

Kolb, D. (1984). Experiential Learning: Experience As The Source Of Learning And Development. In Journal of Business Ethics (Vol. 1).

Koohang, A., Ooi, K.-B., Tan, G., Al-Emran, M., Aw, E., Baabdullah, A., Buhalis, D., Cham, T.-H., Dennis, C., Dutot, V., Dwivedi, Y., Hughes, L., Mogaji, E., Pandey, N., Phau, I., Raman, R., Sharma, A., Sigala, M., & Wong, L.-W. (2023). Shaping the Metaverse into Reality: A Holistic Multidisciplinary Understanding of Opportunities, Challenges, and Avenues for Future Investigation. Journal of Computer Information Systems, 63, 1–31. https://doi.org/10.1080/08874417.2023.2165197

Kover, A. (2020). A new perspective on hospitality: How Hilton uses VR to teach empathy, Available: https://tech.facebook.com/reality-labs/2020/3/a-new-perspective-on-hospitality-how-hilton-uses-vr-to-teach-empathy/ (Accessed on 18.10.2024)

Lau, K. (2022). Rethinking the knowledge transfer process through the use of metaverse: a qualitative study of organizational learning approach for remote workplace. Presence Virtual and Augmented Reality, 31, 229-244. https://doi.org/10.1162/pres_a_00395

Lee, L.-H., Braud, T., Zhou, P., Wang, L., Xu, D., Lin, Z., Kumar, A., Bermejo, C. and Hui, P. (2021). "All one needs to know about metaverse: a complete survey on technological singularity, virtual ecosystem, and research", arXiv preprint, doi: 10.48550/arXiv.2110.05352.

Lim, D. H., Lee, J. Y., & Park, S. (2024). The Metaverse in the Workplace: Possibilities and Implications for Human Resource Development. Human Resource Development Review, 23(2), 164-198. https://doi.org/10.1177/15344843231217174

Lowes (2022). Lowe's unveils Industry-First Digital Twin, giving associates 'Superpowers' to better serve customers, Available: https://corporate.lowes.com/newsroom/press-releases/lowes-unveils-industry-first-digital-twin-giving-associates-superpowers-better-serve-customers-09-20-22 (Accessed on 02.10.2024)

Lowe's Innovation Labs, We bring the future home, Available: https://www.lowesinnovationlabs.com/ (Accessed on 09.11.2024)

Martins L. B., Wolfe S. G., (2023) Metaversed See Beyond The Hype, p126

Mystakidis, S. (2022). Metaverse. Encyclopedia, 2(1), Article 1. https://doi.org/10.3390/encyclopedia2010031

Mystakidis, S., & Lympouridis, V. (2024). Immersive Learning Design in the Metaverse: A Theoretical Literature Review Synthesis (pp. 55–71). https://doi.org/10.1007/978-981-97-1298-4_4

Polyviou, A. and Pappas, I. (2022). Chasing metaverses: reflecting on existing literature to understand the business value of metaverses. Information Systems Frontiers, 25(6), 2417-2438. https://doi.org/10.1007/s10796-022-10364-4

PwC (2022). What does virtual reality and the metaverse mean for training? Available: www.pwc.com/us/en/tech-effect/emerging-tech/virtual-reality-study.html (Accessed on 10.10.2024)

PwC (2020). PwC's study into the effectiveness of VR for soft skills training- The Effectiveness of Virtual Reality Soft Skills Training in the Enterprise, Available: https://www.pwc.co.uk/issues/technology/immersive-technologies/study-into-vr-training-effectiveness.html (Accessed on 10.10.2024)

Riches, S., Taylor, L., Jeyarajaguru, P., Veling, W., & Valmaggia, L. (2023). Virtual reality and immersive technologies to promote workplace wellbeing: a systematic review. *Journal of Mental Health*, *33*(2), 253–273. https://doi-org.ezphost.dur.ac.uk/10.1080/09638237.2023.2182428

Salas, E., Tannenbaum, S. I., Kraiger, K., & Smith-Jentsch, K. A. (2012). The Science of Training and Development in Organizations: What Matters in Practice. Psychological Science in the Public Interest, 13(2), 74–101. https://doi.org/10.1177/1529100612436661

Singh, M., Fuenmayor, E., Hinchy, E., Qiao, Y., Murray, N., & Devine, D. (2021). Digital Twin: Origin to Future. Applied System Innovation, 4, 36. https://doi.org/10.3390/asi4020036

Souchet, A. D., Lourdeaux, D., Pagani, A., & Rebenitsch, L. (2023). A narrative review of immersive virtual reality's ergonomics and risks at the workplace: Cybersickness, visual fatigue, muscular fatigue, acute stress, and mental overload. Virtual Reality, 27(1), 19–50. https://doi.org/10.1007/s10055-022-00672-0

Statista (2023). Average spend on workplace training per employee worldwide from 2008 to 2022, Available: https://www.statista.com/statistics/738519/workplace-training-spending-per-employee/ (Accessed on 10.10.2024)

Stackpole, B. (2020). How virtual reality can help improve employee retention, Available: https://mitsloan.mit.edu/ideas-made-to-matter/how-virtual-reality-can-help-improve-employee-retention (Accessed in 11.10.2024)

Stephenson, N., (1992). Snow Crash, Bantam Books, New York, NY

Velada, R., Caetano, A., Michel, J. W., Lyons, B. D., & Kavanagh, M. J. (2007). The effects of training design, individual characteristics and work environment on transfer of training. International Journal of Training and Development, 11(4), 282–294. https://doi.org/10.1111/j.1468-2419.2007.00286.x

Vodafone (2018). Vodafone – Working at Height, Available: https://makereal.co.uk/work/vodafone-working-at-height/ (Accessed on 20.10.2024)

Wigert, B., Harter, J., & Agrawal, S. (2023). The Future of the Office Has Arrived: It's Hybrid. Gallup, October 9, 2023, Available: https://www.gallup.com/workplace/511994/future-office-arrived-hybrid.aspx. (Accessed on 03.10.2024)

World Economic Forum (2024). The 2020s will be a decade of upskilling. Employers should take notice, Available: https://www.weforum.org/stories/2024/01/the-2020s-will-be-a-decade-of-upskilling-employers-should-take-notice/ (Accessed on 11.11.2024)

Uncovering Unconscious Biases in Information Systems Design: An Exploratory Study

Katie O'Reilly, Dr Stephen McCarthy & Dr Wendy Rowan

Cork University Business School, University College Cork.

Completed Research

Abstract

As technology advancements continue to transform our daily lives, it is important to acknowledge the challenges and risks they can potentially create for users. One potential source of negative impacts is the perpetuation of developers' unconscious biases within technology. Unconscious bias refers to the belief that we act upon our deeply ingrained ideas that we are unaware of, and these ideas can thus result in discriminatory behaviours towards others. However, despite the significant impacts they can have on users, research on unconscious biases within Information Systems (IS) design remains limited. Our study employed a qualitative approach to gather insights from practitioners into the different forms of unconscious bias that can affect the design process. Our findings point towards the risks associated with authority bias, blind spot bias, assumption bias, and stereotyping. Building on these insights, we present recommendations to mitigate these including diverse user engagement and contextualising the use case.

Keywords: Information Systems Design, Unconscious Bias, Diversity, Digital Inclusion, Ethics

1. Introduction

Unconscious biases are the ingrained assumptions that shape our social reality which we are mostly unaware of (Pritlove et al. 2019). These biases act as 'cognitive shortcuts,' influencing decision-making and affecting those around us without our realization. For instance, it can cause us to make judgements and assign characteristics to specific genders, ethnic groups, or age profiles through assumptions and a lack of knowledge (Lord and Taylor 2009). If left unquestioned, unconscious biases can result in discriminatory behaviours as they influence our actions in the world. Unconscious bias is not only present in humans, but we can also find unconscious bias being replicated in the design of Information Systems (IS) (Kordzadeh and Ghasemaghaei 2022). While some forms of IS have been designed to overcome human biases such as hiring algorithms, they can also fall victim to replicating patterns of bias whether this is from data sets used or human biases leaking into the design process (Pethig and Kroenung 2023). An example of IS replicating societal biases is the use of predictive policing programmes. The software, designed to predictively prevent crime, has served to further biases as the data sets used in its creation used geography as a proxy for race furthering social inequalities where minorities were more likely to be flagged (Mayer 2021). This is an example of bias affecting a user group who are unfairly targeted by the outputs of a system. Our paper focuses on IS more broadly as IS influences many different areas in our life such as in healthcare, employment and education.

To address unconscious biases, the study of ethics encompasses discussing and questioning our moral principles, what we view as morally wrong and morally right (Singer 2021). In IS design, ethics aims to investigate the impact that these technologies can have on a person's life. It can be argued that IS solutions are value-ridden as they are heavily influenced by the ethical beliefs of the designer (Kraemer, Overveld, and Peterson 2010). Ethical choices have been previously studied within medical research, it is important to expand this research to IS development and the role the developers play in the system's outputs. When researching the role developers play, it is important to consider that IS design is often a collaborative process. If developers are unaware of their own biases, it becomes challenging to expect them to recognize the influence these biases may exert. The lack of transparency in the development process of IS and within industry in turn means that there is often a lack of accountability. It is not clear to many where the decision-making occurs in IS, such as in algorithm development (Tsamados et al. 2020) as such, it can be difficult to mitigate biases in the design process. If we cannot trace the steps and stages involved in development, questions arise about who we can hold morally accountable for the technology outputs.

In this research paper, we will address the following research questions:

1. *What are different forms of unconscious bias that may be present in IS design teams?*
2. *How can these unconscious biases affect the decision-making process?*

Using one-on-one interviews with practitioners, we discuss unconscious bias at various stages of IS development. Inductive methods were used to discuss examples of unconscious bias witnessed in others. The remainder of this paper is structured as follows. Section 2 will first introduce the theoretical background, while section 3 will then introduce the research methods used for data collection. In examining the findings in section 4, the key examples of unconscious bias that were discussed in interviews will be presented. The discussion will place these findings into current research in section 5 before bringing the paper to a conclusion in section 6.

2. Background

The following section will introduce the theoretical background and provide an overview of current research on biases present in IS.

2.1. Unconscious Bias in Information Systems (IS) Design

Unconscious bias can seep into our daily lives, impacting others around us in ways we often are unaware of. Unconscious bias can affect our perception of others when we are unfairly influenced by things such as gender, age, ethnicity, and socioeconomic status, to list but a few. Stereotyping is one of the more common themes of unconscious bias. Stereotyping can be found to influence our interactions with IS, such as AI chatbots. One study found that the perceived gender of an AI chatbot influenced users' opinions of the products (Ahn, Kim, and Sung 2022). For practical products, the "male" agent was seen as more trustworthy, whereas the "female" agent was more effective for pleasure-oriented products. Another study on ride sharing apps highlighted additional forms of bias perpetuated by these apps, such as price differentials for neighbourhoods with larger non-white populations and those with higher poverty levels (Pandey and Caliskan 2021). While it is important to look within and acknowledge our own unconscious bias, it is important not to allow it to function as an excuse for discriminatory behaviour (Bourne 2019). Our unconscious biases are something we must acknowledge; understand the impact they have and work to correct them.

In recent years, IS have been used in certain tasks to help eliminate the impact that unconscious bias can have on human decision-making (Raghavan 2023). When selecting candidates for interviews, hiring algorithms are often used that scan potential candidates' CVs for compatibility and reject those that do not fit with what the company is looking for. These algorithms were introduced to help increase diversity in teams. Using these predictive hiring tools can help overcome some human biases, but they can also allow systemic and institutional biases to be replicated (Bogen and Rieke 2018). While IS can play a part in helping to eliminate human bias, questions remain as to whether they hold their own bias, and whether they reproduce existing bias? Google faced criticism in 2015 for its targeted advertising practices. A study into targeted job advertisements found that male accountants were shown high-paying executive job advertisements 1,852 times compared to female accountants who were shown these adverts 318 times (Carpenter 2015). Another issue with newer forms of technology is that they may not be inclusive in their functions. IS can serve to replicate and preserve preexisting biases if their outputs are not monitored and assessed.

2.2. Recognising Different Forms of Unconscious Bias

Unconscious bias does not affect our judgement in one uniform linear way, it can appear in many different forms. For this paper, the following forms of bias will be discussed in depth, authority bias, bias blind spot, assumption bias, stereotyping and gender bias. This is not a definitive list of biases, but due to the research, these are the ones addressed in this paper, as they were to appear most prevalent in the data analysis.

The idea that people in a position of authority make the best decisions and have the best knowledge to make these decisions is referred to as authority bias (Tansey 1998). When people in a position of authority are making decisions, we tend to be less critical of their decisions and more likely to accept what they are saying (Gültekin 2024). Prior research has found that those in authority are often viewed as having a responsibility to confront biases, and these figures also feel pressure to critique these biases (Ashburn-Nardo et al. 2020). Authority bias and its effects have been previously researched in medical studies, discussing how you differentiate between learning from someone in authority, to blindly following their guidance (Silvester 2021). Authority bias has also been researched on its effect on our perception of politicians, and how it affects our view of professors (Raviv et al. 1993). These two groups are traditionally viewed as having a higher-ranking position in society. However, in IS research there is little prior research into authority bias and the effect it has on IS design. Where authority bias exists, this can work to stiffen innovation as people fail to move from the status quo.

When we can recognize bias in others, but are unable to recognize these same biases within ourselves, this is referred to as a bias blind spot (Scopelliti et al. 2015). Bias blind spot was first discussed by Pronin, Lin, and Ross (2002). Pronin et al (2002) carried out a study to investigate the ability of participants to recognise their own biases compared to other people's biases. Three studies were conducted asking participants to compare their biases to first, the average American and then, to their classmates and finally, comparing themselves to airport travellers. This study found participants ranked themselves as less biased than others (Pronin, Lin, and Ross 2002). Our bias blind spot can influence our decision-making as we are unaware of these biases or the effect they have on how we think and process information (Ehrlinger, Gilovich, and Ross 2005). Whether we are more critical of others or tend to view ourselves in the most socially favourable way, we fail to recognize our own biases. Research shows that people who have these bias blind spots are less likely to act on methods to reduce their own biases, these biases do not come from an egotistical place but rather a lack of self-awareness (Scopelliti et al., 2015). Bias bind spots can influence IS design teams when evaluating their team members' decisions and views, while failing to recognize their own shortcomings.

When we meet a new person, we are likely to assume they are similar to us in different ways, such as their needs and wants. Assumed similarity bias can influence how we view others, and means we often assume that others hold the same beliefs, qualities and attributes as we do (Cronbach 1955). When we lack information about a person's traits, we tend to rely on assumed similarity bias, especially with less visible traits like agreeableness, conscientiousness, and neuroticism (Beer and Watson 2008). These assumptions can be both positive and negative, and can relate to prior experiences (Lord and Taylor 2009). This can affect the composition of teams that work in the development of IS. Teams that exhibit shared perspectives and viewpoints are more likely to experience enhanced problem-solving and communication dynamics within the group (Kang, Yang, and Rowley 2006). However, this can also result in not all user groups being represented in the design process. When the gender of a developer working to create an IS is the same as the user, the product created is more likely to be representative of the user even in cases of fair language use (Zabel and Otto 2021). A team's composition is important if assumed similarity bias is present, as a team made up of like-minded people that lacks diversity will only continue to perpetuate the same biases. Previous research in business studies has looked at the effect of assumed similarity bias in venture capitalist teams. When looking to invest, venture capitalists tend to choose teams who are similar to themselves in their education, age, and background (Franke et al. 2006). In enterprise crowdfunding, this can appear in a hierarchical bias where employees tend to favour those who are on the same hierarchical level that they are within the organisation (Simons, Kaiser, and Vom Brocke 2019). This can lead to stunted innovation if there is a lack of challenges for growth and development, and the same themes keep appearing. The diversity of a team member's ideas is important as it allows for innovation to flourish. Assumption bias can mean members see a lack of value in external users' ideas and needs and focus heavily on their own. Stereotyping is a commonly discussed form of bias and can be considered as assigning a false overview to a group (Blum 2004). Stereotyping can negatively affect representation and participation in the IS industry (Vainionpää et al. 2021). Low levels of female representation in industry can be linked to a lack of female role models within the industry to encourage girls to pursue careers and study Science, Technology, Engineering and Mathematics (STEM) subjects (Anderson et al. 2017). Women often rate their levels of self-efficacy as lower when compared to their male classmates (Marshman et al. 2018). Female students' lower level of self-efficacy means they are less likely to pursue careers in these subjects. Stereotypes can also affect performance as female employees feel at a higher risk of conforming to negative stereotypes about their ability to perform in the IT industry (Rheingans et al. 2018). The IT industry has historically been a male-dominated field. Stereotyping not only affects the role of women in the IT industry and how they interact with technology, but it can also affect older users. Older users are often stereotyped as having poor knowledge of technology and this can affect their likelihood of adopting new technology (Mariano et al. 2022; Mitzner et al. 2010). Stereotypes influencing IS development can negatively impact many users' relationships with technology as well as their likelihood to pursue careers in the IS field.

There have previously been attempts to link social theories to biases. Optimal Distinctiveness Theory is a social theory on the ingroup-outgroup phenomena in which it states that social identity is defined by the need of an individual to assimilate to their social group while also needing differentiation. The theory was first introduced by Dr Marilynn B. Brewer in 1991 in 'The Social Self: On Being the Same and Different at the Same Time'. This theory is often used in research surrounding stereotyping and prejudice. When we look at the effect of the over-representation of men in technology startups, we can see an example of intergroup bias affecting female founders who struggle to find the line between being like their male counterparts for legitimacy, but also finding a balance and a chance to highlight their own distinctiveness (Sperber and Linder 2023). Optimal Distinctiveness Theory can help to explain the stereotypes women in tech are faced with to overcome and challenge if they are to find a place of legitimacy within the group, but also stand out enough from their male counterparts to attract support. Social dominance theory looks at various levels within society, and conflicts that can occur (Sidanius et al. 2004). It looks at how groups organize themselves in different hierarchies based on context. Impression management theory has examined how individuals maintain others' opinions of themselves. It assumes one wants to be seen favourably by others (Rosenfeld, Giacalone, and Riordan 1994). These theories can help us better understand the emergence of biases in IS design. In particular, these theories were utilised to help us analyse the responses of participants in our study.

3. Methodology

To answer the research question, qualitative research methods were used (Smythe and Giddings 2007). This helped the researcher to develop an understanding of social contexts and people in the world around them (Kaplan and Maxwell 1994). Qualitative research helps to enhance our understanding of the research topic as it allows for an inductive approach and relies on the researcher's understanding and interpretation of the data collected. This provides a deeper understanding of the role that unconscious bias and diversity play in IS design.

Epistemology is the study of how we as humans create our knowledge (Carter and Little 2007). Qualitative research, rooted in interpretivism, aligns with epistemology by emphasizing the understanding of how knowledge is constructed through human experiences rather than through purely empirical observation. This research relies on interpretation as a method of understanding taken from qualitative research methods. Interpretivism allows for the participants to share their own lived experiences as known, and the researcher can interpret knowledge from these social realities (Avenier and Thomas 2015).

Purposive sampling was used when selecting the sample population. Purposive sampling involves deliberately picking certain participants due to qualities they possess that the researcher deems important to their study (Alkassim et al. 2016). When selecting candidates to interview, their area of employment was a crucial criterion, known as criterion sampling. The area of employment was a crucial factor, as candidates who understood IS design/development and worked within the tech industry, academia or citizen participation in technology were approached. These different areas of employment were selected as it would help the researcher to gauge the level of understanding of diversity within IS design and development and the different understandings of unconscious bias. When talking to the developers, this allowed a first-hand account of the design process, and the decisions involved within and across stages. Academics were chosen as students within IS courses are potentially the future of IS developers, so it is important to assess how unconscious bias is framed in teaching. Those working in citizen participation with technology were able to help reveal the public's levels of understanding of unconscious bias and the effect it has on users within their own work. Criterion sampling along with purposive sampling meant that there was a select pool of candidates targeted for interviewing (see Figure 1). Their years of experience working within IS development and their knowledge of the design process were invaluable to the research.

The candidates approached were selected using the researcher's own judgement, selecting participants who possess specific characteristics or experiences relevant to the research question. Candidates were approached via email where they were informed of some of the outlined information of the research and the process involved. Interested candidates were then given more information and a suitable date and time was organized. After reaching out to multiple different potential participants, twelve candidates agreed to be interviewed. Efforts were taken to strive for a gender balance in the sample, however, there was a low response rate with female participants. Out of 20 females reached out to, only 5 replied and 4 agreed to take part in interviews. This impeded the researchers' attempt to minimize gender imbalance. Overall, twelve participants took part, the participant demographics are described below.

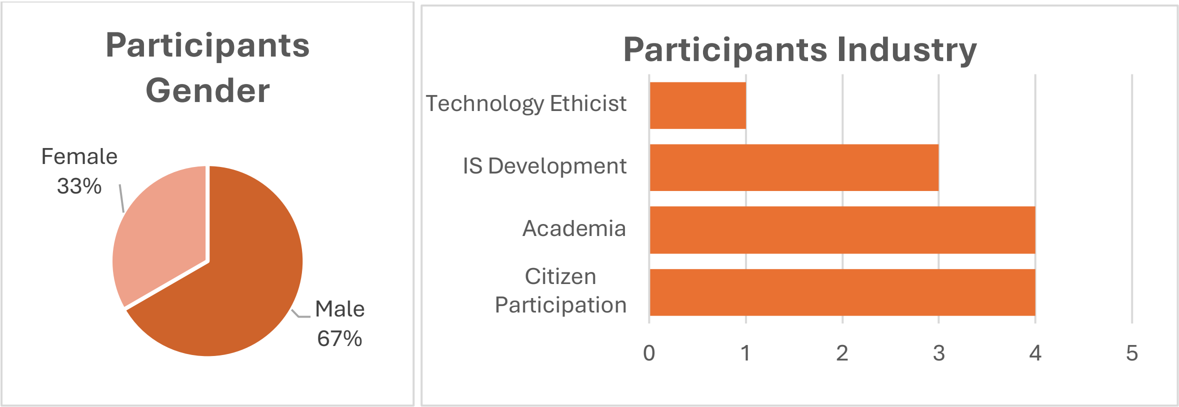


Figure 1 ***Participant Information.***

3.1. Data Collection

Semi-structured interviews were conducted to gather data from the participants, which allowed for flexibility within data collection, as the structure can be changed to suit the research process. Interviews allow for the development of a relationship between the candidate and researcher, as they allow for a free-flowing conversation (Galletta 2013). The interviews consisted of a set of predetermined questions that had been written with the participants' backgrounds and prior knowledge in mind. The questions were designed to help the researcher develop a deeper understanding of the research topic from the participants' real experiences. One-on-one interviews allow for the participants to freely share information exempt from group influence or the perceived judgment of others. The interviews were conducted over Microsoft Teams and lasted between 45 minutes to an hour. Following completion of the interviews, the recordings were transcribed and anonymised

Ethical approval to conduct this study was obtained from the relevant social research ethics committee. This was obtained prior to participant recruitment and data collection.

3.2. Data Analysis

Following data collection, thematic analysis was undertaken to group the findings. The most discussed types of unconscious bias were then selected to be discussed in the findings. Thematic analysis is a popular method for coding qualitative data and aggregating similar terms together (Terry et al. 2017). The Researcher first read through the data to familiarise themselves with it. Inductive methods of analysis were used, meaning the researcher looked for codes within the text (Azungah 2018). The key themes were defined, with existing theories and literature used to cement these definitions. The themes that were finalised were stereotyping as the most popular code, assumption bias, authority bias and bias blind spot. From reading the data and highlighting different forms of bias, different codes were selected as the most popular.

4. Findings

This section presents empirical examples of the four most commonly appearing themes in our data analysis.

4.1. Authority Bias

The first theme that appeared during data analysis was authority bias. Three participants (Participant 1, Participant 7, and Participant 5), when asked to list examples of how they first became aware of unconscious bias, used an example of authority bias to answer the question. Authority biases involve the assumption that someone in power makes the best decisions or has the best knowledge of a topic that can affect us in many areas of life. If we assume IT leaders are the most informed decision-makers who act diligently on our behalf or that our boss has the strongest decision-making and delegation skills, we can find ourselves falling victim to authority bias. *"Which is that decision-makers are the most suited to make the best decisions regarding our lives... You know decision makers. Politicians have the best diplomas. They have, you know, they have put in the most years of studies."* (*Participant 1*). Participant One highlights the example of trusting politicians as an example of authority bias in IS design decisions. Politicians hold a level of authority over the citizens much like how developers hold authority over decisions made in the design process on behalf of users. However, it is important to critically assess the decisions they are making during IS development, it is important to be aware of any influences that may have swayed their decisions.

When Siri was first introduced, there were many issues surrounding the functions programmers developed and those not included. The first beta version of Siri was able to point users to the nearest pharmacy for Viagra, but it could offer no results when asked about birth control (Alegria 2020). Apple blamed this on a simple oversight in early versions. When working with a team, authority bias can affect team members when they fail to critically analyse the team leader and their tasks and expectations. *"I would have an outside certain trust in like my superiors. I'd be like, hey, this must be right because they said it." (Participant 8).* This participant discusses this in their work environment. They highlight how often they fail to challenge or question a superior's decision simply because they believe that it must be correct, simply because of the level of authority a supervisor has over them. This can allow team members in IS development to follow their supervisors' or leaders' demands without assessing the impact the outcome could have on the user.

4.2. Bias Blind Spot

Throughout data analysis bias blind spot was another key topic. Bias blind spots appeared in conversations with participants and through the researchers' observations of participants' recognition of biases. Often, we can recognise faults in others and pick out the biases influencing their decisions and judgements, but we can overlook our own. As these biases are unconscious, we are unaware that they are influencing us *"I cannot imagine. Violence against women. Because I'm always surprised when someone, women in my surroundings from time to time tell me ... but yeah that was also a way to realize the world can be somewhat different to different people." (Participant 5)*. Participant Five discussed an example where they realized their own bias when it was pointed out to them. They could not relate to the fear their female friends and family had when travelling alone at night because to them this was not a scary process, so they did not fully empathize or understand their female friends' experiences. We can see an example of how this can affect IS design when we access the location-sharing aspect of some apps. Snapchat and Uber for example, have both come under fire for dangerous approaches to sharing users' locations (Zreik 2019). This issue is one faced predominantly by women. The participant tells how they became aware that they had this bias blind spot as they could recognize sexist attitudes in others, but only when it was clearly pointed out to them did they recognize their own.

Although this research was an explicit study of unconscious bias with a conversation with participants surrounding their own experiences and understandings of unconscious bias through deductive analysis examples of bias blind spots in some participants were becoming clear. *"It's just, ...Regardless of the solution I produce, there wouldn't be an effect on it that could be my own unconscious biases." (Participant 8)*. Participant Eight states that they have no biases that would affect their work, which is a bias blind spot. The failure to recognize or when asked to analyse your own biases is a bias blind spot. Bias blind spots would affect teams if team members failed to recognize their own biases and carry on letting them influence their decisions.

4.3. Assumption Bias

Assumption bias was a key theme that emerged during data analysis. Participants discussed the bias that can form from assuming another has the same level of knowledge and understanding as you. Participants discussed examples of this. This can affect different areas of IS development if IS designers assume users hold the same understanding as them. This may result in their products being inaccessible. '*You need someone to translate from the language of science to the language ... of the average citizens. If we want ... these people to fully comprehend what they will be using in the future? ... I guess just would citizens be able to understand really the whole process of it? Like how would we make it more understandable and transparent for everyday users? This is where I think is the biggest challenge,*' *(Participant 6)*. Participant Six shares the importance of acknowledging that users have a different level of understanding of IS compared to developers. When a developer assumes that users have the same understanding that they do of IS development, this is an example of assumption bias.

Participant Six also highlights the importance of language and how sometimes it is as simple as using more basic language to allow others to understand. This is also important in IS development; developers cannot assume that users, subject matter experts, and other developers hold the same level of knowledge as them. When making decisions, it is important to consider the people who will be using these systems and if they will have the same level of understanding as they do. '*They had also designed I don't know a guide with 20 pages or so on how you should use it install it and so and I suppose they were conscious, but it was really unconscious, nobody could understand the guide,' (Participant 3)*. Participant Three discusses an example where designers were trying to create a basic guide for the user to follow for the installation of a product. For the designers, the guide was clear and concise, however, when it was distributed to users, they were unable to follow the guide as the language and explanations used were too complex. The designers assumed that the users would hold the same levels of understanding, thus making the guide inaccessible for those without the same level of knowledge in this specific topic as the designers.

4.4. Stereotyping

Stereotyping is the most understood form of unconscious bias, which includes gender bias and classism. Stereotyping was the most widely discussed and commonly understood example throughout the data analysis. When asked to discuss the first time they learned of unconscious bias, gender stereotyping was the most common example among participants. Participants recognised how this can affect decision-making from an early age as we are conditioned by society to follow gender roles. Gender stereotyping is a generic form of unconscious bias, one participant described how at an early age we form these gendered stereotypes. *'If you know a girl is not that good in mathematics because it's just how it works. And yeah, I think all those* *and many other similar examples were the first examples of biases or stereotypes,' (Participant 2)*. These preconceived notions of gender roles can impact a person's decision-making without them realising. As previously discussed in the literature review, this can negatively influence women's role in the IS industry and the likelihood of young girls pursuing careers in IS if we associate certain school subjects and even careers with genders, this can then result in a lack of diversity within this field. Safiya Noble, author of 'Algorithms of Oppression,' discusses the bias in search engines and the effect that the results they produce have. Noble highlights how the search results for 'Black girls' which were mainly pornographic websites drastically compare to the results for 'White girls'. These biased results influence the views of those who google the terms and cause associations between the terms and the results (Noble 2018).

Stereotypes have crept into IS development. A participant shared an example of innovation catered towards men. *'Man-made technology in car security, the fact that car seats are made for men their weight is a bit bigger than women and also that they are tested with those mannequins, men mannequins and the not woman mannequins. So that's the everywhere you look you have a man-made technology made by men,' (Participant 7).* This participant shared an example of technology made with only one user group in mind, the safety of female users was not considered. Participants felt this was because the technology is often made by men who only consider their needs and wants, there is still a lack of female representation, women make up 26% of employees in computer and mathematical careers in America (DuBow and Gonzalez 2020). This can affect the chances of someone deciding to study IS fields, as they may come from a background where technology was not readily available to them during their education (Scott, Sheridan, and Clark 2014). This means that this group can go unrepresented in the IS design process. When making decisions within a team, if all members represent the same social class, there is a lack of understanding of what those in the outgroup would need or want. The ingroup's decisions can be affected by limiting stereotypes as they fail to acknowledge what is really needed.

5. Discussion

In this research, we explored the implications of unconscious biases in IS design. Bias can affect users when they are not represented in the design of the IS they use. However, despite their potential negative impacts such as marginalisation and exclusion, studies on the effects of biases have been limited in the IS field to date.

Our qualitative study addressed the research questions by identifying the types of biases discussed by participants surrounding IS design, drawing on their experiences and relevant examples. It is evident from the findings that the forms of bias highlighted by this research need to be addressed to create more inclusive IS solutions going forward. In terms of the types of biases yielded by this study, assumption bias, authority bias, bias blind spot, and stereotyping were the most frequent types of bias. These forms of bias require interventions to create more accessible IS and to minimise any future negative consequences for users. In the paragraphs that follow, we present a set of recommendations as to how developers can address these biases.

Authority bias can impact individuals and working groups when they accept what their leaders say without questioning the demands. In teams, this can have both a positive and negative effect on team members when working under a supervisor or perceived IT team leader. There is an expectation that the leader is making the correct decision, which relates back to the literature surrounding our perceived belief that a leader will recognise bias and prejudice and work to address them (Ashburn-Nardo et al. 2020). Team members can find themselves assuming that their leader is making the best decision for all, which can lead to a lack of innovation and inaccessible products if shortcomings are not picked up. This can be related to Social Dominance Theory, where the biases within a team allow for the promotion of intergroup hierarchies, meaning there are few challenges faced by leaders from those within their in-group (Hewstone, Rubin, and Willis 2001). This theory can help us to understand authority bias further. To address authority bias within teams, we recommend that *participatory dialogue* be used to support democratic decision-making between individuals at different levels of seniority. When team members feel empowered to question assumptions, they can call out biases they recognise within those in authority (Gültekin 2024). They should also be able to provide alternative suggestions to prompts given to them by those in positions of leadership to prevent authority bias from arising.

Bias blind spot was another ethical concern during IS development and refers to our inability to recognise biases within ourselves that we can see in others (Pronin, Lin, and Ross 2002). Participants discussed how they could recognize common biases in others but sometimes failed to see these biases within. Bias blind spot does not come from an egotistical place but rather from naivety (Scopelliti et al. 2015). During interviews, when the participants were asked to discuss examples of biases, they had recognised in others with whom they interacted, they produced plentiful examples of different forms of bias and different instances they had seen colleagues acting with a bias affecting their judgement. However, participants were slower to recognize any of their own biases. Interestingly, Participant Eight went as far as to state that unconscious biases could not affect their work, a key example of a bias blind spot. To mitigate this, we recommend *critical thinking tools* such as journey mapping (McCarthy et al. 2024) to ensure developers continually question their decisions to ensure that the biases they are unaware of do not affect their decision-making process. When developers have higher levels of awareness of biases, they can actively work to critically analyse their work and recognise these biases.

Assumption biases can be recognised when team members assume that the users, they are making decisions on behalf of, have the same levels of knowledge, wants, and needs as themselves (Cronbach 1955). This can result in the users' needs and wants being overlooked in the IS design process. As discussed, prior research has shown that when the gender of the developer matches the gender of the user group, then the technology is better designed for the user group (Zabel and Otto 2021). When many workers within IS development are male, this can result in technology being designed for male users (Kenny and Donnelly 2020; DuBow and Gonzalez 2020). Assumption bias means that developers could fail to consider the outgroup's needs and wants, which relates back to the Optimal Distinctiveness Theory (Brewer 1991). To help developers avoid assumption bias, we recommend that *diverse user engagement* is undertaken to guide knowledge integration in the IS design process (McCarthy et al. 2025). This promotes developers to think of their diverse user base when designing IS. It is important to ensure they are not working from the assumption that users will have the same requirements and wants as they do. This could be done through the use of personas to help developers picture future users. Personas can be used to advance social good such as increasing accessibility and encouraging cross-cultural communication (Chang, Lim, and Stolterman 2008). It is important to consider the diverse needs of users, as if development teams are made up of individuals with similar backgrounds, it can result in certain user groups being left out. If developers consider a diverse range of user personas, they are working to ensure the IS design can be more inclusive.

The most recognised form of bias was stereotyping. Participants discussed growing up with gendered roles for girls and boys, this can be linked to literature on the lack of female representation in IS development (Anderson et al. 2017). Our childhood perceptions of gender pushed onto us by society can heavily influence our careers later in life (Yansen and Zukerfeld 2014). In order to find a position of legitimacy within the group for support but also stand out from their male counterparts, women in tech must overcome and confront stereotypes, this juxtaposition can be linked to Optimal Distinctiveness Theory (Brewer 1991; Sperber and Linder 2023). Many development teams see a lack of female representation in the workplace, which can consequently lead to the creation of IS that fail to adequately represent female users. This can affect team composition as well as the recognition of biases of team members if there is a lack of representation within the group. We recommend *contextualising the use case* to encourage developers to consider many diverse scenarios when designing IS. Scenarios can serve as work-oriented design objects to ensure developers are user focused and address stereotypes that may seep in during the design process (Rosson and Carroll 2007). Developers should consider diverse user scenarios to help them identify and prevent potential stereotypes that could appear.

Table 1 summarises the recommendations emanating from this exploratory study. These key takeaways will be used to assemble an artefact as part of the design science phase of the research and to continue to address the issue of unconscious bias in IS design and development.

Recommendations	**Description**
1. Participatory dialogue	IS design teams should promote inclusive group dialogue to prevent authority bias.
2. Critical thinking tools	IS design teams should raise critical questions that address bias blind spots.
3. Diverse user engagement	IS design teams should visualise different personas to challenge assumed similarity bias.
4. Contextualising the use case	IS design teams should work with diverse scenarios of IS usage to tackle stereotype bias.

Table 1. Recommendations to mitigate unconscious bias.

In terms of future research, the intention is to take a deeper look into the forms of bias present in IS design and identify strategies to mitigate the impact of unconscious bias on the user. This includes further research into unconscious bias in IS design and development teams with the aim of using design science research to create a self-assessment rubric.

6. Conclusion

In this paper, we investigated the different types of unconscious biases present in IS design through qualitative interviews. The findings assessed different examples of bias from participants who worked in IS design and development, academia, and citizen participation. Through the participants' shared experiences, the researchers were able to highlight the four most discussed examples of unconscious bias: authority bias, assumption bias, bias blind spot, and stereotypes. These forms of bias affect the IS design process in different ways, which, in turn, can affect users. We present recommendations to guide developers on how to mitigate unconscious bias in the design process. This research offers a novel view of unconscious bias in teams designing and developing IS.

References

Ahn, Jungyong, Jungwon Kim, and Yongjun Sung. 2022. “The Effect of Gender Stereotypes on Artificial Intelligence Recommendations.” *Journal of Business Research* 141 (March):50–59. https://doi.org/10.1016/j.jbusres.2021.12.007.

Alegria, Sharla N. 2020. “What Do We Mean by Broadening Participation? Race, Inequality, and Diversity in Tech Work.” *Sociology Compass* 14 (6). https://doi.org/10.1111/soc4.12793.

Alkassim, Rukayya S., Xuankiem Tran, Jason D. Rivera, Ilker Etikan, Sulaiman Abubakar Musa, and Rukayya Sunusi Alkassim. 2016. “Comparison of Convenience Sampling and Purposive Sampling.” *American Journal of Theoretical and Applied Statistics* 5 (1): 1–4. https://doi.org/10.11648/j.ajtas.20160501.11.

Anderson, Lisa, Dana Edberg, Adama Reed, Mark G. Simkin, and Debra Stiver. 2017. “How Can Universities Best Encourage Women to Major in Information Systems?” *Communications of the Association for Information Systems* 41:734–58. https://doi.org/10.17705/1CAIS.04129.

Ashburn-Nardo, Leslie, Alex Lindsey, Kathryn A. Morris, and Stephanie A. Goodwin. 2020. “Who Is Responsible for Confronting Prejudice? The Role of Perceived and Conferred Authority.” *Journal of Business and Psychology* 35 (6): 799–811. https://doi.org/10.1007/s10869-019-09651-w.

Avenier, Marie-José, and Catherine Thomas. 2015. “Finding One’s Way around Various Methodological Guidelines for Doing Rigorous Case Studies: A Comparison of Four Epistemological Frameworks.” *Systèmes d’information &amp; Management* Volume 20 (1): 61–98. https://doi.org/10.3917/sim.151.0061.

Azungah, Theophilus. 2018. “Qualitative Research: Deductive and Inductive Approaches to Data Analysis.” *Qualitative Research Journal* 18 (4): 383–400. https://doi.org/10.1108/QRJ-D-18-00035.

Beer, Andrew, and David Watson. 2008. “Personality Judgment at Zero Acquaintance: Agreement, Assumed Similarity, and Implicit Simplicity.” *Journal of Personality Assessment* 90 (3): 250–60. https://doi.org/10.1080/00223890701884970.

Blum, Lawrence. 2004. “Stereotypes And Stereotyping: A Moral Analysis.” *Philosophical Papers* 33 (3): 251–89. https://doi.org/10.1080/05568640409485143.

Bogen, Miranda, and Aaron Rieke. 2018. “Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias.” Upturn.

Bourne, Jenny. 2019. “Unravelling the Concept of Unconscious Bias.” *Race & Class* 60 (4): 70–75. https://doi.org/10.1177/0306396819828608.

Brewer, Marilynn B. 1991. “The Social Self: On Being the Same and Different at the Same Time.” *Personality and Social Psychology Bulletin* 17 (5): 475–82. https://doi.org/10.1177/0146167291175001.

Carpenter, Julia. 2015. "Google's Algorithm Shows Prestigious Job Ads to Men, but Not to Women. Here's Why That Should Worry You." *The Washington Post*, June 7, 2015, Washington edition. https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/.

Carter, Stacy M., and Miles Little. 2007. "Justifying Knowledge, Justifying Method, Taking Action: Epistemologies, Methodologies, and Methods in Qualitative Research." *Qualitative Health Research* 17 (10). https://doi.org/10.1177/1049732307306927.

Chang, Yen-ning, Youn-kyung Lim, and Erik Stolterman. 2008. "Personas: From Theory to Practices." In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, 439–42. Lund Sweden: ACM. https://doi.org/10.1145/1463160.1463214.

Cronbach, Lee J. 1955. "Processes Affecting Scores on 'Understanding of Others' and 'Assumed Similarity.'" *Psychological Bulletin* 52 (3): 177–93. https://doi.org/10.1037/h0044919.

Ehrlinger, Joyce, Thomas Gilovich, and Lee Ross. 2005. "Peering Into the Bias Blind Spot: People's Assessments of Bias in Themselves and Others." *Personality and Social Psychology Bulletin* 31 (5): 680–92. https://doi.org/10.1177/0146167204271570.

Franke, Nikolaus, Marc Gruber, Dietmar Harhoff, and Joachim Henkel. 2006. "What You Are Is What You like—Similarity Biases in Venture Capitalists' Evaluations of Start-up Teams." *Journal of Business Venturing* 21 (6): 802–26. https://doi.org/10.1016/j.jbusvent.2005.07.001.

Galletta, Anne. 2013. *Mastering the Semi-Structured Interview and beyond: From Research Design to Analysis and Publication.* Mastering the Semi-Structured Interview and beyond: From Research Design to Analysis and Publication. New York, NY, US: New York University Press. https://doi.org/10.18574/nyu/9780814732939.001.0001.

Gültekin, Duygu Güner. 2024. "Understanding and Mitigating Authority Bias in Business and Beyond:" In *Advances in Human Resources Management and Organizational Development*, edited by Enis Siniksaran, 57–72. IGI Global. https://doi.org/10.4018/979-8-3693-1766-2.ch004.

Hewstone, Miles, Mark Rubin, and Hazel Willis. 2001. "INTERGROUP BIAS." https://doi.org/10.1146/annurev.psych.53.100901.135109.

Kang, Hye-Ryun, Hee-Dong Yang, and Chris Rowley. 2006. "Factors in Team Effectiveness: Cognitive and Demographic Similarities of Software Development Team Members." *Human Relations* 59 (12): 1681–1710. https://doi.org/10.1177/0018726706072891.

Kaplan, Bonnie, and Joseph A. Maxwell. 1994. "Qualitative Research Methods for Evaluating Computer Information Systems." In *Evaluating the Organizational Impact of Healthcare Information Systems*, edited by James G. Anderson and Carolyn E. Aydin, 30–55. New York, NY: Springer New York. https://doi.org/10.1007/0-387-30329-4_2.

Kenny, Etlyn J, and Rory Donnelly. 2020. "Navigating the Gender Structure in Information Technology: How Does This Affect the Experiences and Behaviours of Women?" *Human Relations* 73 (3): 326–50. https://doi.org/10.1177/0018726719828449.

Kordzadeh, Nima, and Maryam Ghasemaghaei. 2022. “Algorithmic Bias: Review, Synthesis, and Future Research Directions.” *European Journal of Information Systems* 31 (3): 388–409. https://doi.org/10.1080/0960085X.2021.1927212.

Kraemer, Felicitas, • Kees Van Overveld, and Martin Peterson. 2010. “Is There an Ethics of Algorithms?” https://doi.org/10.1007/s10676-010-9233-7.

Lord, Charles G., and Cheryl A. Taylor. 2009. “Biased Assimilation: Effects of Assumptions and Expectations on the Interpretation of New Evidence.” *Social and Personality Psychology Compass* 3 (5): 827–41. https://doi.org/10.1111/j.1751-9004.2009.00203.x.

Mariano, João, Sibila Marques, Miguel R. Ramos, Filomena Gerardo, Cátia Lage Da Cunha, Andrey Girenko, Jan Alexandersson, et al. 2022. “Too Old for Technology? Stereotype Threat and Technology Use by Older Adults.” *Behaviour & Information Technology* 41 (7): 1503–14. https://doi.org/10.1080/0144929X.2021.1882577.

Marshman, Emily M., Z. Yasemin Kalender, Timothy Nokes-Malach, Christian Schunn, and Chandralekha Singh. 2018. “Female Students with A’s Have Similar Physics Self-Efficacy as Male Students with C’s in Introductory Courses: A Cause for Alarm?” *Phys. Rev. Phys. Educ. Res.* 14 (2): 020123. https://doi.org/10.1103/PhysRevPhysEducRes.14.020123.

Mayer, Claude-Hélène. 2021. “Bias, Prejudice and Shame in Predictive Policing: State-of-the-Art and Potential Interventions for Professionals.” In *Shame 4.0: Investigating an Emotion in Digital Worlds and the Fourth Industrial Revolution*, edited by Claude-Hélène Mayer, Elisabeth Vanderheiden, and Paul T. P. Wong, 109–28. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-59527-2_6.

McCarthy, Stephen, Titiana Ertiö, Ciara Fitzgerald, and Nina Kahma. 2024. “Digital Sustainability for Energy-Efficient Behaviours: A User Representation and Touchpoint Model.” *Information Systems Frontiers*, July. https://doi.org/10.1007/s10796-024-10509-7.

McCarthy, Stephen, P O’Raghallaigh, C Kelleher, and F Adam. 2025. “A Socio-Cognitive Perspective of Knowledge Integration in Digital Innovation Networks,.” *Journal of Strategic Information Systems*.

Mitzner, Tracy L., Julie B. Boron, Cara Bailey Fausset, Anne E. Adams, Neil Charness, Sara J. Czaja, Katinka Dijkstra, Arthur D. Fisk, Wendy A. Rogers, and Joseph Sharit. 2010. “Older Adults Talk Technology: Technology Usage and Attitudes.” *Computers in Human Behavior* 26 (6): 1710–21. https://doi.org/10.1016/j.chb.2010.06.020.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* Algorithms of Oppression: How Search Engines Reinforce Racism. New York, NY, US: New York University Press.

Pandey, Akshat, and Aylin Caliskan. 2021. “Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy’s Price Discrimination Algorithms.” In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 822–33. Virtual Event USA: ACM. https://doi.org/10.1145/3461702.3462561.

Pethig, Florian, and Julia Kroenung. 2023. “Biased Humans, (Un)Biased Algorithms?” *Journal of Business Ethics* 183 (3): 637–52. https://doi.org/10.1007/s10551-022-05071-8.

Pritlove, Cheryl, Clara Juando-Prats, Kari Ala-leppilampi, and Janet A Parsons. 2019. "The Good, the Bad, and the Ugly of Implicit Bias." *The Lancet* 393 (10171): 502–4. https://doi.org/10.1016/S0140-6736(18)32267-0.

Pronin, Emily, Daniel Y. Lin, and Lee Ross. 2002. "The Bias Blind Spot: Perceptions of Bias in Self versus Others." *Personality and Social Psychology Bulletin* 28 (3): 369–81. https://doi.org/10.1177/0146167202286008.

Raghavan, Manish. 2023. *The Societal Impacts of Algorithmic Decision-Making*. 1st ed. New York, NY, USA: ACM. https://doi.org/10.1145/3603195.

Raviv, Amiram, Daniel Bar-Tal, Alona Raviv, and Reuven Abin. 1993. "Measuring Epistemic Authority: Studies of Politicians and Professors." *European Journal of Personality* 7 (2): 119–38. https://doi.org/10.1002/per.2410070204.

Rheingans, Penny, Erica D'Eramo, Crystal Diaz-Espinoza, and Danyelle Ireland. 2018. "A Model for Increasing Gender Diversity in Technology." In *SIGCSE 2018 - Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, 2018-January:459–64. Association for Computing Machinery, Inc. https://doi.org/10.1145/3159450.3159533.

Rosenfeld, PAUL, Robert Giacalone, and Catherine Riordan. 1994. "Impression Management Theory and Diversity: Lessons for Organizational Behavior." *American Behavioral Scientist - AMER BEHAV SCI* 37 (March):601–4. https://doi.org/10.1177/0002764294037005002.

Rosson, Mary Beth, and John M. Carroll. 2007. "Scenario-Based Design." In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, edited by Andrew Sears and Julie A Jacko, 1032–50. Boca Raton: CRC Press.

Scopelliti, Irene, Carey K. Morewedge, Erin McCormick, H. Lauren Min, Sophie Lebrecht, and Karim S. Kassam. 2015. "Bias Blind Spot: Structure, Measurement, and Consequences." *Management Science* 61 (10): 2468–86. https://doi.org/10.1287/mnsc.2014.2096.

Scott, Kimberly A., Kimberly M. Sheridan, and Kevin Clark. 2014. "Culturally Responsive Computing: A Theory Revisited." *Learning, Media and Technology* 40 (4): 412–36. https://doi.org/10.1080/17439884.2014.924966.

Sidanius, Jim, Felicia Pratto, Colette van Laar, and Shana Levin. 2004. "Social Dominance Theory: Its Agenda and Method." *Political Psychology* 25 (6): 845–80. https://doi.org/10.1111/j.1467-9221.2004.00401.x.

Silvester, Carly. 2021. "Authority Bias." In *Decision Making in Emergency Medicine*, edited by Manda Raz and Pourya Pouryahya, 41–46. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-16-0143-9_7.

Simons, Alexander, Lena Franziska Kaiser, and Jan Vom Brocke. 2019. "Enterprise Crowdfunding: Foundations, Applications, and Research Findings." *Business & Information Systems Engineering* 61 (1): 113–21. https://doi.org/10.1007/s12599-018-0568-7.

Singer, Peter. 2021. "Ethics." In . https://www.britannica.com/topic/ethics-philosophy.

Smythe, Liz, and Lynne S. Giddings. 2007. “From Experience to Definition: Addressing the Question ‘What Is Qualitative Research?’” *Nursing Praxis in New Zealand Inc* 23 (1): 37–57.

Sperber, Sonja, and Christian Linder. 2023. “Gender Bias in IT Entrepreneurship: The Self-Referential Role of Male Overrepresentation in Digital Businesses.” *European Journal of Information Systems* 32 (5): 902–19. https://doi.org/10.1080/0960085X.2022.2075801.

Tansey, Michael M. 1998. “How Delegating Authority Biases Social Choices.” *Contemporary Economic Policy* 16 (4): 511–18. https://doi.org/10.1111/j.1465-7287.1998.

Terry, Gareth, Nikki Hayfield, Victoria Clarke, and Virginia Braun. 2017. “Thematic Analysis.” In *The SAGE Handbook of Qualitative Research in Psychology*, edited by Carla Willig and Wendy Stainton Rogers, Second edition. Thousand Oaks, California: SAGE Publications Inc.

Tsamados, Andreas, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, · Luciano Floridi, and Luciano Floridi. 2020. “The Ethics of Algorithms: Key Problems and Solutions” 1:3. https://doi.org/10.1007/s00146-021-01154-8.

Vainionpää, Fanny, Marianne Kinnula, Netta Iivari, and Tonja Molin-Juustila. 2021. “Girls in IT: Intentionally Self-Excluded or Products of High School as a Site of Exclusion?” *Internet Research* 31 (3): 846–70. https://doi.org/10.1108/INTR-09-2019-0395.

W. DuBow and Gonzalez. 2020. “NCWIT Scorecard: The Status of Women in Technology.” National Center for Women & Information Technology (NCWIT). https://wpassets.ncwit.org/wp-content/uploads/2021/05/20221741/ncwit_scorecard_data_highlights_10082020.pdf.

Yansen, Guillermina, and Mariano Zukerfeld. 2014. “Why Don’t Women Program? Exploring Links between Gender, Technology and Software.” *Science, Technology and Society* 19 (3): 305–29. https://doi.org/10.1177/0971721814548111.

Zabel, Sarah, and Siegmar Otto. 2021. “Bias in, Bias Out – the Similarity-Attraction Effect Between Chatbot Designers and Users.” In *Human-Computer Interaction. Design and User Experience Case Studies*, edited by Masaaki Kurosu, 184–97. Cham: Springer International Publishing.

Zreik, Jean-Pierre. 2019. “GEO-LOCATION, LOCATION, LOCATION.” *Rutgers Computer and Technology Law Journal* 45 (2): 135–68.

Actionable Data for Climate Health: A Literature Review

Yamikani Phiri[1], Silvia Masiero[1], Anders Nielsen[2]
[1] University of Oslo
[2]Norwegian Institute of Bioeconomy Research (NIBIO)

Completed Research

Abstract

The notion of climate health refers to the multiplicity of health effects and risks posed by different dimensions of climate change. Climate health data are data that capture multiple dimensions of climate health, and that can be used, aggregated and mapped in order to tackle global health challenges. In this paper we use the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework to conduct a literature review on climate health data, focusing on their actionability in relation to global health challenges. Our literature review identifies themes of (a) early warning and emergency response, (b) assessment of planning interventions, and (c) preventive and future planning as central to research on climate health data. With this literature review, we frame climate health data as an object of interest for the IS field, exploring ways in which IS research can uniquely contribute to knowledge on the global challenge of climate health.

Keywords: climate health; climate data; global challenges; literature review

1. Introduction

Over the last 10-15 years, two concomitant phenomena have affected how climate change, and its multiple impacts on life on the planet, have been brought to the attention of Information Systems (IS) research. A first phenomenon lies in the turn of the IS field to societal challenges, framed as complex social problems combined with the ecosystems that perpetuate them (Majchrzak et al. 2016). This has been followed by an explicit turn of the field to social justice, defined as ‘a state of fairness, moderation, and equality in the distribution of rights and resources in society’ (United Nations 2006, cited in Aanestad et al. 2021). In the landscape of an IS discipline that takes social justice as a theme of concern, the structural vulnerabilities exposed by climate change, and the ways IS can address them, become prominent themes of IS research.

A second phenomenon, unfolding at the same time, lies in the positioning of data as a central, independent object of IS research (Aaltonen & Stelmaszak 2023). This development has marked a novel trajectory in the field: originally scattered across multiple bodies of literature, data have emerged as a self-standing research object, generating a new bibliographic stream in IS (cf. Aaltonen et al. 2021; Alaimo & Kallinikos 2022; Järvenpää & Essén 2023). In the light of this, data have become a central device to understand the societal challenges that, with a focus on issues of justice, IS research has embraced (Aanestad et al. 2021). This puts us in the position of viewing data in relation to such global societal challenges: first and foremost, that of climate health, an expression that indicates the multiplicity of health risks posed by different dimensions of climate change (WHO 2023a).

Against this backdrop, it becomes important to learn how climate health data are used in the production of health-related outcomes, and how such use can be put to the service of tackling the severe challenges posed by climate-related phenomena. In this paper, we use the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework to conduct a literature review on climate health data, interrogating their actionability towards challenges of global health including early warning and emergency response, assessment of planning interventions, preventive and future planning. In doing so, we accomplish two objectives: we map extant knowledge on climate health data, and we position climate health data as an object of interest for IS research. We conclude by drawing the implications of our study for research on climate health as a global challenge for the IS field.

2. Rationale: IS and Climate Health Data

Climate change poses a significant threat to mental and physical human health. Exposure to heat waves, wildfires, and changes in precipitation leading to drought and floods are direct causes of vulnerabilities, injuries, illness, and even death (Di Napoli et al. 2022; Masselot et al. 2023; WHO 2023a). Climate and climate change can also have an indirect effect on health outcomes through natural systems; for example, Malaria incidence prevalence is more related to rainfall and temperature (Mafwele & Lee 2022) and human systems, for instance, agriculture is sensitive to climate, and climate change affects agriculture systems as changes in precipitation, temperature, season, pest and disease dynamics affect the quantity and quality of the food which has an indirect effect on human nutrition (Owino et al. 2022; von Braun 2020).

Climate health focuses on the relationship between climate and climate change and its implications on human health outcomes (Casson et al. 2023; Cunsolo Willox et al. 2012; Shea et al. 2020). The changing climatic conditions have enhanced the transmission of many climatically sensitive infectious vectors-food-and water-borne diseases, and non-communicable diseases, particularly in developing countries (Hunter

2003; WMO 2023). Dengue fever is now the world's fastest-growing vector-borne disease and has epidemic potential due to recent sustained warmer temperatures caused by climate change (Braun et al. 2023). Weather- and climate- hazards have contributed to mortality and the global burden of disease (Di Napoli et al. 2022). Approximately 250,000 additional people die every year due to climate-related cases like undernutrition, malaria, diarrhoea and heat stress alone (WHO 2023). Climate health data is essential for effectively preparing, intervening, and managing climate-sensitive diseases and reducing adverse health outcomes (Ghebreyesus et al. 2009; Nissan et al. 2022). Climate health data brings together streams of work to help shed light on how climate can impact society's health outcomes.

This study frames climate health data as an object of IS research. To do so, it conceives of climate and weather based on the typology of Bruno Soares et al. (2018). In this typology, weather is characterised as observations on a short time scale, where forecasts make predictions ranging from minutes, hours, days, and weeks to the present. Climate describes weather patterns over longer time scales, from months to decades or even longer. Despite the distinction between climate and weather, they are interconnected in a continuum, necessitating that decision-makers consider both elements (Georgeson et al. 2017). Technological advancements in observations, data management, and modelling have improved the global forecasting and availability of weather and climate data (Bruno Soares et al. 2018; Georgeson et al. 2017). Although vast quantities of data are currently available from different sources, this does not necessarily mean better data for decision-making. Most research in the past 40 years has focused on underlying scientific prediction or observation systems, the uptake and the extent to which climate information is used and how it influences decision-making in climate-sensitive sectors like health and agriculture is not well-documented, particularly in developing countries (Bruno Soares et al. 2018; Lemos et al. 2012).

In the context of climate and health, weather and climate data are crucial in health surveillance, outbreak investigation, health risk assessment, health service delivery, policy long-term planning and programmatic decision-making (Thomson et al. 2018; WHO/WMO 2023). Effective responses to climate change require actionable climate data integrated into decision-making processes (Jagannathan et al. 2023). Data is actionable if it is correct, consistent, accessible, understandable and timely to be applied to solve a real-world problem in a given context (CARE Climate Change 2014; Evans et al. 2017; Jagannathan et al. 2023; Sarkar 2022). Actionable climate health data is essential for effective preparation, intervention and management of climate-sensitive diseases and the reduction of negative health outcomes (Ghebreyesus et al. 2009; Nissan et al. 2022). Integrating climate data in health has contributed to uncovering hidden challenges that could have negatively affected public health gains of the previous decades; for example, the association of increased temperature and adverse maternal and perinatal outcomes like Preeclampsia, preterm birth, low birth weight and stillbirth which can affect the mother and the children's brain and body developments (WHO 2023b). Climate data can also be used for climate-related forecasts of floods and heat waves, as well as early warning systems for vector-borne diseases (WHO/WMO 2023). It is crucial to invest in actionable data to adapt to and mitigate the effects of climate change (UNDRR 2015).

Our literature review aims to understand how climate health data is used in health-related outcomes. The challenge of climate change and the importance of climate health data and their novel insights that are essential in mitigating and adapting to climate change cannot be understated. While climate variabilities are inevitable, climate health data can be used to mitigate their consequences, such as loss of life and socioeconomic changes, through advance preparation and contingency plans (Buizer et al. 2000). The heterogeneity in climate data and uptake of climate data is relevant to understanding how climate data is applied in health-related actions and outcomes. Climate data and services can enhance our understanding of how and when the health system and population can be impacted and how this can be managed (WMO 2023). The use of climate data is increasingly essential in early warning, managing and responding to health burdens like dengue fever (Lowe et al. 2017). Climate data is essential for the improvement of climate-sensitive decisions that contribute to mitigation and adaptation goals (Findlater et al. 2021). Despite the profound impact of climate change on public health and healthcare systems, the health sector is significantly under-represented in policies, planning, and programming climate change adaptation initiatives worldwide (Cunsolo & Harper 2019). In addition, three-quarters of National Meteorological and Hydrological Services provide climate data, only a quarter of the Ministry of Health in the world incorporated climate and weather data in their climate-sensitive sectors (WMO 2023), this has contributed to neglected, underreported and underestimation of climate-related maternal and child health events (WHO 2023b). The need to bring together the understanding of how climate data is used in health-related outcomes is essential for adaptation to health efforts.

3. Methodology

We conducted the literature review using the PRISMA framework. PRISMA provides a comprehensive and transparent way of reporting on a literature review (Sarkis-Onofre et al. 2021). The articles were retrieved from the Scopus database on 18 January 2024. The search terms used were based on the concepts of climate data, health, and decision in their titles, abstracts or keywords. We included only journal articles from the identified articles that presented empirical studies about health-related decisions based on weather or climate data. We run the following query

(TITLE-ABS-KEY ({climate information} OR {climate Knowledge} OR {weather information} OR {weather Knowledge} OR {weather data} OR {climate data})
AND TITLE-ABS-KEY ({health} OR {Well-being} OR {Wellbeing} OR {Wellness})
AND TITLE-ABS-KEY ("decision making"))

The search on Scopus returned 104 hits; after filtering, 68 articles remained. Of the remaining articles, the distribution of the journal were quite dispersed, the journals with most hits were *Climate Change* (4), *International Journal of Environmental Research And Public Health* (3), and *Climate Services* (3). The filtering process removed 1 article that was not in English, 7 book series, 10 books, and 18 conference proceedings, which we have not included to keep the review to a dataset of journals. The remaining articles' abstracts were read, and 32 were excluded because they focused on nonhuman-related health outcomes. Thirty-six articles were fully reviewed. Out of the 36, eleven articles were excluded because they did not have any empirical relationship between health, climate and weather data or information and decision-making. Articles that were unclear to the individual reviewer were shared with another or all the reviewers and discussed before they were added or excluded. Articles were selected if they reported on one or more health-related actions that were made in part using climate and weather data, resulting in 25 articles.

Data about decisions were extracted together with the climate/weather data used. The extracted data included the type of weather data or climate event, country of study, the study's approach, information products produced, target consumers or users, and actions that were made or could be made using the data. These were then analyzed using thematic analysis, a flexible way to identify, analyze, report and interpret data within the literature (Braun & Clarke 2006). In the first round, we found common words such as "early warning," "prevent," "planning," and "resource allocation." In the second round, these were grouped into themes based on patterns related to the actions taken.

4. Results and Discussion

4.1 Climate and Weather: Data and Events

Our study established that most papers included more than one climate and weather variable. Though not mutually exclusive, temperature and precipitation are the most common factors appearing in 20 and 14 research articles, respectively, as shown in Figure 1.

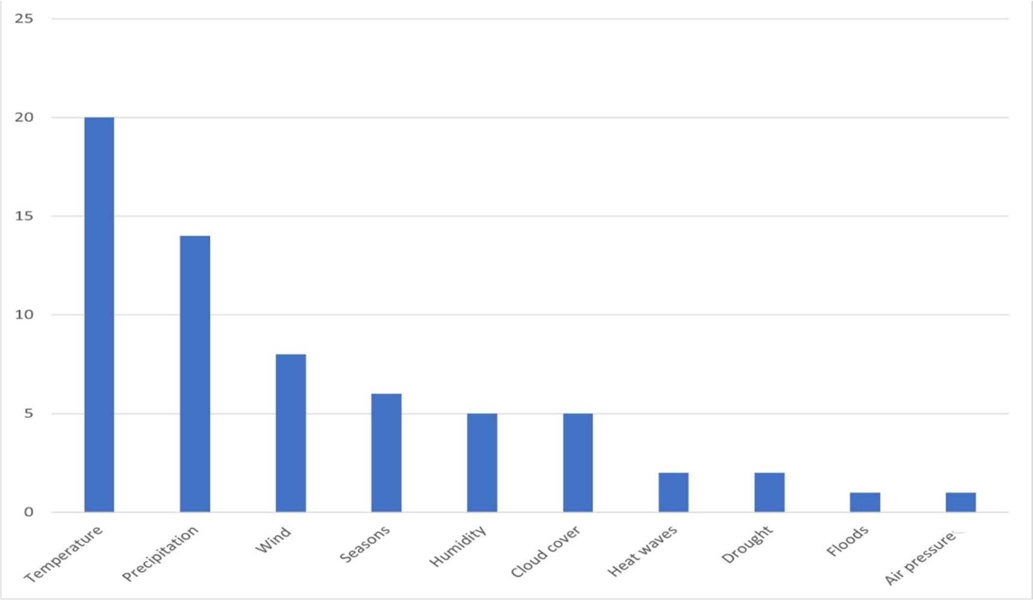


Figure 1. Climate and weather variables

Heat waves, droughts, and floods are extreme weather events that are also studied in the papers included for the study, but their appearance is minimal.

4.2 Climate and Weather Data and Health-Related Decision Making

We grouped the decision-making based on weather and climate data into three themes based on how the data was used. These themes were decisions related to early warning and emergency response, assessment and planning intervention, and preventive and future planning. In this case, this implies how climate and health data inform or could inform actions that impact a person's or community's health outcomes.

4.2.1 Early warning and emergency response

This theme was associated with using climate and weather information to take action about upcoming health implications like epidemics and quickly respond to an existing health situation. Climate and weather data characteristics are commonly associated with short timescales, from hours to a season. Temperature and precipitation were the most common data used for early warning and emergence response, with wind and humidity separating the rest. Seasons and seasonal forecasts were the longer timeframes associated with these themes, and drought was the only extreme climate event, as shown in Table 1.

Author(s)	Weather/ climatic factors	Applications	Region
Li et al. 2023	Temperature, Air pressure, Wind speed, Precipitation	Immediate warning regarding regional heat wave events in terms of where the heatwave will originate and where it will propagate its pathways and response.	North America (USA)
Usmani et al. 2023	Temperature, Rainfall	Quick response to cholera outbreaks by risk prediction that aims to anticipate when an early response in order to change the shape of an epidemiological curve.	Asia (Yemen)
Stewart-Ibarra et al. 2022	Rainfall, Temperature, Seasons	Early warning for dengue based on meteorological predictions due to wet conditions.	North America (Barbados)
Quinn et al. 2022	Temperature	An early detection platform for malaria is essential in determining system-related supply chain and personnel management. Providing user-tailored response actions appropriate to the threat and the location	Africa (Mozambique, Ethiopia, Sub-Sahara Africa)
Kim et al. 2020	Temperature, Humidity	Identifying asthma risk/severity by predictive Peak expiratory flow rate (PEFR) based on particulate matter (PM) previous PEFR, temperature, and humidity.	Asia (South Korea)

Skelton 2020	Seasonal temperature	Caregivers of the elderly and infants should be aware of the danger of heat waves; prolonged asthma suffering due to the longer pollen season.	Europe (Switzerland)
Bruno Soares et al. 2018	Seasonal forecast	Respond to emergencies in health like outbreaks (daily time scale).	Europe
Worrall et al. 2008	Seasonal forecast	Provide advance warnings of the potential for malaria epidemics to occur.	Africa (Sub-Sahara Africa)
Kenemaru et al. 2017	Temperature, Wind, Cloud	Provide prediction to dispatch a safe flight for an emergency. This leads to decreased morbidity and mortality of patients by providing transport capability for more patients in the mountains who rely on it.	Asia (Japan)
Boulanger et al. 2010	Precipitation, Temperature	Early warning for dengue epidemic based on meteorological predictions.	South America

Table 1. Early warning and emergency response

Temperature plays a role in providing early warning for health practitioners on vector-borne diseases, like malaria (for example Quinn et al. 2022; Worrall et al. 2008) and dengue fever (for example Boulanger et al. 2010; Lowe et al. 2018; Stewart-Ibarra et al. 2022). Climate variables affect vectorial capacity, the ability of vector species to survive and reproduce in an environment and the ease with which they acquire, carry, and transmit pathogens (Braun et al. 2023). The effects of weather conditions (temperature and precipitation) do play a role in the incubation and survival of mosquitos, their transmission capacity and geographic range of e.g. malaria and dengue fever (Thomson et al. 2018). Weather and climate data, like rainfall and temperature, are used to predict when the probability of an outbreak increases by understanding the causal relationships between weather conditions and vector- or water-borne disease dynamics. We found that temperature was commonly used as an early warning signal in the southern hemisphere, such as sub-Saharan Africa (Quinn et al. 2022), South America (Stewart-Ibarra et al. 2022) and Barbados (Stewart-Ibarra et al. 2022).

In the northern hemisphere, temperature data was also used to inform action on non-communicable health-related outcomes like heat waves and respiratory diseases such as asthma. Temperature is essential for understanding heat waves' origin, time, and path in Europe and North America (for example, Bruno Soares et al. 2018; P. Li et al. 2023; Skelton 2020). Weather forecasts are crucial for raising awareness of vulnerable people so they can prepare for heat waves (P. Li et al. 2023; Skelton 2020). Temperature and humidity data are also used as early warnings to predict the severity of weather-induced asthma (Kim et al. 2020) and increased asthma suffering due to a prolonged season of higher temperature that contributes to lengthening the pollen season (Skelton 2020).

Weather data is also used in responding to emergencies; factors like wind and cloud cover provide key information in predicting the dispatch of safe emergency flights in Japan's mountainous regions (Kanemaru et al. 2017). This has provided those living in the mountains with an emergency transport system to rely on. Temperature and rainfall data have also been used to respond to cholera cases quickly and anticipate cases with the aim of changing the shape of the epidemiological curve in Yemen (Usmani et al. 2023). In all these cases, the application of weather data has led to better health outcomes.

4.2.2. Assessment and Planning Interventions

The studies investigating assessment and planning looked at actions that affect current or probable upcoming health scenarios. The characteristics of climate and weather data used in studies have a short timescale, for example, temperature, precipitation, wind and humidity, and medium to long timescale, such as seasons (see Table 3). Seasonal data appeared more frequently in this category than early warning and emergence responses. Climate and weather data have been used in the planning and intervention of water-borne diseases (for example Bornemann et al. 2019), vector-borne diseases (for example Boulanger et al. 2010; Ceccato et al. 2014; Lowe et al. 2018; Quinn et al. 2022), airborne diseases (for example Agapito et al. 2020) and other health outcomes like malnutrition (for example Bornemann et al. 2019), paediatric disease (for example H. Li et al. 2019) and physical fitness (Flynn et al. 2012) as shown in Table 2.

Source	Weather/ climatic factors	Applications	Region
Koch et al. 2023	Temperature, Precipitation	Adaptive measures of coping in rural areas that are highly exposed to climatic changes protect people from heat, especially during the night and outdoor work; heat associated with numerous effects on the cardiovascular system and heat-related sleep disruption, which is an issue for the general public's health.	Africa (Burkina Faso)
Quinn et al. 2022	Temperature, Precipitation	12 weeks of malaria caseload prediction based on climate and epidemiological data; this allows officials to prepare for the incoming cases.	Africa (Mozambique, Ethiopia, Sub-Sahara Africa)
Stewart-Ibarra et al. 2022	Rainfall, Temperature Seasons, drought	Providing timely and accurate information to guide intervention for the public health sector by early detection of potential and endemic dengue outbreaks; identifying the risk of dengue fever at different times and climatic events like drought and precipitation.	North America (Barbados)
Kanti et al. 2022	Temperature, Precipitation, Humidity, Wind	Provide a good definition of what is a heat wave to find a proper threshold for heat wave warning that can reduce mortality as mortality attributable to heat wave.	Europe (France)

Ageno et al. 2020	Temperature, Precipitation, Humidity, Wind	Importance of weather condition and their potential impact on transmission of COVID-19 and taking appropriate measures to mitigate the risk of infection.	Europe (Italy)
Bornemann et al. 2019	Seasonal forecast	Selection of most cost-effective sanitation systems (changes in total water demand).	Europe
Li et al. 2019	Temperature, Precipitation, Humidity, Wind	Planning when to allocate resources in the paediatric department of health personnel and clinics; Planning when to allocate resources in the paediatric department of health personnel and clinics.	Asia (China)
Bornemann et al. 2019	Seasonal forecast	Food security (malnutrition) planning, ensuring safe activities like harvesting (agriculture), and Developing groundwater irrigation techniques in available shallow water storage for irrigation.	Europe
Lowe et al. 2018	Rainfall, Temperature, Drought, Seasonal	Generate forecasts of febrile illness using the seasonal disease forecast process. Increase preparation time to better allocate scarce resources and reduce the risk of climate outbreaks.	North America (Barbados)
Lowe et al. 2018	Rainfall, Temperature, Drought	Understanding and assessment of the lag between rainfall and drought and onset of febrile outbreaks	North America (Barbados)
Cecchi et al. 2014	Seasons, Temperature, Precipitation	Assessing risk and allowing Ministries of Health and WHO to have an informed decision predicting and	Africa

		preventing meningitis and malaria epidemics.	
Flynn et al. 2012	Precipitation, Wind, Temperature, Seasonal	Support the development of methods to mitigate the adverse effect of weather on bicycle commuting and promote increased use of active transport for routine purposes.	North America (USA)
Boulanger et al. 2010	Precipitation, Season	Seasonal forecasting of climate-related dengue epidemic.	South America

Table 2. Assessing and planning interventions

Climate and weather data are key in predicting the anticipated number of cases, risks, developing cost-effective strategies and allocating resources to adapt or mitigate climate and weather-related health outcomes (Boulanger et al. 2010; Kanti et al. 2022; Lowe et al. 2018; Quinn et al. 2022; Stewart-Ibarra et al. 2022). Climate data from satellite images and epidemiological data predicted the number of malaria cases to be anticipated (Quinn et al. 2022), what paediatric disease and what age will be impacted more (Li et al. 2019). Climate health data provided insights on disease trends that are essential for stakeholders in the action to be taken. This information plays a significant role in planning and assessing the current disease burden and resources. Climate data is applied in the selection and cost-efficient sanitation system (Bornemann et al. 2019), efforts to mitigate the effects of weather on commuting (Flynn et al. 2012), understanding the time it takes from a climate event to a disease outbreak (Lowe et al. 2018; Stewart-Ibarra et al. 2022), assessing risk and preventing of meningitis and malaria outbreaks (Ceccato et al. 2014), heat coping mechanisms, potential COVID-19 transmission (Agapito et al. 2020; Koch et al. 2023) and providing threshold to understand develop interventions and warning to reduce motility (Kanti et al. 2022).

Weather forecasts and climate change projections are powerful tools for predicting environment conditions on both short and long-term scales. Climate and weather data, in this case, is essential in understanding the time from a climate event and health-related outcomes like outbreaks and pandemics. While some climate events can have an immediate impact that requires an emergency response, climate health data is used to model and understand the time it takes (lag) based on the nature of vectors and pathogens, vectorial capacity and socio-anthropological factors (Boulanger et al. 2010;

Braun et al. 2023). This is modelled and used to predict expected cases and climate health outcome lags. Climate data, for example, seasonal data, is also a good proxy for predicting the amount of yield, as agriculture is sensitive to climate variation and essential to food security, and the amount of water, as precipitation is one of the main methods of replenishing major freshwater sources. This is essential for assessment and planning interventions.

4.2.3 Preventive and Future Planning

Preventive actions and future planning are decisions that are associated with preventing future health impacts by ensuring that appropriate structures are in place, as shown in Table 3. Due to the nature of the theme, weather and climate data used have a significantly longer timescale than the other two themes, usually from months, seasons, years and decades. This involved the use of a long timescale of data collection, ranging from 2 years of temperature data collection in infrastructure planning (Short et al. 2015) to season(s) to decades (10 to 30 years) (Thomson et al. 2018). This category also included more extreme climate events, such as drought, flood, and heat waves, than the other two categories.

This involved defining thresholds to identify heat waves (Vescovi et al. 2009), prioritisation of decisions and investments by governments and donors (Quinn et al. 2022), planning infrastructure such as hospitals for better heat wave adaptation and ventilation (Short et al. 2015), policy planning (Cheng et al. 2012), establishing seasonal disease trends, transmission intensity and shifting geographies of infection diseases (Quinn et al. 2022; Thomson et al. 2018), coping mechanisms (Koch et al. 2023). All these studies target policymakers or institutions that have an influence on policymaking and long-term preventive measures.

Author(s)	Weather/ climatic factors	Applications	Region
Quinn et al. 2022	Temperature seasonality	Predict the shifting geographies and seasonality of malaria transmission suitability because of increase temperature has created new places where the vector can survive.	Africa (Mozambique, Ethiopia Sub-Sahara Africa)

Cheng et al. 2012	Temperature, Wind, Cloud cover	Health policy planning and preventive measures for today's weather and future climate conditions.	North America (Canada)
Kanti et al. 2022	Temperature, Precipitation, Humidity, Wind	Consideration of the choice of heatwave definition in terms of the number of warnings it could trigger and the scale of heatwave related health implications, perceived risk, and attributable mortality.	Europe (France)
Venus et al. 2022	Precipitation, Temperature, Drought, Floods	Improvements on infrastructure like permanent housing, latrines, and health centre networks are essential in dealing with more frequent climate shocks and is key to combat flood-related diseases like diarrhoea and cholera. Promoting public health by recommending flood-proof latrines as no water is needed.	Europe
Quinn et al. 2022	Temperature	Future public health decisions and priorities for donors and governments.	Africa (Mozambique,Ethiopia Sub-Sahara Africa)
Bornemann et al. 2019	Seasonal forecast	Selection of the most cost-effective sanitation systems (changes in total water demand); selection of the most cost-effective sanitation systems (changes in the aetiology of pathogens and critical disease groups).	Europe
Thomson et al. 2018	Seasons, Temperature, Humidity	Determining directly or indirectly transmission dynamics of vector-borne disease with extreme weather	Africa

		events, Decadal (10-30 years) and long-term shifts in the climate may have an impact on vector-borne diseases by changing their geographic range.	
Short et al. 2015	Cloud cover, Temperature	Planning for hospital infrastructure as heat waves are risky for young, elderly, and seriously ill.	Europe (England)
Cheng et al. 2012	Temperature , Wind, Cloud cover	Consideration of infrastructure planning to consider climate extremes over the lifespan of the building.	North America (Canada)
Corral et al. 2012	Precipitation , Temperature	Exploring the impacts of climate on disease trends and testing methodologies for impact evaluation for malaria interventions in Ethiopia.	Africa (Ethiopia)
Vescovi et al. 2009	Temperature	Characterisation of current and future climate hazard in terms of heatwaves e.g., number of hot days.	North America (Canada)

Table 3. Preventive and future planning

4.3 Accessibility of climate health data

Climate health data in these studies targeted health practitioners, policymakers, planners and other stakeholders, as shown in Figure 2. This information was accessed through online platforms, publications, bulletins, campaigns, alerts, news, models, databases, charts and data warehouses.

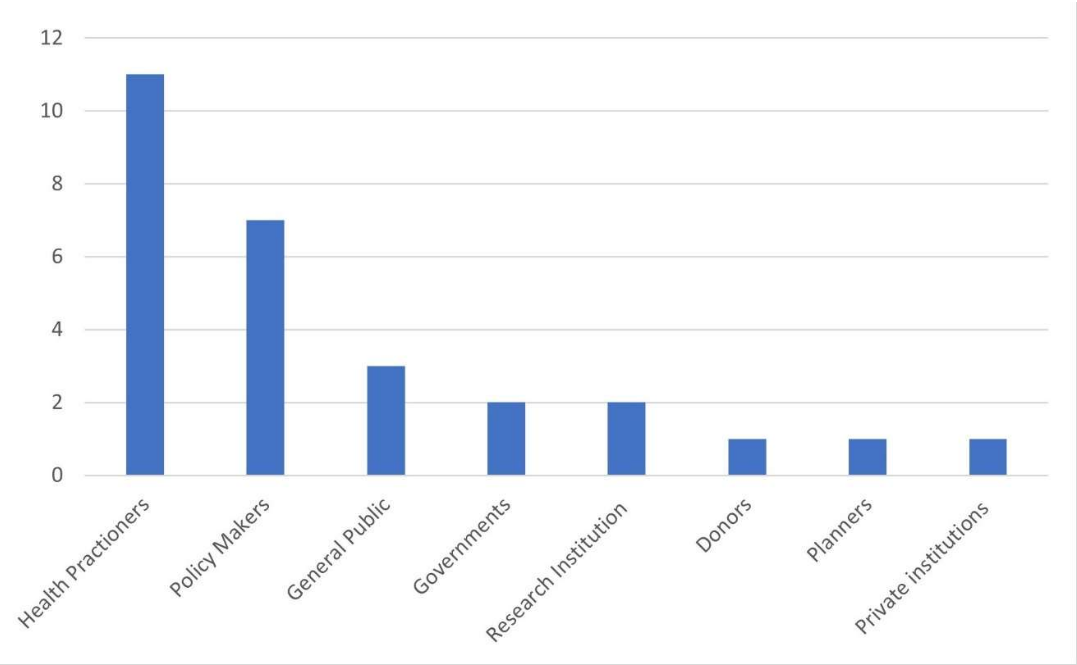


Figure 2. Accessibility of data

Health practitioners and policymakers were the key targets of the data. Climate health data forms part of planning, interventions and future policy direction in the health system (Ceccato et al. 2014; Li et al. 2019; Lowe et al. 2018; Quinn et al. 2022; Short et al. 2015; Stewart-Ibarra et al. 2022; Thomson et al. 2018; Usmani et al. 2023; Venus et al. 2022; Worrall et al. 2008). While targeting policymakers and health practitioners is essential, it is also important to target the public directly with relevant climate health data products. People will have to face the health implications of climate and adapt with or without the help of the government (Roy & Venema 2002). Data concerning early warning, the need for change in socio-anthropological activities, for example, storing water during drought seasons (Stewart-Ibarra et al. 2022), changes in disease seasonal trends and shifts in disease from one geography to another are essential for the general public. If the public is largely excluded as a target of data products, then climate change and its implications will be perceived as an issue of a specific demographic, in this case, health practitioners and policymakers; therefore, the public will be less engaged (Phadke et al. 2015). This will affect people's ability to prepare and change their behaviour to mitigate and adapt to climate change.

4.4 From Data to Decision

Approaches used to transform climate and weather data into information were mostly quantitative, as shown in Figure 3. Most studies used climate and weather data to produce models, simulations, projections and forecasts to inform decision-making on health outcomes. While this is sufficient, there is also a need for an in-depth understanding of people and society's climate health decision-making on health outcomes, adaptation measures, and mitigation impact through qualitative studies. Qualitative studies help in understanding aspects of human values, culture and relationships, which are the essential health outcomes and how and by whom decisions are made (Cypress 2015). Need for tailored information (for example Bornemann et al. 2019), lack of social burden data (Usmani et al. 2023), and social practices like keeping water in uncovered containers at home during drought seasons (Stewart-Ibarra et al. 2022) requires an in-depth understanding of the social aspects and practise of the people in their local context. This is essential in effective action as climate change health impacts are not top-down, can be specific to a particular community, and there is not a one-size-fits-all solution (Phadke et al. 2015).

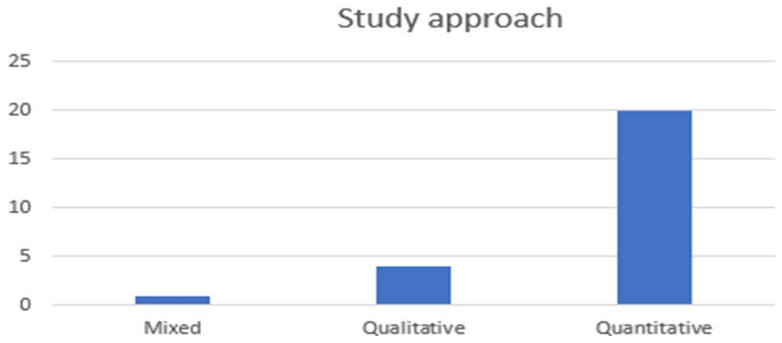


Figure 3. Study approaches

4.5 User-Provider Communication

It was noted that there is a gap between climate scientists and health workers; on the one hand, models and predictions need to be translated to be understandable to the public and health practitioners so that they can understand the risks and take action. On the other hand, some requirements from health practitioners need to be "interpreted" by climate scientists (Stewart-Ibarra et al. 2022). This could also be part of the reason why, for example, in Switzerland, while other professionals like building technicians and green space managers incorporated heat waves in their decision-making, for health specialists, this is met with mixed reactions except for the elderly and children, as others argue that this is a personal choice (Skelton 2020). Some recent studies have highlighted the disconnect between climate information producers and actual end users. Porter & Dessai expressed the "Mini-me" idea of producers who sometimes think that the end-user has the same rationale, creates misalignment as the actual end-user (Porter & Dessai 2017), this has a negative effect on the actionability of the climate data in health and other sectors (Bruno Soares et al., 2018; Jagannathan et al. 2023; Porter & Dessai 2017; Vaughan & Dessai 2014). Producers often overlook the specific requirements of different user groups, making broad, generalised information products. It is essential to package climate information in a format that is not only scientifically credible but also usable for decision-making on adaptation to climate change (Porter & Dessai 2017).

There is a need for communication and collaboration between climate scientists and users to bridge the gap between scientific knowledge and practical application in the context of climate change adaptation. Actionable climate health data can help societies to prepare better, mitigate and adapt to the risks and opportunities posed by climate variability and climate change (Bruno Soares et al. 2018). The collaborative effort between scientists and decision-makers is essential for ensuring that climate information is not only accurate but also comprehensible and directly applicable to a wide array of stakeholders.

5. Conclusion

The IS discipline has recently witnessed a turn to global societal challenges, along with the framing of data as an independent object of research. Against this backdrop, we have used the PRISMA method to review the literature on climate health data, interrogating their actionability towards challenges of global health. We have found that the themes of (a) early warning and emergency response, (b) assessment of planning interventions, and (c) preventive and future planning inform three central bodies of literature on the topic, which enables us to position such themes at the core of the engagement of IS research with climate health.

Our literature review accomplishes two primary goals. First, it has enabled us to map extant knowledge on climate health data, knowledge that is produced predominantly outside the IS field. This finding enables us to imagine routes through which IS research can contribute knowledge on climate health data: first, framing data as an independent research object (Aaltonen & Stelmaszak 2023), IS research affords disentangling the specific contribution of data to climate health. Secondly, with its recent turn to social justice (Aanestad et al. 2021), IS research is well-positioned to conceive climate health data as part of a process of addressing structural vulnerabilities, bridging the climate health discourse with streams of work on data governance and data sustainability, among others (Aaltonen et al. 2021; Järvenpää & Essen 2023).

Second and finally, this paper has enabled us to position climate health data as an object of interest for IS research. On the one hand, the early-stage engagement of the discipline with this new object implies that the literature reviewed here comes mostly from other domains, from which much learning can be generated. On the other, the paper has illuminated substantial routes through which the IS discipline, with its focus on the social study of technology, can contribute research on climate health data. We hope this work can serve as a basis for the devising of such novel, societally important engagements for the IS discipline.

References

Aaltonen, A., Alaimo, C., & Kallinikos, J. 2021. "The making of data commodities: Data analytics as an embedded process," *Journal of Management Information Systems* (38:2), pp. 401-429.

Aaltonen, A., & Stelmaszak, M. 2023. "The Performative Production of Trace Data in Knowledge Work," *Information Systems Research*, online 20 September 2023.

Aanestad, M., Kankanhalli, A., Maruping, L., Pang, M.-S., & Ram, S. 2021. "Digital Technologies and Social Justice". *MIS Quarterly* Special issue call for paper, 1-9.

Agapito, G., Zucco, C., and Cannataro, M. 2020. "Covid-Warehouse: A Data Warehouse of Italian Covid-19, Pollution, and Climate Data," *International Journal of Environmental Research and Public Health* (17:15), p. 5596.

Bornemann, F. J., Rowell, D. P., Evans, B., Lapworth, D. J., Lwiza, K., Macdonald, D. M. J., Marsham, J. H., Tesfaye, K., Ascott, M. J., and Way, C. 2019. "Future Changes and Uncertainty in Decision-Relevant Measures of East African Climate," *Climatic Change* (156:3), pp. 365-384.

Boulanger, J.-P., Brasseur, G., Carril, A. F., De Castro, M., Degallier, N., Ereño, C., Le Treut, H., Marengo, J. A., Menendez, C. G., Nuñez, M. N., Penalba, O. C., Rolla, A. L., Rusticucci, M., and Terra, R. 2010. "A Europe–South America Network for Climate Change Assessment and Impact Studies," *Climatic Change* (98:3-4), pp. 307-329.

Braun, M., Andersen, L. K., Norton, S. A., and Coates, S. J. 2023. "Dengue: Updates for Dermatologists on the World's Fastest-Growing Vector-Borne Disease," *International Journal of Dermatology* (62:9), pp. 1110-1120.

Braun, V., and Clarke, V. 2006. "Using Thematic Analysis in Psychology," *Qualitative Research in Psychology* (3:2), pp. 77-101.

Bruno Soares, M., Alexander, M., and Dessai, S. 2018. "Sectoral Use of Climate Information in Europe: A Synoptic Overview," *Climate Services* (9), pp. 5-20.

Buizer, J. L., Foster, J., and Lund, D. 2000. "Global Impacts and Regional Actions: Preparing for the 1997–98 El Niño," *Bulletin of the American Meteorological Society* (81:9), pp. 2121-2140.

CARE. 2014. "Facing Uncertainty: The Value of Climate Information for Adaptation, Risk Reduction and Resilience in Africa."

Casson, N., Cameron, L., Mauro, I., Friesen-Hughes, K., and Rocque, R. 2023. "Perceptions of the Health Impacts of Climate Change among Canadians," *BMC Public Health* (23:1), p. 212.

Ceccato, P., Trzaska, S., Pérez García-Pando, C., Kalashnikova, O., Del Corral, J., Cousin, R., Blumenthal, M. B., Bell, M., Connor, S. J., and Thomson, M. C. 2014. "Improving Decision-Making Activities for Meningitis and Malaria," *Geocarto International* (29:1), pp. 19-38.

Cheng, C. S., Auld, H., Li, Q., and Li, G. 2012. "Possible Impacts of Climate Change on Extreme Weather Events at Local Scale in South–Central Canada," *Climatic Change* (112:3), pp. 963-979.

Crimmins, A., Balbus, J., Gamble, J. L., Beard, C. B., Bell, J. E., Dodgen, D., Eisen, R. J., Fann, N., Hawkins, M. D., Herring, S. C., Jantarasami, L., Mills, D. M., Saha, S., Sarofim, M. C., Trtanj, J., and Ziska, L. 2016. "The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment," U.S. Global Change Research Program.

Cuartas, J., Bhatia, A., Carter, D., Cluver, L., Coll, C., Donger, E., Draper, C. E., Gardner, F., Herbert, B., Kelly, O., Lachman, J., M'jid, N. M., and Seidel, F. 2023. "Climate Change Is a Threat Multiplier for Violence against Children," *Child Abuse & Neglect*), p. 106430.

Cunsolo, A., and Harper, S. L. 2019. "Editorial Climate Change and Health: A Grand Challenge and Grand Opportunity for Public Health in Canada," *Health Promotion and Chronic Disease Prevention in Canada : Research, Policy and Practice* (39:4), pp. 119-121.

Cunsolo Willox, A., Harper, S. L., Ford, J. D., Landman, K., Houle, K., and Edge, V. L. 2012. ""From This Place and of This Place:" Climate Change, Sense of Place, and Health in Nunatsiavut, Canada," *Social Science & Medicine* (75:3), pp. 538-547.

Cypress, B. S. 2015. "Qualitative Research: The "What," "Why," "Who," and "How"!," *Dimensions of Critical Care Nursing* (34:6), p. 356.

Di Napoli, C., McGushin, A., Romanello, M., Ayeb-Karlsson, S., Cai, W., Chambers, J., Dasgupta, S., Escobar, L. E., Kelman, I., Kjellstrom, T., Kniveton, D., Liu, Y., Liu, Z., Lowe, R., Martinez-Urtaza, J., McMichael, C., Moradi-Lakeh, M., Murray, K. A., Rabbaniha, M., Semenza, J. C., Shi, L., Tabatabaei, M., Trinanes, J. A., Vu, B. N., Brimicombe, C., and Robinson, E. J. 2022. "Tracking the Impacts of Climate Change on Human Health Via Indicators: Lessons from the Lancet Countdown," *BMC Public Health* (22:1), p. 663.

Evans, K. J., Terhorst, A., and Kang, B. H. 2017. "From Data to Decisions: Helping Crop Producers Build Their Actionable Knowledge," *Critical Reviews in Plant Sciences* (36:2), pp. 71-88.

Findlater, K., Webber, S., Kandlikar, M., and Donner, S. 2021. "Climate Services Promise Better Decisions but Mainly Focus on Better Data," *Nature Climate Change* (11:9), pp. 731-737.

Flynn, B. S., Dana, G. S., Sears, J., and Aultman-Hall, L. 2012. "Weather Factor Impacts on Commuting to Work by Bicycle," *Preventive Medicine* (54:2), pp. 122-124.

Georgeson, L., Maslin, M., and Poessinouw, M. 2017. "Global Disparity in the Supply of Commercial Weather and Climate Information Services," *Science Advances* (3:5), p. e1602632.

Ghebreyesus, T. A., Tadesse, Z., Jima, D., Bekele, E., Mihretie, A., Yihdego, Y. Y., Dinku, T., Connor, S. J., and Rogers, D. P. 2009. "Using Climate Information in the Health Sector," *Field Actions Science Reports. The journal of field actions*:Vol. 2).

Hunter, P. R. 2003. "Climate Change and Waterborne and Vector-Borne Disease," *Journal of Applied Microbiology* (94 Suppl), pp. 37S-46S.

IPCC. 2023. *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group Ii Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Jagannathan, K., Pathak, T. B., and Doll, D. 2023. "Are Long-Term Climate Projections Useful for on-Farm Adaptation Decisions?," *Frontiers in Climate* (4), p. 1005104.

Järvenpää, S. L., & Essén, A. 2023. "Data sustainability: Data governance in data infrastructures across technological and human generations," *Information and Organization* (33:1), 100449.

Kanemaru, K., Katzer, R., Hanato, S., Nakamura, K., Matsuoka, H., and Ochiai, H. 2017. "Weather Webcam System for the Safety of Helicopter Emergency Medical Services in Miyazaki, Japan," *Air Medical Journal* (36:2), pp. 71-76.

Kanti, F. S., Alari, A., Chaix, B., and Benmarhnia, T. 2022. "Comparison of Various Heat Waves Definitions and the Burden of Heat-Related Mortality in France: Implications for Existing Early Warning Systems," *Environmental Research* (215), p. 114359.

Kim, D., Cho, S., Tamil, L., Song, D. J., and Seo, S. 2020. "Predicting Asthma Attacks: Effects of Indoor Pm Concentrations on Peak Expiratory Flow Rates of Asthmatic Children," *IEEE Access* (8), pp. 8791-8797.

Koch, M., Matzke, I., Huhn, S., Sié, A., Boudo, V., Compaoré, G., Maggioni, M. A., Bunker, A., Bärnighausen, T., Dambach, P., and Barteit, S. 2023. "Assessing the Effect of Extreme Weather on Population Health Using Consumer-Grade Wearables in Rural Burkina Faso: Observational Panel Study," *JMIR mHealth and uHealth* (11), p. e46980.

Lemos, M. C., Kirchhoff, C. J., and Ramprasad, V. 2012. "Narrowing the Climate Information Usability Gap," *Nature Climate Change* (2:11), pp. 789-794.

Li, H., Yu, G., Dong, C., Jia, Z., An, J., Duan, H., and Shu, Q. 2019. "Pedmap: A Pediatric Diseases Map Generated from Clinical Big Data from Hangzhou, China," *Scientific Reports* (9:1), p. 17867.

Li, P., Yu, Y., Huang, D., Wang, Z.-H., and Sharma, A. 2023. "Regional Heatwave Prediction Using Graph Neural Network and Weather Station Data," *Geophysical Research Letters* (50:7), p. e2023GL103405.

Lowe, R., Gasparrini, A., Van Meerbeeck, C. J., Lippi, C. A., Mahon, R., Trotman, A. R., Rollock, L., Hinds, A. Q. J., Ryan, S. J., and Stewart-Ibarra, A. M. 2018. "Nonlinear and Delayed Impacts of Climate on Dengue Risk in Barbados: A Modelling Study," *PLOS Medicine* (15:7), p. e1002613.

Lowe, R., Stewart-Ibarra, A. M., Petrova, D., García-Díez, M., Borbor-Cordova, M. J., Mejía, R., Regato, M., and Rodó, X. 2017. "Climate Services for Health: Predicting the Evolution of the 2016 Dengue Season in Machala, Ecuador," *The Lancet Planetary Health* (1:4), pp. e142-e151.

Mafwele, B. J., and Lee, J. W. 2022. "Relationships between Transmission of Malaria in Africa and Climate Factors," *Scientific Reports* (12:1), p. 14392.

Majchrzak, A., Markus, M.L. and Wareham, J., 2016. "Designing for digital transformation". MIS quarterly, (40:2), pp. 267-278.

Masselot, P., Mistry, M., Vanoli, J., Schneider, R., Iungman, T., Garcia-Leon, D., Ciscar, J.-C., Feyen, L., Orru, H., and Urban, A. 2023. "Excess Mortality

Attributed to Heat and Cold: A Health Impact Assessment Study in 854 Cities in Europe," *The Lancet Planetary Health* (7:4), pp. e271-e281.

Nissan, H., Simmons, W., and Downs, S. 2022. "Building Climate-Sensitive Nutrition Programmes," *Bulletin of the World Health Organization* (100:01), pp. 70-77.

Owino, V., Kumwenda, C., Ekesa, B., Parker, M. E., Ewoldt, L., Roos, N., Lee, W. T., and Tome, D. 2022. "The Impact of Climate Change on Food Systems, Diet Quality, Nutrition, and Health Outcomes: A Narrative Review," *Frontiers in Climate* (4), p. 941842.

Phadke, R., Manning, C., and Burlager, S. 2015. "Making It Personal: Diversity and Deliberation in Climate Adaptation Planning," *Climate Risk Management* (9), pp. 62-76.

Porter, J. J., and Dessai, S. 2017. "Mini-Me: Why Do Climate Scientists' Misunderstand Users and Their Needs?," *Environmental Science & Policy* (77), pp. 9-14.

Quinn, C., Quintana, A., Blaine, T., Chandra, A., Epanchin, P., Pitter, S., Thiaw, W., Shek, A., Blate, G. M., Zermoglio, F., Pleuss, E., Teka, H., Gudo, E. S., Dissanayake, G., Colborn, J., Trtanj, J., and Balbus, J. 2022. "Linking Science and Action to Improve Public Health Capacity for Climate Preparedness in Lower- and Middle-Income Countries," *Climate Policy* (22:9-10), pp. 1146-1154.

Rodell, M., and Li, B. 2023. "Changing Intensity of Hydroclimatic Extreme Events Revealed by Grace and Grace-Fo," *Nature Water* (1:3), pp. 241-248.

Roy, M., and Venema, H. D. 2002. "Reducing Risk and Vulnerability to Climate Change in India: The Capabilities Approach," *Gender & Development* (10:2), pp. 78-83.

Sarkar, I. N. 2022. "Transforming Health Data to Actionable Information: Recent Progress and Future Opportunities in Health Information Exchange," *Yearbook of Medical Informatics* (31:1), pp. 203-214.

Sarkis-Onofre, R., Catalá-López, F., Aromataris, E., and Lockwood, C. 2021. "How to Properly Use the Prisma Statement," *Systematic Reviews* (10:1), p. 117.

Shea, B., Knowlton, K., and Shaman, J. 2020. "Assessment of Climate-Health Curricula at International Health Professions Schools," *JAMA Network Open* (3:5), p. e206609.

Short, C. A., Renganathan, G., and Lomas, K. J. 2015. "A Medium-Rise 1970s Maternity Hospital in the East of England: Resilience and Adaptation to Climate Change," *Building Services Engineering Research and Technology* (36:2), pp. 247-274.

Skelton, M. 2020. "How Cognitive Links and Decision-Making Capacity Shape Sectoral Experts' Recognition of Climate Knowledge for Adaptation," *Climatic Change* (162:3), pp. 1535-1553.

Stewart-Ibarra, A. M., Rollock, L., Best, S., Brown, T., Diaz, A. R., Dunbar, W., Lippi, C. A., Mahon, R., Ryan, S. J., Trotman, A., Van Meerbeeck, C. J., and Lowe, R. 2022. "Co-Learning During the Co-Creation of a Dengue Early Warning System for the Health Sector in Barbados," *BMJ Global Health* (7:Suppl 7), p. e007842.

Thomson, M. C., Muñoz, Á. G., Cousin, R., and Shumake-Guillemot, J. 2018. "Climate Drivers of Vector-Borne Diseases in Africa and Their Relevance to Control Programmes," *Infectious Diseases of Poverty* (7:1), p. 81.

UNDRR. 2015. "Global Assessment Report on Disaster Risk Reduction 2015 | Undrr."

United Nations. (2006). "Social Justice in an Open World: The Role of the United Nations." Available at https://www.un.org/esa/socdev/documents/ifsd/SocialJustice.pdf, March 9, 2023.

Usmani, M., Brumfield, K. D., Magers, B. M., Chaves-Gonzalez, J., Ticehurst, H., Barciela, R., McBean, F., Colwell, R. R., and Jutla, A. 2023. "Combating Cholera by Building Predictive Capabilities for Pathogenic Vibrio Cholerae in Yemen," *Scientific Reports* (13:1), p. 2255.

Vaughan, C., and Dessai, S. 2014. "Climate Services for Society: Origins, Institutional Arrangements, and Design Elements for an Evaluation Framework," *WIREs Climate Change* (5:5), pp. 587-603.

Venus, T. E., Bilgram, S., Sauer, J., and Khatri-Chettri, A. 2022. "Livelihood Vulnerability and Climate Change: A Comparative Analysis of Smallholders in the Indo-Gangetic Plains," *Environment, Development and Sustainability* (24:2), pp. 1981-2009.

Vescovi, L., Bourque, A., Simonet, G., and Musy, A. 2009. "Transfer of Climate Knowledge Via a Regional Climate-Change Management Body to Support Vulnerability, Impact Assessments and Adaptation Measures," *Climate Research* (40), pp. 163-173.

von Braun, J. 2020. "Climate Change Risks for Agriculture, Health, and Nutrition," *Health of People, Health of Planet and Our Responsibility: Climate Change, Air Pollution and Health*), pp. 135-148.

WHO. 2023a. "Climate Change."

WHO. 2023b. "Protecting Maternal, Newborn and Child Health from the Impacts of Climate Change: A Call for Action."

WHO/WMO. 2023. "Implementation Plan Advancing Integrated Climate, Environment and Health Science and Services 2023-2023."

WMO. 2023. "Climate Change Is Bad for Health but Climate Services Save Lives." *World Meteorological Organization*.

Worrall, E., Connor, S. J., and Thomson, M. C. 2008. "Improving the Cost-Effectiveness of Irs with Climate Informed Health Surveillance Systems," *Malaria Journal* (7:1), p. 263.

The Role of Digital Traceability in Boosting Technological Innovation: Empirical Evidence from Chinese Manufacturing Companies

Xiongyong Zhou
Fuzhou University, School of Economics and Management, China
Haiyan Lu*
Newcastle University Business School
Luca Mora
Edinburgh Napier University Business School, UK
Angelo Natalicchio
Department Of Mechanics, Mathematics & Management, Politecnico Di Bari
Antonio Messeni Petruzzelli
Department Of Mechanics, Mathematics & Management, Politecnico Di Bari

Abstract

Digital traceability is acknowledged to improve visualization in the supply chain. However, it is still unclear to what extent manufacturing companies can enhance technological innovation through digital traceability. From the perspective of knowledge management, the study develops a conceptual model to explore the impact of digital traceability on product and process innovation in manufacturing companies. Using 296 manufacturing companies samples across four digital traceability demonstration provinces in China, hierarchical regression analysis results indicate that, first, the implementation of digital traceability significantly promotes both product and process innovation in manufacturing companies. Second, knowledge absorption serves as a mediating role between digital traceability and technological innovation. Specifically in this regard, digital traceability partially mediating the effect on process innovation and fully mediating the effect on product innovation. Third, market competition moderates the direct link between chain traceability and product innovation and also moderates the latter part of the mediation effects (knowledge absorption-technological innovation).

Keywords: Digital traceability, Product innovation, Process innovation, Knowledge absorption, Market competition, Manufacturing companies

1.0 Introduction

In the aftermath of the pandemic, traditional centralized traceability solutions no longer work well in current product supply chains. This exposes multiple issues, such as data manipulation and reliance on a single point of failure (Sunny et al., 2020). Additionally, the existence of data silos across various links of the supply chain has obstructed efficient information sharing, resulting in information islands and data garbage (Behnke and Janssen, 2019). Therefore, it is crucial to urgently transform the traditional traceability model through digital means. One potential benefit of this transformation driven by digital technology is the stimulation of technological innovation (Hastig and Sodhi, 2020). Technological innovation is key to creating economic value and gaining a competitive advantage for manufacturing companies (Liu et al., 2020). Amid the emergence of new technologies and the evolution of markets, organizations are increasingly embracing innovation as a key strategic differentiator from their competitors (Gunday et al., 2011; Hervas-Oliver et al., 2021). However, it is still unclear whether the digital traceability introduced by manufacturing companies can actually trigger innovation. Given the rapid evolution of digital traceability technologies, supply chain participants often struggle to determine whether investing in digitally-enabled traceability systems is worthwhile (Hew et al., 2020; Zhou et al., 2023). Hence, this research endeavors to empirically analyze the potential of manufacturing companies to spur technological innovation, encompassing both product and process enhancements, through the digital transformation of their traceability systems, and to discern the underlying mechanisms involved. This study supports manufacturing companies in building sustainable competitive advantages, but also contributes to the innovative development of the manufacturing sector within the digital economy.

The field of digital traceability research is still in its early stages when it comes to exploring the connection between digital traceability and technological innovation. Currently, only a limited number of studies have ventured into this territory, offering preliminary insights into the matter (Epelbaum and Martinez, 2014; Hastig and Sodhi, 2020; Shou et al., 2021; Yi et al., 2022), and the empirical evidence in these papers has yet to draw definitive conclusions on the relationship. For example, Hastig and Sodhi (2020) assert that the traceability enabled by blockchain can facilitate rapid product development, thereby promoting business innovation. Epelbaum and Martinez (2014) insisted that the proactive adoption of traceability technology by food companies can result in varying degrees of technological innovation. Shou et al. (2021) found that manufacturing firms can achieve operational innovation, leading to excellent operational performance and customer satisfaction, by appropriately matching traceability and supply chain coordination. Innovation serves as a pivotal driver for sustaining long-term competitive advantage for enterprises (Gunday et al., 2011). However, limited literature has not yet made a clear conclusion on the relationship between traceability and innovation, and little is known about how different dimensions of digital traceability promote technological innovation in areas such as product development or process optimization, particularly in China, where traceability practices have been launched relatively late.

Simply implementing digital traceability may not directly lead to technological innovation. In the current digital era, manufacturing firms require prompt access to internal and external organizational information, and must be able to effectively process this information to attain knowledge that can propel technological advancements and innovations (Cousins et al., 2019). Thus, the ability to effectively absorb dispersed knowledge from both internally and externally may be necessary for manufacturing companies. Knowledge absorption, according to the knowledge management perspective, involves the acquisition, integration, and utilization of heterogeneous knowledge resources (Sjödin et al., 2019). Technological innovation can be facilitated when organizations can obtain and employ information, knowledge, and expertise possessed by their supply chain members, and integrate both new and existing knowledge (Lee et al., 2018). It is essential to recognize a gap that warrants attention, as the knowledge management literature suggests that there is likely to be a mediator between traceability and innovation in tracing and tracking processes. In manufacturing companies, digital traceability may capture vast amounts of real-time information and data, and efficient absorption of this information can support product development and process improvement (Hastig and Sodhi, 2020). As a result, knowledge absorption processes may function as a critical bridge for explaining the connection between digital traceability and technological innovation.

To gain a comprehensive understanding, the relationship between digital traceability and technological innovation cannot be separated from an examination of the market environment (Zhou et al., 2022). Currently, the global supply chain is undergoing deep integration, and the competition within industries and at the international level is becoming increasingly intense. The strong substitutability of any product or process necessitates that manufacturing companies drive all-round innovation through digital management practices. Prior research has documented the importance of market competition (Damanpour, 2010; Feng et al., 2019). When it comes to digital traceability, market competition is a contingency factor that may impact the connections among digital traceability, knowledge absorption, and technological innovation. In highly competitive market environments, market demand is constantly changing, and frequent interactions by consumers on traceability platforms can stimulate new ideas for product development and process optimization by manufacturing companies. Additionally, the process of knowledge absorption is closely related to the external environment. In the knowledge economy era, the rapidly changing and increasingly complex external environment may influence the link between digital traceability practices and technological innovations. Faced with a challenging market environment, manufacturing companies with a high degree of knowledge absorption can better convert scattered information into valuable professional knowledge through digital traceability practices, which may strengthen their innovation advantages (Zhou et al., 2022).

Building upon the aforementioned discussions, we endeavor to tackle the following research question: *To what extent can digital traceability impact technological innovation, with consideration of knowledge absorption and market competition?*

The purpose of this study is to scrutinize the effects that digital traceability exerts on the technological innovation processes within manufacturing companies, with knowledge absorption as the mediator and market competition in the exterrnal environment as the moderator through the lens of the posed research question. Through statistical analysis of effective survey data from 296 manufacturing companies in four demonstration areas of traceability system construction, this study further analyzes the role and boundary conditions of knowledge absorption and market competition in triggering technological innovation through digital traceability. This paper aims to approach the issue from a comprehensive perspective of the interconnected innovation system within organizations, treating knowledge absorption as a key component. By defining the core elements of digital traceability and empirically analyzing the interactions among these elements, the study seeks to gain a deeper understanding of how digital traceability facilitates the process of technological innovation. The anticipated research outcomes are poised to bridge the gaps in current scholarship regarding supply chain traceability and its impact on innovation performance. They will shed light on the heretofore elusive "black box" of the intricate relationship between digital traceability and technological innovation, elucidating the dynamic contextual factors that modulate this relationship. Ultimately, this study offers actionable insights for manufacturing enterprises seeking to bolster their technological innovation.

2.0 Literature Review and Hypotheses Development

2.1 Theoretical Framework

Absorption pertains to an organization's inherent ability to methodically analyze, process, interpret, and comprehend information acquired from external sources (Smuts and Van der Merwe, 2022). Knowledge absorption is a perpetual and dynamic process in which organizations engage in reconfiguring and integrating their existing internal knowledge systems. They also eliminate knowledge that is no longer relevant or has diminished value. Furthermore, they actively acquire new knowledge and potential value from external sources (Duchek, 2015). Knowledge absorption is a crucial element in an enterprise's knowledge management system, as it directly impacts its effectiveness by determining the quantity and quality of available knowledge resources. The more abundant the knowledge resources, the greater the opportunity for the enterprise to integrate, innovate, and apply knowledge (Sjödin et al., 2019). From a knowledge management perspective, the implementation of a new technology in an organization often presents challenges for users to understand and effectively utilize it (Lee and Choi, 2009). By leveraging insights gained from previous experiences with adopting other technologies, users can expedite the process of adopting the newly implemented technology. The knowledge management viewpoint suggests that a firm's knowledge resources are not only derived from internal sources within the organization, but also from external stakeholders such as suppliers, customers, competitors, and regulators (Sjödin et al., 2019). Knowledge assets, such as skills, experience, and technology integrated within the supply chain, are primary drivers of added value and competitive advantage. Effective knowledge absorption is manifested through the integration and optimization of every stage in the knowledge value chain, rendering knowledge a valuable and enduring strategic asset for manufacturing enterprises (Agyabeng-Mensah et al., 2020).

Digital traceability is an essential technology for knowledge search (Engelseth, 2009). It uses various digital technologies to assess a firm's knowledge of product location and processes, from origin to the end customer, in order to inform operational decision-making (Skilton and Robinson, 2009). Full-chain traceability involves sharing knowledge across the supply chain network, including external stakeholders, rather than relying solely on internal knowledge (Cousins et al., 2019). For manufacturing companies, knowledge absorption through digital traceability refers to the utilization of digital technology to effectively integrate, process, and organize extensive quantities of traceability information that is dispersed across supply chain networks. This process may generate valuable knowledge and experiences that enrich both product and process innovation (Zahra and George, 2002). Therefore, An in-depth assessment of an organization's capacity to absorb knowledge is crucial for comprehending the role of digital traceability in driving technological innovation (Abbas, 2020). In order for a company to attain exceptional performance through its knowledge assets, the seamless exchange of valuable knowledge both internally and externally is of utmost importance. To obtain a comprehensive grasp of knowledge management, it is imperative to integrate perspectives on both product and process innovation (Khan et al., 2022). Knowledge absorption is a mechanism that allows firms to absorb and utilize knowledge resources in an innovative way, considering both product development and process optimization (Abbas, 2020). Traceability allows for the acquisition of a vast amount of valuable data from the entire supply chain. Certain pieces of this information have the potential to directly inform operational decision-making, while others necessitate further analysis and incorporation into existing knowledge in order to cultivate expertise that can enhance operational practices (Cousins et al., 2019). However, merely gathering fragmented knowledge through digital traceability does not automatically result in enhanced product and process innovation in manufacturing companies. It is essential for manufacturing companies to thoroughly integrate and assimilate knowledge across organizational boundaries to contribute to both types of innovation and maximize existing and new knowledge (Smuts and Van der Merwe, 2022). Therefore, manufacturing companies must actively promote knowledge absorption to cultivate valuable expertise that drives improved technological innovation.

In a competitive market environment, information is abundant and complex, posing challenges in its collection and effective processing (Tsai and Yang, 2013). Depending on one-sided or outdated information for decision-making can result in operational failures (Soto-Acosta et al., 2018). Therefore, it is imperative for manufacturing companies to prioritize the acquisition, integration, and utilization of information and knowledge in order to make well-informed decisions and generate value within a highly competitive market (Song and Yang, 2019). Digital traceability can provide technical support for accessing and integrating information into valuable knowledge (Aung and Chang, 2014). In a competitive market, manufacturing companies can capture more information through digital traceability practices to promote technological innovation while also absorbing knowledge to facilitate operational decisions and improve technological innovation. Drawing from the knowledge management perspective, a theoretical model, as shown in Fig.1, is presented herein in order to investigate the interconnections among digital traceability, knowledge absorption, market competition, and technological innovation.

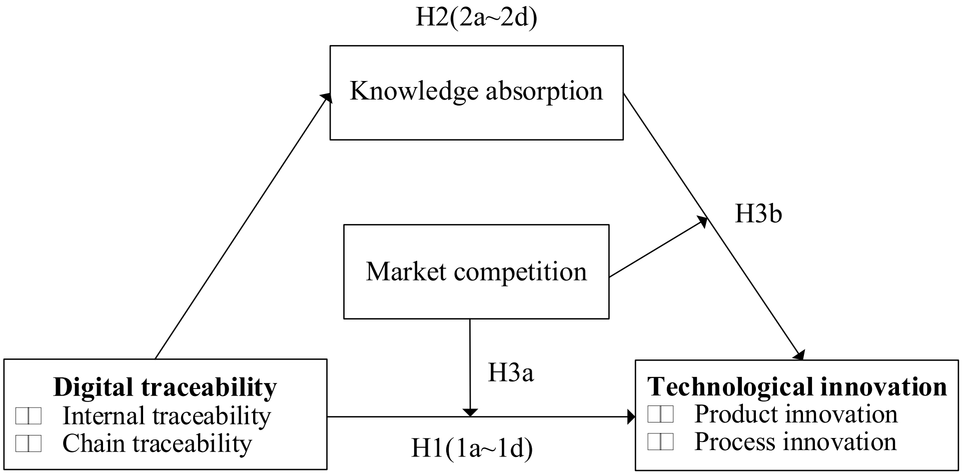


Figure 1. Theoretical framework.

2.2 Digital Traceability and Technological Innovation

2.2.1 Digital Traceability

As a crucial endeavor for manufacturing firms to adopt digital transformation, digital traceability entails the utilization of emerging digital technologies, such as the Internet of Things, artificial intelligence, machine learning, blockchain, and cloud computing, to monitor the movement of a product across the supply chain (Gillani et al., 2020).Through the capture of diverse data, this technology empowers consumers and stakeholders within the supply chain to discern and track the historical background, manufacturing origin, and distribution trajectory of a given product. The implementation of digital traceability endeavors to bolster efficient, productive, and transparent supply chains by facilitating the exchange of information and the surveillance of product quality (Hastig and Sodhi, 2020). However, the value of digital traceability extends beyond quality assurance. These systems possess the capacity to collect and analyze data related to supply chains, thereby providing manufacturing companies with crucial knowledge and insights obtained from both internal and external sources. This amalgamated data is consolidated and made available for all functions to see, making it much easier for sourcing, production, legal, and compliance teams to collaborate and act on the information, thereby fostering technological innovation in companies (Casino et al., 2021).

Regarding the extensional definition of digital traceability, multiple categorizations have been proposed in the existing literature (Moe, 1998; Olsen and Borit, 2013). This paper adopts the widely recognized categorization proposed by (Moe, 1998), which divided traceability practices into two dimensions: internal traceability and chain traceability. Internal traceability encompasses the diligent surveillance of a manufacturing company's internal production processes, procedures, and interconnections. Chain traceability, also known as external traceability, refers to an external traceability network composed of supply chain node enterprise information sharing with raw materials or products as the traceability point. A well-established blockchain traceability system can achieve information sharing for all nodes, including internal and chain, and ensure data integrity (Cui et al., 2023).

2.2.2 Technological Innovation

Innovation is a crucial process that involves creating and applying new ideas or behaviors (Gunday et al., 2011). Technological innovations, in particular, involve products, services, and production process technologies (Schilling, 2005), which are closely linked to the primary work activities of organizations. This study centers on technological innovation within manufacturing firms, which can be further classified into two categories: product and process innovations (Damanpour, 2010). Product innovations involve introducing new products or services that meet the needs of external users, while process innovations involve incorporating new elements into a company's production or service operation to enhance product quality or service delivery (Bessant et al., 2005). Product innovations affect the organization's external offerings (Damanpour and Gopalakrishnan, 2001), while process innovations improve production efficiency or effectiveness and can lead to cost reductions (Schilling, 2005).

2.2.3 The Direct Impact of Digital Traceability on Technological Innovation

Digital traceability can provide a source of information to advance product innovation. Specifically, within the context of internal traceability in manufacturing companies, digital traceability management can facilitate the enhancement of quality management approaches, the utilization of quality monitoring tools, and the development of structured problem-solving capabilities among cross-functional teams (Zeng et al., 2015). These improvements, in turn, can expedite the adoption of new technologies and products, thereby fostering product innovation at scale (Hong et al., 2019). Moreover, digital traceability systems are essential for addressing the imperative for improved data collection throughout the production, processing, and distribution stages, enabling manufacturing companies to conduct internal data analyses for predictive and data-driven decision-making purposes (Casino et al., 2021). Additionally, in the context of chain traceability, the use of blockchain-enabled traceability may enhance the connectivity among supply chain members and provide manufacturing companies with access to valuable intelligence related to sales patterns for new products or upstream supply disruptions, enabling rapid response (Hastig and Sodhi, 2020).

Early involvement and knowledge sharing with key suppliers can facilitate new product development (Song and Yang, 2019). A digital traceability system enables manufacturing companies to collaborate with suppliers across different tiers in real-time, allowing them to exchange information about their supply chain, products, and materials (Casino et al., 2021). Complex information may provide valuable support for suppliers in generating new ideas, technologies, and products by suppliers, while also increasing their willingness to adjust their production processes or product technology prototypes in response to evolving manufacturer demands and market changes, thereby providing better services to manufacturers for innovation (Lee et al., 2021). Moreover, traceability systems create a platform for communication, interaction, and innovative suggestions from consumers (Alfaro and Rábade, 2009). This allows manufacturing companies to grasp the latest feedback on consumer product demands, providing new insights and market positioning for the development of products (Hastig and Sodhi, 2020). Customer requirements serve as the primary source of external knowledge and innovation for enterprises. By utilizing traceability technology and leveraging big data analysis, manufacturing companies can gain comprehensive insights into the current and future needs and expectations of customers and effectively meet their various functional, quality, and performance requirements (Behnke and Janssen, 2020). Manufacturing companies ensure that consumers have access to the most comprehensive quality information possible by promptly sharing and updating relevant traceability data (Zhou et al., 2022).

Digital traceability not only promotes product innovation, but also drives continuous improvement and renewal of business processes. First of all, digital traceability technology provides manufacturing companies with the ability to share information across different functional departments and effectively organize and analyze the collected data, internalizing the interpreted information as knowledge for the basis of internal production and management processes (Sunny et al., 2020). For example, keeping information flowing and closely collaborating between R&D and production departments can significantly improve the efficiency of product design and manufacturing. Secondly, digital traceability promotes the application of emerging information technologies in transactions between enterprises, which helps to promote process innovation (Shou et al., 2021). If products can be more easily identified and traced, and greater transparency provided in their production methods, then participants are more motivated to improve their processes (Kim et al., 2012). Driven by intelligent sensors, AI controllers, and advanced data management software, digital traceability systems can make automated decisions based on collected data, optimizing equipment and process efficiency, including predicting maintenance (Hastig and Sodhi, 2020). Lastly, digital traceability can also optimize factory operations and simplify business processes by sharing information with suppliers and customers (Shou et al., 2021), thereby reducing inventory, minimizing stockouts, and shortening delivery cycles. By identifying non-value-added processes through such effective use of quality information, manufacturing companies can continuously enhance their business process management (Corallo et al., 2020). Given the preceding discussion, the following hypothesis is proposed:

H1: Digital traceability is positively related to technological innovations.

H1a: Internal traceability is positively related to product innovation.

H1b: Internal traceability is positively related to process innovation.

H1c: Chain traceability is positively related to product innovation.

H1d: Chain traceability is positively related to process innovation.

2.3 The Mediation of Knowledge Absorption

2.3.1 Knowledge Absorption

Grant (1996) acknowledged knowledge as a pivotal strategic resource, whereas the knowledge management perspective prescribes a set of processes aimed at converting data into knowledge or useful information to drive organizational progress (Duchek, 2015). A key element of knowledge management is knowledge absorption, which involves the acquisition, assimilation, transformation, and application of external knowledge (Purvis et al., 2001).

Corporate innovation results from the combination of different knowledge elements (Kamasak et al., 2017). However, not all information or knowledge is valuable to a company, and information needs to be effectively absorbed to serve the enterprise's needs (Lane et al., 2006). For effective innovation and achieving a competitive edge, it is essential to absorbe both existing and new information into useful knowledge (Cohen and Levinthal, 1990). Enterprise knowledge systems can be optimized by actively engaging in knowledge absorption activities, which involve promoting the intentional flow of knowledge along the supply chain to effectively integrate both existing and novel knowledge. Sjödin et al. (2019) pointed out that acquiring knowledge relies not only on extensively exploring a firm's internal resources but also on the vast amount of data and information produced through interactions with external stakeholders. Firms are required to undertake the crucial task of filtering, organizing, and absorbing information resources into valuable knowledge assets.

Knowledge absorption is crucial in the process of innovation, serving as both input and output (Khan et al., 2022). Innovation-related knowledge can be acquired from multiple sources, including internally developed knowledge embedded within the organization, which serves as a key source of competitive advantage (Lee and Choi, 2009). However, research in strategy has demonstrated the substantial contribution of knowledge from external sources to a firm's success as well (Hervas-Oliver et al., 2021). To drive performance improvement, firms must integrate existing and new knowledge into valuable knowledge. Empirical studies on knowledge absorption have established a positive association with innovation (Abbas, 2020; Shahzad et al., 2020; Smuts and Van der Merwe, 2022). Therefore, manufacturing companies need to implement robust strategies for knowledge absorption that facilitate the smooth acquisition and integration of essential knowledge across the supply chain. This approach will enable firms to transform their business models and become more innovative (Kavalić et al., 2021). For manufacturing companies, effectively leveraging digital traceability to absorb knowledge involves utilizing digital technologies to access, assimilate, consolidate, and apply vast amounts of traceability data distributed across supply chain partners.. Such a process may produce new knowledge and experience that support product development and process enhancement within the enterprise.

2.3.2 The Mediating Effect of Knowledge Absorption on the Digital Traceability-Technological Innovation Link

The digital traceability system is an esteemed investment in terms of knowledge and technological resources (Engelseth et al., 2014). It functions as a platform for the reciprocal exchange of information and knowledge between focal companies and their partners. This system fosters the advancement of organizational routines based on optimal practices and can be utilized to establish mechanisms for learning and facilitate innovative endeavors (Zhou et al., 2022). A digital traceability system constitutes a strategic investment in knowledge and technological resources, creating a platform that facilitates the bidirectional flow and sharing of information and knowledge between focal companies and their partners (Song and Yang, 2019). This system encourages the development of organizational routines formed by best practices and can be leveraged to establish learning mechanisms and support innovative activities (Zhou et al., 2022).

Hastig and Sodhi (2020) suggested that focal companies need to recognize that promoting blockchain adoption necessitates a focus on knowledge development. Knowledge management enhances a firm's operational efficiency by employing traceability systems that enable rapid and dependable access to its existing knowledge stocks. Engelseth (2009) found that maintaining traceability of various knowledge elements used during the product development phase can facilitate knowledge sharing and effective absorption of knowledge to achieve product innovation. Song and Yang (2019) suggested that implementing food traceability practices can enable food companies to gain more creative knowledge and experience. Existing literature also supports the idea that suppliers and customers play a crucial role in acquiring knowledge within the supply chain (Kamasak et al., 2017). By absorbing knowledge from different partners and applying it to business practices, manufacturing companies can foster more innovative thinking and approaches (Zahra and George, 2002).

Digital traceability is expected to drive technological innovation, emphasizing the importance of an efficient knowledge absorption process. Merely possessing vast and complex information resources is insufficient; the ability to effectively absorb novel knowledge is crucial for firms' technological innovation.

The process of absorbing knowledge is vital in advancing technological innovation in manufacturing firms (Liu et al., 2020; Yam et al., 2011), providing a necessary mechanism to transform digital traceability resources into innovation benefits. The effective acquisition and utilization of external innovation resources, such as knowledge and technology, are crucial to developing new products and improving process efficiency for firms. From a knowledge management perspective, effectively managing knowledge requires its integration into daily work activities within organizations, which in turn facilitates the development of new products and the improvement of business process efficiency, ultimately boosting organizational effectiveness. By organically combining newly acquired knowledge with existing knowledge, and continuously accumulating and absorbing knowledge, firms can improve their technological innovation (Khan et al., 2022). As digital traceability practices deepen, the knowledge obtained and integrated by manufacturing companies becomes more abundant, and this knowledge can help them construct a technological innovation to respond to changing market demands (Kamasak et al., 2017). Epelbaum and Martinez (2014) emphasized that traceability information and knowledge require effective processing and handling to form professional expertise and experience. When advancements in traceability technology are converted into a knowledge asset and a competitive edge, they can lead to improved performance.

Kamasak et al. (2017) argued that knowledge absorption should be viewed as a critical strategic resource for innovation activities, encompassing closely related product and process innovations. Technological innovation is primarily fueled by an organization's capacity to integrate knowledge by forming new combinations and relationships within its existing knowledge base, as well as by merging newly acquired knowledge with what it already possesses (Engelseth et al., 2014). Absorbing knowledge through these methods can increase the value of the knowledge created, thereby driving new product development and process improvement (Aliasghar et al., 2020). Within a traceability supply chain, effective sharing and exchange of traceability data and diverse knowledge among supply chain members can invigorate collective learning. This process promotes the internalization of information, enriching the organization's knowledge base and expanding its intellectual capital. As Matta et al. (2013) has thoroughly analyzed, this approach not only broadens the organization's knowledge reserves but also reinforces the strength of its intellectual assets. Innovation thus emerges as an interactive social exchange process, where various professional knowledge resources are shared, merged, and assimilated (Lane et al., 2006). For instance, effective processing and integration of traceability information and technological resources by manufacturing companies can lead to useful knowledge that predicts market demand changes (Hastig and Sodhi, 2020). By sharing information technology, professional knowledge, and applied research with partners, scattered information within and outside the organization can be identified, digested, transformed, and utilized to build new knowledge and experience (Malhotra et al., 2005), during which digital traceability implemented by manufacturing companies could trigger new product development and innovative business processes. Engelseth et al. (2014) proposed that companies should utilize the knowledge resources generated, including cutting-edge traceability technologies, to strengthen their knowledge management and thereby boost their innovation advantage.

Synthesising the above literature, it is hypothesized that:

H2: Knowledge absorption mediates the relationship between digital traceability and technological innovations.

H2a: Knowledge absorption mediates the relationship between internal traceability and product innovation.

H2b: Knowledge absorption mediates the relationship between internal traceability and process innovation.

H2c: Knowledge absorption mediates the relationship between chain traceability and product innovation.

H2d: Knowledge absorption mediates the relationship between chain traceability and process innovation.

2.3.3 The Moderation of Market Competition

Market competition refers to the extent of competition among firms within an industry that are competing for limited market resources (Tsai and Yang, 2013). The globalization of trade has resulted in increased competition among incumbents, new business models, and markets (Damanpour, 2010). Greater external market competition leads to significant changes in customer needs. Consequently, manufacturing companies must proactively identify and promptly respond to market information, adapting measures to address market changes. Market competition may also moderate the direct impact of digital traceability on technological innovation. In highly competitive markets, blockchain traceability could improve supply chain responsiveness to market movements or trends and increase its resilience against market disruptions (Hastig and Sodhi, 2020). Manufacturing companies can leverage digital traceability to effectively monitor, search, acquire, and analyze potential customer needs and feedback (Casino et al., 2021), enabling them to capture fleeting market opportunities and make decisions that are conducive to new product development and process improvement (Santos-Vijande and Álvarez-González, 2007).

In highly competitive market environments, digital traceability activities of manufacturing companies may face competition in their existing business areas (Dai et al., 2021), which compels managers to take timely actions such as continuously upgrading current production processes and technologies, thereby expanding their innovation advantages (Shou et al., 2021). Conversely, when the intensity of market competition is weak, manufacturing companies prioritize efficiency and production capacity over product improvements and process changes (Zhou et al., 2022).

Market competition may play a moderating role in the path from knowledge absorption to technological innovation. As competition intensifies, manufacturing companies face increased external competitive pressures and uncertainties, prompting them to seek information from external stakeholders such as supply chain members or competitors to gain a competitive edge (Song and Yang, 2019). At this time, digital traceability implementation by manufacturing companies can increase the depth of knowledge search, enabling them to effectively absorb and utilize information for technological innovation improvement and respond to changes in market demand through innovation (Hastig and Sodhi, 2020). However, in relatively stable market competition, manufacturing companies may receive similar information and knowledge from external sources, leading to homogenization of their understanding of issues and filtering out differentiated information. This may cause them to be content with the status quo and lack the motivation to develop new products or optimize their business (Zhou et al., 2022).

Building on the previous points, it is hypothesized that:

H3a: Market competition positively moderates the linkage between digital traceability and technological innovations.

H3b: Market competition positively moderates the linkage between knowledge absorption and technological innovations.

3.0 Methodology

3.1 Survey instruments

This research employed a questionnaire-based survey methodology to gather empirical data. In order to examine our hypotheses, we constructed an initial survey instrument grounded in prior research, which was subsequently modified in light of feedback from industry managers who considered the current conditions and specific characteristics of the Chinese manufacturing sector.

The eleven items used to measure digital traceability were derived from two previous empirical studies (Cousins et al., 2019; Zhou et al., 2022), and encompassed two distinct domains: internal traceability and chain traceability. Participant were requested to evaluate the degree to whose organization has implemented each digital traceability practice across its entire supply chain. The digital traceability items were measured using a five-point Likert scale, with those survey responses ranging from 1 (not taken into consideration) to 5 (successfully implemented).

This study utilized a measurement instrument comprised of five items that assessed knowledge absorption, which were adapted from a prior investigation (Lane et al., 2006). Respondents were asked to assess the level of knowledge absorption they observed within their respective organizations. Their answers to the knowledge absorption items were recorded on a five-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree).

Market competition was evaluated using a five-item measurement tool, derived from a prior investigation (Jansen et al., 2006). Participant were queried about their perceptions on the competitive landscape external to their organization. Responses to the market competition items were evaluated using a five-point Likert scale, from 1 (strongly disagree) to 5 (strongly agree).

Drawing on a previous study (Gunday et al., 2011), this study conceptualized technological innovation as encompassing two dimensions: product innovation and process innovation. Respondend were asked to assess the degree to whose organization had pursued technological innovation across these two dimensions, using a five-point Likert scale that ranged from 1 (very low extent) to 5 (very high extent).

Control variables employed in this study were primarily informed by insights gained during interviews with industry managers in the process of developing the survey instrument. Control variables consisted of three factors: firm age (the duration of the company’s operation), size (measured by the number of employees), and ownership structure. Ownership was classified into three categories: state-owned, privately-owned, and foreign or joint ventures, with each group represented by its respective classification dummy variables.

Initially, the survey questionnaire was created in English. A team of native English and Chinese speakers then translated the instruments into Chinese, ensuring that the meaning was accurately conveyed using the standard back-translation method (Brislin, 1980). To further ensure the questionnaire’s relevance and comprehensibility to the Chinese respondents, we consulted experts in the fields of digital traceability practices, knowledge absorption, market competition, and technological innovation. The team of experts who reviewed the initial draft of the questionnaire comprised five senior researchers with significant academic expertise in supply chain traceability, three government officials responsible for developing product traceability strategies, and six experienced professionals with years of involvement in traceability operations with their companies receiving traceability system certification from the China Quality Certification Center as early as 2017. Using the experts' valuable feedback, we refined the core constructs to enhance the scale's relevance and measurement precision in the Chinese context. This process involved three rounds of detailed face-to-face discussions, during which we meticulously followed the experts' advice, particularly focusing on improving question clarity and avoiding potential ambiguity.

3.2 Data collection

Located in a global manufacturing centre, many Chinese firms have experienced rapid market changes and more and more challenges in promoting technological innovations and advancing digital traceability related to their supply chain partners. With the development of national strategies such as the 14th Five-Year Plan and the Vision 2035 outline, China's manufacturing industry is undergoing a transformation towards efficient automation, digitalisation and intelligence. Driven by modern technology and under the background of expanding opening up, Chinese manufacturing companies are using the latest digital technology to promote all-round innovation, such as digital traceability technologies. Drawing on sound digital traceability practices to promote high-level technological innovation has become a crucial strategy for China's development of an innovation supply chain. As a result, survey data from China is highly relevant for addressing our research questions.

Specifically, this study collected data from several provinces, including Shandong, Shanghai, Ningxia, and Fujian, which have taken the lead in launching pilot programs for traceability systems supported by digital technologies since 2016. In the same year, the General Office of the State Council of China released the "Opinions on Accelerating the Construction of Traceability System for Important Products," which clearly identified the nationwide implementation of product quality traceability system construction across seven critical product categories, including food, pharmaceuticals, edible agricultural products, agricultural production materials, special equipment, rare earths, and dangerous goods, as a necessary task for all regions of the country. Moreover, the document urged manufacturers to take the forefront in the creation of the traceability system. After a period of four years devoted to traceability initiatives, several manufacturing companies in the aforementioned regions have successfully implemented digital traceability pilot programs. Hence, these four regions were chosen as the focus for gathering data efforts. The local departments of commerce in these targeted areas furnished a directory containing the contact information of 2,176 manufacturing companies, as well as a letter of endorsement. We employed a stratified sampling methodology to select 500 manufacturing companies from this inventory for inclusion in our study. These surveyed firms represent a diverse range of industries, such as food, agriculture, pharmaceuticals, and medicine, and so on.

The intended participants were top or middle managers in charge of purchasing, operations, production, logistics, or IT within manufacturing companies. Respondents were expected to have a solid grasp of their company's overall situation, particularly regarding digital traceability practices, knowledge absorption, market competition, and technological innovations. Anonymity was guaranteed throughout the survey process, and all participants received a confidentiality statement. During this period, the questionnaires were distributed between July and November of 2019, and were sent via email along with prepaid return envelopes. Each questionnaire also included a website link for respondents to complete the survey online. To prompt non-respondents to complete and return the survey, follow-up actions such as reminder emails, phone calls, and site visits were implemented.

In total, 324 questionnaires were distributed via email, online posting, and site visits, with an additional 25 questionnaires being sent via traditional mail. Following exclusion criteria related to incomplete or inadequately filled-out questionnaires (i.e., those with more than 6 blanks), we ended up with 296 questionnaires suitable for analysis. These responses were obtained from a variety of sources, including 103 from Shanghai, 76 from Shandong, 63 from Ningxia, and 54 from Fujian. Descriptive information pertaining to the age, size, and ownership of the surveyed firms is summarized in Table 1.

Category	Sample	Percentage (%)
Firm age (years)		
Less than 5	34	11.49%
5-10	125	42.23%
11-30	101	34.12%
More than 30	36	12.16%
Size (employees)		
Less than 201	54	18.24%
201-500	95	32.09%
501-1200	42	14.19%
More than 1200	105	35.47%
Ownership		
State-owned	91	30.74%

Private	176	59.46%
Foreign or joint ventures	29	9.80%

Table 2. Profile of sample firms.

3.3 Non-response bias and common method bias (CMB)

To determine the likelihood of non-response bias, a t-test was performed to compare the questionnaires from early (209) and late (87) respondents, following the approach recommended by Armstrong and Overton (1977). The results of the t-test indicated that there were no significant differences ($p > 0.05$) in the average scores of all constructs and items between the two groups. Therefore, the findings of this study are unlikely to have been influenced by non-response bias, allowing us to conclude with confidence.

To assess the potential for common method bias (CMB), this study utilized Harman's one-factor test following the recommendation of Podsakoff et al. (2003). The results indicated that the largest factor could only account for 35.282% of the variance, suggesting that no single factor could account for the majority of the variation in the data. A confirmatory factor analysis of the one-factor model was performed to further examine the presence of CMB. The results revealed a poor fit for the model, with a chi-square statistic of $\chi2$ (326) = 2914.600, CFI = 0.463, TLI = 0.422, and RMSEA = 0.164. Given these results, it is reasonable to infer that CMB is unlikely to have significantly influenced the findings of this study.

4.0 Statistical Analysis and Results

4.1 Factor Analysis and Results

To identify the underlying dimensions (factors) of digital traceability, knowledge absorption, market competition, and innovation performance, an exploratory factor analysis (EFA) with varimax rotation and Kaiser normalization was performed on the survey data. The scree test and the initial eigenvalue test both indicated the presence of two factors for digital traceability, which together explained 72.308% of the total variance. Table 2 displays the loadings of the digital traceability items, each of which has a high loading (above 0.60) on one factor and low loadings (below 0.30) on the other factors, further supporting the construct validity of the identified factors. Based on item characteristics, the two digital traceability factors were labeled as internal traceability and chain traceability. To ensure the internal consistency and validity of the latent constructs, a reliability test was conducted on items within the same factor, with a benchmark value of 0.70 (Nunnally et al., 1978). The reliability coefficient values for the two digital traceability factors were high, at 0.917 and 0.915 for internal traceability and chain traceability, respectively.

Item description	Factors	
	1	2
Know the processes involved in producing products across the complete supply chains.	0.785	0.211
Improve digital records and storage of traceability information.	0.803	0.200
Establish mechanisms and tools for timely integration of traceability information.	0.813	0.312
Optimize whole-process data management and digital analysis.	0.762	0.232
Establish a cross-functional team for digital product traceability management.	0.830	0.201
Provide training on digital traceability for the supply chain.	0.837	0.239
Trace the source of our purchases throughout entire supply chains.	0.299	0.789
Track product distribution channel and sales process.	0.279	0.822
Quickly exchange and transmit traceability information.	0.205	0.868
Collect feedback from customers regarding updates and improvements to key products.	0.137	0.843
Share information related to market demand trends and forecasts.	0.297	0.806

Note. Extraction method: principal component analysis. Rotation method: varimax with Kaiser normalization.

[a] Rotation converged in two iterations.

Table 2. Rotated component matrix[a] on digital traceability.

This study employed a consistent method to investigate the prospective factors of knowledge absorption and market competition, which were found to account for 67.714% and 63.094% of the inherent variation, separately. Table 3 and Table 4 respectively present the factors underlying these two constructs. Furthermore, the reliability coefficient alpha values for knowledge absorption and market competition were determined to be high, measuring at 0.913 and 0.895, respectively.

Item description	Factor
Quickly identifying valuable new knowledge.	0.803
Promptly recognize how new technological knowledge may bring changes to the organization.	0.836
Absorb new knowledge and make it useful for the organization.	0.824

Promptly introducing process innovations based on new technological knowledge	0.857
Quickly using absorbed new technology for new product development.	0.793

Note: Extraction method: principal component analysis.

[a] One component extracted.

Table 3. Rotated component matrix[a] on knowledge absorption.

Item description	Factor
The competition in our local market is intense.	0.787
Our customers demand new products on a regular basis.	0.808
Upon the introduction of a new product, other companies in the industry quickly follow suit.	0.796
Rapid responses to the actions of peer competitors are required.	0.800
The quantity of products required by the market fluctuates rapidly and frequently.	0.781

Note: Extraction method: principal component analysis.

[a] One component extracted.

Table 4. Rotated component matrix[a] on market competition.

The results of the EFA revealed two factors for technological innovations, with the loadings of all items presented in Table 5. These two factors were labeled as product innovation and process innovation, accounting for 74.163% of the total inherent variance. The reliability coefficient alpha values for the two types of innovation factors were determined to be high, measuring at 0.902 and 0.863, respectively.

Item description	Factors	
	1	2
Develop new products with a higher degree of novelty than existing products	0.845	0.200
Develop new products that are more competitive in the market	0.871	0.143
Develop new products with great market potential	0.888	0.133
Update or improve production and business processes	0.175	0.837
Update or improve tools and equipment	0.193	0.830
Update or improve the application of information technology	0.091	0.803

Note. Extraction method: principal component analysis. Rotation method: varimax with Kaiser normalization.

[a] Rotation converged in three iterations.

Table 3. Rotated component matrix[a] on technological innovations.

Based on the findings from the EFA, as presented in Table 6, all composite reliability (CR) values were determined to be above 0.7, indicating an ideal level of model quality. The average variance extracted (AVE) values for each construct were found to be above

0.5, demonstrating a high degree of convergent validity. Additionally, Table 7 indicates that the square root of the AVE for each construct surpassed the correlation coefficient of any two variables, providing support for discriminant validity (Fornell and Larker, 1981).

Constructs	Reliability	CR (> 0.6)	AVE (> 0.5)
Internal traceability	0.915	0.917	0.649
Chain traceability	0.913	0.915	0.682
Knowledge absorption	0.879	0.913	0.677
Market competition	0.852	0.895	0.631
Product innovation	0.858	0.902	0.754
Process innovation	0.788	0.863	0.678

Table 6. Reliability and validity assessment.

4.2 Descriptive Analysis and Results

Table 7 provides a comprehensive overview of the mean, standard deviation, and correlation coefficient of each variable. The correlation observed among the key constructs are in alignment with the research hypothesis, which offers initial empirical support for the validation of the proposed hypotheses.

	Mean	S.D.	1	2	3	4	5	6
1. Internal traceability	3.680	0.586	(0.806)[a]					
2. Chain traceability	3.596	0.543	0.537**	(0.826)				
3. Knowledge absorption	3.634	0.550	0.452**	0.440**	(0.823)			
4. Market competition	3.701	0.610	0.362**	0.195**	0.480**	(0.794)		
5. Product innovation	3.592	0.731	0.269**	0.277**	0.365**	0.277**	(0.868)	
6. Process innovation	3.658	0.652	0.476**	0.408**	0.418**	0.363**	0.357**	(0.823)

Note: Pearson correlations.
N = 296; * p < 0.05, ** p < 0.01.
[a] Square root of AVE reported along diagonal in bold.

Table 7. Descriptive statistics and correlations.

4.3 Regression Analysis and Results

4.3.2 The Mediating Effect Test

To examine the mediating function of knowledge absorption in the connection between digital traceability and technological innovation, the causal steps approach was applied. As shown in Table 8, the results reveal that internal traceability is significantly and positively related to both product innovation (M2, $\beta = 0.203$, $p < 0.05$) and process innovation (M4, $\beta = 0.387$, $p < 0.001$). Similarly, chain traceability is found to be significantly and positively linked to product innovation (M2, $\beta = 0.245$, $p < 0.01$) and process innovation (M4, $\beta = 0.252$, $p < 0.01$), respectively. These findings provide support for H1.

Both internal traceability (M1, $\beta = 0.261$, $p < 0.001$) and chain traceability (M1, $\beta = 0.289$, $p < 0.001$) have a positive impact on knowledge absorption. Moreover, knowledge absorption demonstrates a positive correlation with product innovation (M3, $\beta = 0.345$, $p < 0.001$) and process innovation (M5, $\beta = 0.243$, $p < 0.01$). However, we found that the direct effects of internal traceability (M3, $\beta = 0.113$, n.s) and chain traceability (M3, $\beta = 0.145$, n.s) on product innovation were not significant, suggesting that knowledge absorption fully mediates the relationship between digital traceability and product innovation, providing full support for H2a and H2c. On the other hand, the direct effects of internal traceability (M5, $\beta = 0.323$, $p < 0.001$) and chain traceability (M5, $\beta = 0.182$, $p < 0.05$) on process innovation were found to be lessened but still significant, indicating that the effects of digital traceability on process innovation are partially mediated by knowledge absorption, partially supporting H2b and H2d.

This study utilized the Bootstrapping method to validate the mediating role of knowledge absorption in the relationship between digital traceability and technological innovation. As shown in Table 9, the direct effects of digital traceability on process innovation were statistically significant (excluding zero) within the 95% confidence interval, and the indirect effects mediated by knowledge absorption were also significant (excluding zero) within the same interval. Therefore, H2a and H2c were further confirmed. On the other hand, the direct effects of digital traceability on product innovation were not significant (including zero) within the 95% confidence interval, while the indirect effects through knowledge absorption were significant (excluding zero). As a result, H2b and H2d were fully supported.

	Knowledge absorption	Product innovation		Process innovation	
	M1	M2	M3	M4	M5
Firm year	-0.007**	-0.006	-0.004	-0.001	0.001
Size	-0.037	-0.125	-0.112	-0.103	-0.094
State-owned	0.165	0.109	0.052	0.243*	0.203
Private	0.022	0.051	0.043	0.150	0.144
Internal traceability	0.261***	0.203*	0.113	0.387***	0.323***
Chain traceability	0.289***	0.245**	0.145	0.252**	0.182*
Knowledge absorption			0.345***		0.243**
R2	0.293	0.123	0.171	0.276	0.305
Adjusted R2	0.278	0.105	0.151	0.261	0.288
F	19.972***	6.761***	8.472***	18.321***	18.085***

Note. *** p < 0.001, ** p < 0.01, * p < 0.05.

Table 8. The mediating effect test of knowledge absorption (N = 296).

Path	Effect	BootSE	Boot LLCI	Boot ULCI
Total				
Internal traceability → Product innovation	0.119	0.049	0.023	0.215
Chain traceability → Product innovation	0.133	0.049	0.038	0.228
Internal traceability → Process innovation	0.226	0.039	0.149	0.304
Chain traceability → Process innovation	0.137	0.039	0.060	0.214
Indirect				
Internal traceability → Product innovation	0.053	0.019	0.018	0.093
Chain traceability → Product innovation	0.054	0.026	0.014	0.112
Internal traceability → Process innovation	0.037	0.018	0.009	0.079
Chain traceability → Process innovation	0.038	0.015	0.011	0.071
Direct				
Internal traceability → Product innovation	0.066	0.049	-0.030	0.163

Chain traceability → Product innovation	0.079	0.049	-0.018	0.175
Internal traceability → Process innovation	0.189	0.040	0.110	0.268
Chain traceability → Process innovation	0.099	0.040	0.020	0.178

Table 9. The mediating effect of knowledge absorption.

4.3.2 The Moderated Mediating Effect Test

Given the potential role of market competition as a crucial moderator in multiple pathways, we conducted a test of the moderated mediating model.

According to Table 10, the primary analysis involved assessing the moderating impacts on the direct path between digital traceability and technological innovation. The interaction term (internal traceability * market competition) did not demonstrate any significant effects on product innovation (M6, β = 0.061, n.s) and process innovation (M9, β = 0.064, n.s). The interaction term (chain traceability * market competition) was found to be insignificant in relation to process innovation (M10, β = 0.023, n.s); however, it was significantly and positively associated with product innovation (M7, β = 0.092, $p < 0.05$), indicating a significant and positive moderating effect on the link between chain traceability and product innovation, thereby partially supporting H3a. Secondly, this study examined the moderating effects on the second half of the indirect path, which encompasses knowledge absorption and technological innovation. The findings indicated significant positive effects of the interaction term (knowledge absorption * market competition) on both product innovation (M8, β = 0.096, $p < 0.01$) and process innovation (M14, β = 0.126, $p < 0.001$), thereby indicating a significant and positive moderating effect on the link between knowledge absorption and technological innovation, thus partially supporting H3b.

In summary, market competition serves as a moderator in both the direct connection between chain traceability and process innovation and in the second half of the indirect path, which includes knowledge absorption and product/process innovation within the mediation model.

	Product innovation	Process innovation

	M6	M7	M8	M9	M10	M11
Firm year	-0.005	-0.005	-0.004	0.001	0.000	0.001
Size	-0.096	-0.105	-0.095	-0.075	-0.080	-0.073
State-owned	0.060	0.098	0.029	0.195	0.202	0.174
Private	0.030	0.052	0.018	0.129	0.132	0.112
Internal traceability	0.113	0.097	0.023	0.318***	0.303***	0.207**
Chain traceability	0.229**	0.258**	0.199*	0.235**	0.256***	0.252**
Market competition	0.220**	0.226**	0.145*	0.217***	0.217***	0.182**
Internal traceability * market competition	0.061			0.064		
Chain traceability * market competition		0.092*			0.023	
Knowledge absorption			0.309**			0.200**
Knowledge absorption * market competition			0.096**			0.126***
R2	0.157	0.168	0.199	0.318	0.311	0.365
Adjusted R2	0.134	0.145	0.174	0.299	0.292	0.345
F	6.690***	7.230***	7.907***	16.749***	16.188***	18.278***

Note. *** p < 0.001, ** p < 0.01, * p < 0.05.

Table 10. The moderated mediating effect test.

Following the simple slope method introduced by Aiken et al. (1991), we plotted a diagram to represent the moderating effect. As seen in Figs. 2~4, low market competition is indicated as one standard deviation below the mean, and high market competition as one standard deviation above the mean.

Additionally, the study examined the moderating influence on the direct link (i.e., chain traceability-product innovation) as depicted in Fig. 2. Results demonstrated that the relationship between chain traceability and product innovation was insignificant in a low-competition market environment (β = 0.047, n.s). However, when market competition is high, chain traceability exhibits a significant positive relationship with product innovation (β = 0.232, $p < 0.001$), indicating an upward trend in the predictive effect of chain traceability on product innovation as market competition increases.

Additionally, the study examined the moderating effect on the direct pathway (i.e., chain traceability-product innovation) as depicted in Fig. 2. Results demonstrated that chain traceability was insignificantly related to product innovation when market competition is low (β = 0.047, n.s). However, when market competition is high, chain traceability has a significant positive relationship with product innovation (β = 0.232, p < 0.001), indicating an upward trend in the predictive effect of chain traceability on product innovation as market competition increases.

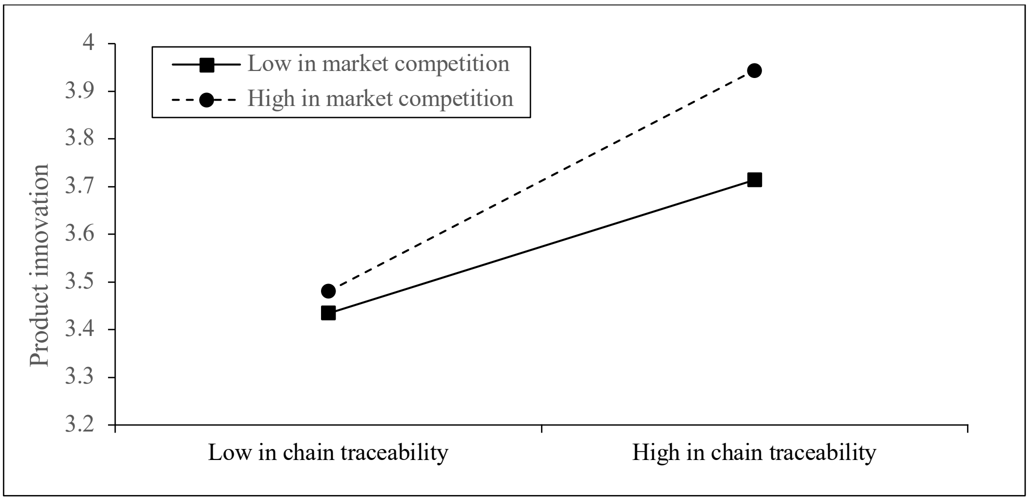


Figure 2. Moderating effect of market competition on the chain traceability-product innovation link.

Additionally, the research investigated the moderating effects on the second phase of the indirect relationships, specifically the connections between knowledge absorption and product innovation, and knowledge absorption and process innovation. As illustrated in Figure 3, the results showed that knowledge absorption is only weakly associated with product innovation when market competition is low (β = 0.074, n.s), but becomes significantly more positively related to product innovation under high competition (β = 0.266, p < 0.001).

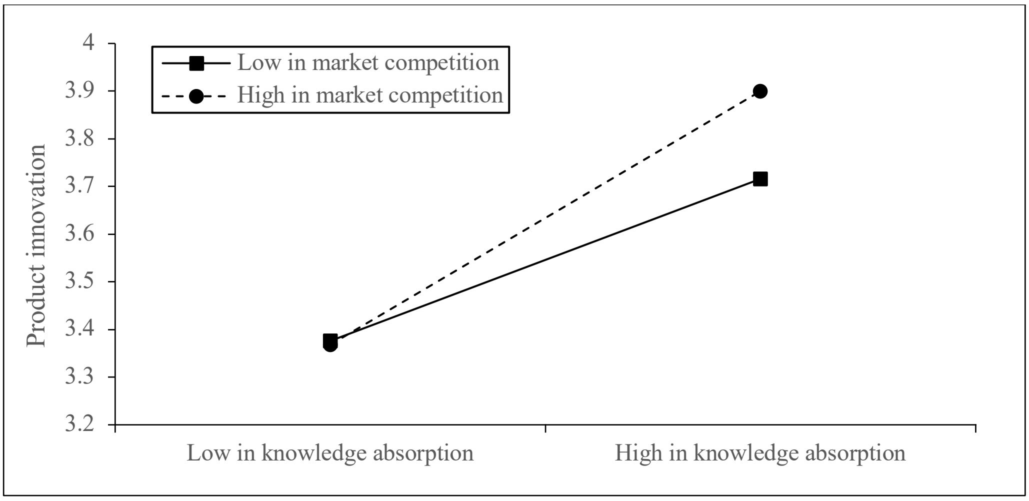


Figure 3. Moderating effect of market competition on the knowledge absorption-product innovation link.

According to Fig. 4, the relationship between knowledge absorption and process innovation is insignificant in a low-competition environment (β = -0.016, n.s), but becomes significantly positive in a high-competition environment (β = 0.236, $p < 0.001$).

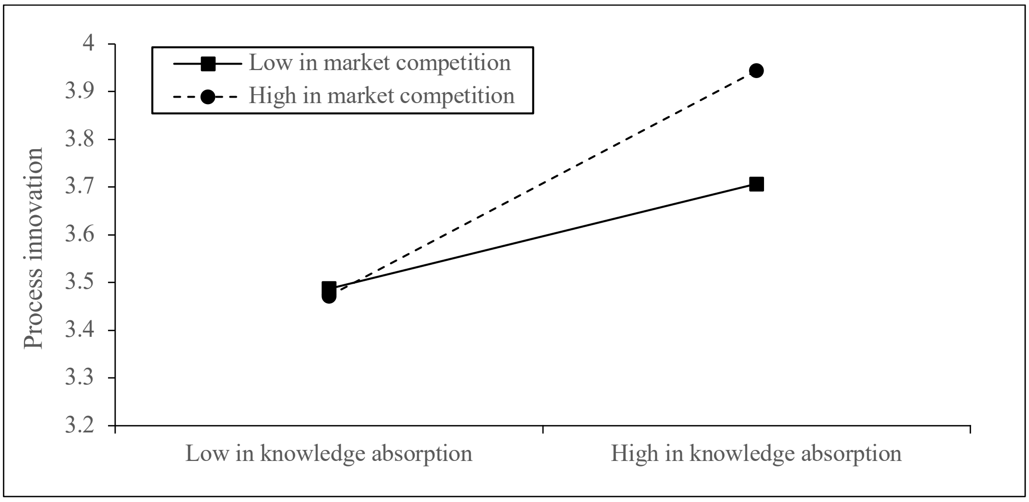


Figure 4. Moderating effect of market competition on the knowledge absorption-process innovation link.

Furthermore, Table 11 demonstrates the significance of the mediating effect after introducing the moderating variables. The results demonstrate that the mediating effect of knowledge absorption in the relationship between digital traceability and technological innovation remains significant at all three market competition levels, exhibiting a rising trend. These findings imply that as market competition intensifies, digital traceability becomes more effective in driving product and process innovation by improving knowledge absorption.

Path		Market competition	Effect	BootSE	Boot LLCI	Boot ULCI
Internal traceability → Product innovation	Direct	M-1SD	0.015	0.062	-0.106	0.137
		M	0.013	0.052	-0.090	0.116
		M+1SD	0.010	0.075	-0.138	0.158
	Indirect	M-1SD	0.021	0.023	-0.033	0.060
		M	0.047	0.019	0.012	0.086
		M+1SD	0.074	0.029	0.025	0.138
Chain traceability → Product innovation	Direct	M-1SD	0.124	0.049	0.028	0.221
		M	0.120	0.042	0.038	0.202
		M+1SD	0.116	0.060	-0.002	0.233
	Indirect	M-1SD	-0.005	0.015	-0.032	0.030
		M	0.031	0.016	0.005	0.067
		M+1SD	0.066	0.026	0.024	0.124
Internal traceability → Process innovation	Direct	M-1SD	0.024	0.064	-0.103	0.151
		M	0.105	0.049	0.009	0.202
		M+1SD	0.187	0.063	0.064	0.310
	Indirect	M-1SD	0.031	0.030	0.013	0.101
		M	0.051	0.026	0.010	0.110
		M+1SD	0.071	0.028	0.022	0.131
Chain traceability → Process innovation	Direct	M-1SD	0.154	0.051	0.053	0.255
		M	0.137	0.039	0.060	0.214
		M+1SD	0.121	0.050	0.022	0.219
	Indirect	M-1SD	-0.007	0.016	-0.040	0.024
		M	0.031	0.016	0.005	0.067
		M+1SD	0.068	0.029	0.022	0.132

Table 11. The moderated mediating effect.

5.0 Discussion and Implication

5.1 Main Findings

Existing research highlights the critical role of traceability in driving innovation within manufacturing companies. Our empirical investigation demonstrates that the adoption of digital traceability by manufacturing companies can facilitate both product and process innovation. Our results are consistent with Shou et al. (2021), who asserted that the implementation of traceability initiatives can promote operational innovation; but they did not examine the impact of specific traceability practices on different technological innovations. This study extends the work of Shou et al. (2021) by highlighting that chain traceability contributes more to product innovation, while internal traceability plays a larger role in driving process innovation. This is because manufacturing companies can leverage traceability platforms to gather feedback or demand from external stakeholders on product updates or improvements, or to share information with key supply chain members on market demand trends and forecasts to generate new product ideas that will satisfy consumers. In contrast, internal traceability focuses on digital traceability training for the supply chain, enabling companies to acquire diverse knowledge from different domains, and promoting creative and innovative solutions to enhance processes. Digital traceability management through cross-functional teams can help companies reorganize business processes and identify inherent flaws, thereby fostering process optimization and innovation.

The implementation of digital traceability in manufacturing companies does not exclusively generate technological innovation through its direct impact, as knowledge absorption may serve as a potential mediator in this relationship. The most salient finding in this research is that knowledge absorption plays a complete mediating role in the impact of digital traceability on product innovation, addressing a gap in the existing literature. Our result opposes the argument made by (Epelbaum and Martinez, 2014), who argued that the use of traceability technology application leads directly to product innovation. This is because traceability technology alone does not directly facilitate feedback or information on product and process improvements. From the viewpoint of knowledge management, the primary purpose of knowledge management in organizations is to strengthen their capacity to gather essential knowledge resources and efficiently manage their absorption and utilization.Manufacturing companies invest significantly in traceability to encourage technological innovation, but they fail to concentrate on efficiently assimilating and integrating this data or information into valuable knowledge and solutions, which may result in limited product innovation.

Our findings provide empirical evidence that knowledge absorption partially mediates the link between digital traceability and process innovation. In other words, digital traceability-driven process innovation can be achieved, in part, through the efficient absorption of knowledge. This finding supports the observation noted by (Garcia-Torres et al., 2019) in the literature review domain, which suggests that the integration and digestion of large amounts of information obtained through traceability techniques with pre-existing knowledge can contribute to traceability-related technological innovation. In line with knowledge management theory, the dissemination of knowledge and expertise can foster the integration and cross-fertilization of ideas. For companies to enhance their performance, it is essential to combine newly acquired knowledge with what they already know (Lane et al., 2006). Effective knowledge absorption enables the processing and integration of real-time data from tracking systems, helping manufacturing companies seize market opportunities and increase the probability of process innovation (Wowak et al., 2016).

The findings suggest that market competition amplifies the positive impact of chain traceability on product innovation. Specifically, in highly competitive markets, manufacturing companies that advance their chain traceability practices are likely to experience significant improvements in product innovation, with little effect on process innovation. These results align, to some extent, with the observations made by (Song and Yang, 2019), who argued that firm performance in traceability practices may be influenced by the external competitive environment. However, our study extends their work by demonstrating that digital traceability practices can effectively facilitate product innovation in manufacturing companies through the moderating role of market competition. Such finding further suggests the relationship between digital traceability and product innovation may be more sensitive to external environmental changes than it is for process innovation. In highly competitive markets, manufacturing companies are likely to prioritize customer retention, which can be achieved by dynamically acquiring and responding to customer needs through traceability systems. Consequently, traceability systems have become an indispensable means of ensuring customer satisfaction and loyalty.

Statistical evidence confirms that market competition serves as an essential moderator in the connection between knowledge absorption and technological innovation among manufacturing companies. Specifically, our findings indicate that a highly competitive market amplifies the beneficial influence of knowledge absorption on product and process innovation. Our result supports the findings of (Zhou et al., 2022), which showed that environmental dynamism can enhance the impact of dynamic capabilities on organizational performances in the food industry. Drawing on knowledge management theory, we argue that knowledge absorption is particularly vulnerable to external market demands and technological changes. In response to market competition, firms typically seek to improve their ability to integrate diverse knowledge and develop innovative knowledge systems (Soto-Acosta et al., 2018). We posit that the acquisition and assimilation of extensive information volumes demand formalized learning processes. In this regard, we suggest that in highly competitive market environments, manufacturing companies must invest more in digital traceability resources to enhance their knowledge absorption capabilities, allowing them to better adapt to external environmental changes and avoid being outpaced by competitors. By leveraging this capability, manufacturing companies can innovate new products and refine their business processes, resulting in long-term competitive advantages. Therefore, we conclude that market competition provides a contingency mechanism for the conversion of absorptive capacity into technological innovation within the digital traceability processes of manufacturing companies.

5.2 Theoretical implications

This study aims to examine the role of digital traceability practices within the context of digital transformation. It also investigates how these practices impact product and process innovation in resolving technological innovations within the manufacturing system. The theoretical contributions of this study can be summarized in three main aspects.

Firstly, our study builds upon the findings of (Shou et al., 2021) by expanding the examination of the linkage between traceability and innovations using a large sample size. Additionally, we propose a comprehensive definition for digital traceability that takes into account both categories of practices. Through empirical validation, we establish the differential relationships between digital traceability practices and specific types of technological innovation, thereby making a valuable contribution to the existing literature on supply chain traceability. Our research advances the understanding of the value of traceability by delving deeper into traceability performance and elucidating the intrinsic mechanisms through which digital traceability practices impact technological innovation. Consequently, our findings provide new avenues for the development of theories on supply chain traceability.

Secondly, this paper contributes to a deeper understanding of the relationship between digital traceability, knowledge absorption in manufacturing companies, and two types of innovation. It draws on the knowledge management perspective to address the need for further investigation into the “black box” between traceability and organizational performance (Zhou et al., 2022). By complementing existing evidence in the supply chain traceability literature, which shows a significant and positive relationship between digital traceability and product innovation, we highlight the crucial role of knowledge absorption. It is of great interest to examine in more detail how knowledge absorption links digital traceability with technological innovation. Our findings offer valuable insights for firms on how knowledge absorption strategies can enhance the effectiveness of their digital traceability efforts in promoting technological innovation, particularly in emerging markets where traceability is still in its early stages.

Lastly, the study also contributes to the discussion of digital traceability and technological innovation by exploring the dynamic situational conditions of market competition, given that traceability practices essentially arise from the market's demand for quality transparency from manufacturing companies. Unlike previous studies, such as those by Song and Yang (2019) and Zhou et al. (2022), which focused solely on the moderating effect of environmental dynamism on the indirect pathways of the traceability-performance relationship, our research highlights the moderating role of market competition on both the direct and indirect impact pathways between digital traceability practices and technological innovation. These findings suggest that in highly competitive markets, digital traceability practices and knowledge absorption by manufacturing companies can enhance and reshape various types of innovation. Accordingly, we offer insights into the impact mechanism and contextual limitations of digital traceability practices within dynamic situational conditions.

Overall, this study establishes a foundational framework for researchers interested in comprehending the role of the digital traceability process and its influence on technological innovation within emerging markets such as China, where traceability practices are still in an exploratory phase. Our research aims to empirically validate the potential relevance of digital traceability to technological innovation in manufacturing firms. By doing so, we contribute new practices to the field of supply chain traceability and emerging IT research in the context of Industry 4.0. We emphasize the significance of investing in digital traceability resources and strategies for acquiring knowledge within highly competitive market environments, as this will facilitate firm-level technological innovation and ultimately lead to sustainable competitive advantages.

Overall, our study provides preliminary groundwork for scholars seeking to better understand how digital traceability processes contribute to technological innovation in the Chinese context, as an emerging market where traceability practices are still in the exploratory stage. Specifically, by empirically validating the potential relevance of digital traceability to technological innovation in manufacturing companies, our research offers innovative practices in the field of supply chain traceability for emerging information technology research under Industry 4.0. Our research points to the necessity of investing in digital traceability and knowledge absorption strategies in highly competitive market environments to foster firm technological innovation, ultimately leading to a sustainable competitive advantage.

5.3 Managerial implications

In the digital economy era, digitization and intelligence have become important trends driving transformation and development across all industries, particularly in the manufacturing industry. Digital traceability, by deeply integrating the concept of traceability with emerging information technologies, can significantly bring innovative advantages to manufacturing companies.

Traceability systems offer effective technological support for enterprise innovation. The findings of our study demonstrate a significant positive relationship between digital traceability and technological innovation. Consequently, we encourage manufacturing enterprises to take proactive internal traceability practices to promote process optimization and seek broader chain traceability practices to facilitate product development, ultimately forming sustainable innovation advantages by examining their own traceability resources and conditions. Combining the expanded capabilities of traceability with digital technology research and development to promote supply chain visualization can unlock the immense potential of manufacturing enterprises in product and process improvement. For instance, blockchain traceability systems possess technical characteristics such as anti-counterfeiting, tamper-proofing, and traceability, which can address issues in equipment management, data sharing, trust among multiple parties, and security within manufacturing. These systems play a critical role in enhancing industrial production efficiency, reducing production costs, and improving supply chain collaboration levels for business process innovation.

However, manufacturing companies must recognize that merely implementing a traceability plan is insufficient for achieving strong innovation performance without enhancing knowledge absorption. This study found that knowledge absorption is a necessary condition for manufacturing companies to implement digital traceability to trigger technological innovation. Hence, developing suitable knowledge absorption processes within a company's management structure is essential to support innovation efforts. We encourage manufacturing companies to leverage traceability technology to acquire external knowledge from stakeholders, including customers, suppliers, and competitors. This external knowledge can be internalized through collective learning, enriching the company's existing knowledge base and resources, ultimately fostering innovative ideas.

For instance, we advocate for manufacturing companies to deepen collaboration with a variety of stakeholders, traditional and non-traditional alike, in order to enhance their absorptive capacity through knowledge and technology exchange, thus driving better innovation performance. Besides widely absorbing feedback from external stakeholders to continuously improve product quality and production processes, it is also crucial for managers to cultivate a culture of continuous learning within the organization, to foster an environment conducive to technological innovation in the realm of digital traceability. Companies require the establishment of a formal or informal learning mechanism within their organization to effectively absorb knowledge, identify valuable information along the supply chain, and uncover novel ideas and approaches for product development and process improvement. Learning processes such as establishing internal systems for sharing knowledge and transforming tacit knowledge into explicit knowledge are key. The public sector could assist firms in identifying these learning requirements and promoting the development of broad, transferable learning skills.

China's manufacturing industry is currently undergoing rapid market changes, facing unprecedented challenges in strengthening its digital transformation and technological innovation capabilities. Manufacturing companies should remain highly vigilant and proactively respond to external market environments, dynamically adjusting their digital traceability and knowledge absorption capabilities to effectively drive technological innovation. With the rise of fierce market competition, manufacturing companies need to adopt highly flexible traceability strategies to promote innovation. On the one hand, empirical results from this study suggest that market competition strengthens the positive connection between chain traceability and product innovation. This means that when market competition is intense, the product innovation triggered by implementing chain traceability can be enhanced. Therefore, manufacturing companies need to remain vigilant to market shifts, dynamically adjusting the level of knowledge absorption to match specific contexts, and maximize the improvement of their technological innovation. Manufacturing firms must intensify their chain traceability practices and pursue fresh ideas for product innovation by engaging in information exchange with supply chain partners at both ends. On the other hand, this study revealed that market competition strengthens the positive link between knowledge absorption and technological innovation. This finding indicates that with high market competition, leveraging traceable information, knowledge, and skills from supply chain partners may facilitate innovation in both products and processes. Therefore, in highly competitive external environments, manufacturing companies need to maintain a highly elastic and holistic knowledge absorption strategy to keep pace with rapid external changes. By timely searching for and acquiring external information and effectively processing and utilizing it, manufacturing companies can form professional knowledge, skills, and experience that help with product development and process optimization, thereby enhancing their innovation capabilities. Furthermore, the measures developed in this study offer a valuable benchmark for organizations' knowledge absorption activities, helping practitioners improve their product and process innovation efforts. Additionally, in market environments with lower competition intensity, manufacturing companies should continue to introduce, integrate, and utilize external knowledge in their internal knowledge creation processes while maintaining their existing advantages to achieve incremental innovation.

6.0 Conclusion

Focusing on Industry 4.0 in China, this paper examines how manufacturing companies trigger their technological innovation through digital traceability, drawn upon the perspective of knowledge management. This study also investigates the mediating role of knowledge absorption and the moderating effect of market competition on the relationship between digital traceability practices and technological innovation. A sample of 296 manufacturing companies from four demonstrative areas of traceability system construction in China is employed. By investigating the mechanisms by which digital traceability drives technological innovation in manufacturing companies, this study highlights the value of traceability and offers empirical data from emerging markets to deepen and broaden research on supply chain traceability. The main research findings are as follows:

Firstly, promoting digital traceability in manufacturing companies can bring positive improvements in technological innovation. Of these, chain traceability has a greater positive effect on product innovation than internal traceability, while internal traceability has a stronger positive effect on process innovation relative to chain traceability.

Secondly, the technological innovation brought about by digital traceability practices implementation in manufacturing companies is not entirely due to its direct effects; knowledge absorption could partially mediate the relationship between them. More specifically, digital traceability's impact on process innovation exhibits a partial mediation effect, while the impact on product innovation exhibits a complete mediation effect.

Thirdly, market competition serves as a positive moderator in the link between chain traceability and product innovation. Faced with highly competitive market environments, promoting chain traceability significantly enhances product innovation, but does not affect process innovation.

Finally, the more competitive the market environment, the more manufacturing companies' knowledge absorption can promote improvements in both product and process innovation. Market competition serves as a contingency mechanism in the digital traceability context of manufacturing companies, allowing absorptive capacity to be translated into technological innovation. Our empirical results provide decision support for manufacturing companies, demonstrating how their digital traceability practices can promote technological innovation by better implementing knowledge absorption in competitive market environments.

This research presents a few limitations that require attention in future studies. Firstly, the sample mainly consists of supply chain-related companies, and it is uncertain whether the research findings are applicable to other industries. Notably, leading companies from industries such as clothing, automobiles, and pharmaceuticals have also started building digital traceability systems, and examining the mechanisms by which these firms improve their traceability practices through technological innovation would enhance research on supply chain traceability. Secondly, industry and category segmentation could play a crucial role in shaping the application of digital traceability and technological innovation, and future studies could incorporate them as control variables into the theoretical model. This research, in its third finding, verifies that knowledge absorption mediates the relationship between digital traceability and technological innovation, with a single-dimensional scale applied for its measurement. Future research could further examine different dimensions of absorptive capabilities (such as potential and realized absorptive capabilities) and their distinct effects on the connection between digital traceability and technological innovation, fully revealing that absorptive capability is a necessary process for promoting innovation in digital traceability practices. Fourthly, as a cross-sectional study, this research cannot observe the dynamic changes in technological innovation that are spurred or promoted by digital traceability in manufacturing companies. Future research could use dynamic models to analyze the sustained dynamic changes in technological innovation before and after the implementation of digital traceability in manufacturing companies, to fully reveal the long-term and short-term effects of technological innovation.

References

Abbas, J., (2020). Impact of total quality management on corporate sustainability through the mediating effect of knowledge management. Journal of Cleaner Production, 244, 118806.

Agyabeng-Mensah, Y., Afum, E., Acquah, I.S.K., Dacosta, E., Baah, C., Ahenkorah, E., (2020). The role of green logistics management practices, supply chain traceability and logistics ecocentricity in sustainability performance. The International Journal of Logistics Management, 32, 538-566.

Aiken, L.S., West, S.G., Reno, R.R., (1991). Multiple regression: testing and interpreting interactions. Sage.

Alfaro, J.A., Rábade, L.A., (2009). Traceability as a strategic tool to improve inventory management: A case study in the food industry. International Journal of Production Economics, 118, 104-110.

Aliasghar, O., Sadeghi, A., Rose, E.L., (2023). Process innovation in small-and medium-sized enterprises: The critical roles of external knowledge sourcing and absorptive capacity. Journal of Small Business Management, 61(4):1583-610.

Armstrong, J.S., Overton, T.S., (1977). Estimating nonresponse bias in mail surveys. Journal of Marketing Research, 14, 396-402.

Aung, M.M., Chang, Y.S., (2014). Traceability in a food supply chain: Safety and quality perspectives. Food Control, 39, 172-184.

Behnke, K., Janssen, M., (2020). Boundary conditions for traceability in food supply chains using blockchain technology. International Journal of Information Management, 52, 101969.

Behnke, K., Janssen, M.F.W.H.A., (2019). Boundary conditions for traceability in food supply chains using blockchain technology. International Journal of Information Management , 52, 101969.

Bessant, J., Lamming, R., Noke, H., Phillips, W., (2005). Managing innovation beyond the steady state. Technovation, 25, 1366-1376.

Brislin, R.W., (1980). Translation and content analysis of oral and written materials, in: Triandis, H.C., Berry, J.W.E. (Eds.), Handbook of Cross-cultural Psychology (pp. 389–444). Boston, MA: Allyn Bacon.

Casino, F., Kanakaris, V., Dasaklis, T.K., Moschuris, S., Stachtiaris, S., Pagoni, M., Rachaniotis, N.P., (2021). Blockchain-based food supply chain traceability: A case study in the dairy sector. International Journal of Production Research, 59, 5758-5770.

Cohen, W.M., Levinthal, D.A., (1990). Absorptive capacity: A new perspective on learning and innovation. Administrative Science Quarterly, 128-152.

Corallo, A., Latino, M.E., Menegoli, M., Pontrandolfo, P., (2020). A systematic literature review to explore traceability and lifecycle relationship. International Journal of Production Research, 58, 4789-4807.

Cousins, P.D., Lawson, B., Petersen, K.J., Fugate, B., (2019). Investigating green supply chain management practices and performance: The moderating roles of supply chain ecocentricity and traceability. International Journal of Operational Production Management, 39, 767-786.

Cui, Y., Hu, M., Liu, J., (2023). Value and design of traceability-driven blockchains. Manufacturing & Service Operations Management, 25(3):1099-116.

Dai, B., Nu, Y., Xie, X., Li, J., (2021). Interactions of traceability and reliability optimization in a competitive supply chain with product recall. European Journal of Operational Research, 290, 116-131.

Damanpour, F., (2010). An integration of research findings of effects of firm size and market competition on product and process innovations. British Journal of Management 21, 996-1010.

Damanpour, F., Gopalakrishnan, S., (2001). The dynamics of the adoption of product and process innovations in organizations. Journal of Management Studies, 38, 45-65.

Duchek, S., (2015). Designing absorptive capacity? An analysis of knowledge absorption practices in German high-tech firms. International Journal of Innovation Management, 19, 1550044.

Engelseth, P., (2009). Food product traceability and supply network integration. Journal of Business & Industrial Marketing, 24, 421-430.

Engelseth, P., Wongthatsanekorn, W., Charoensiriwath, C., (2014). Food product traceability and customer value. Global Business Review, 15, 87S-105S.

Epelbaum, F.M.B., Martinez, M.G., (2014). The technological evolution of food traceability systems and their impact on firm sustainable performance: A RBV approach. International Journal of Production Economics, 150, 215-224.

Feng, T., Wang, D., Lawton, A., Luo, B.N., (2019). Customer orientation and firm performance: The joint moderating effects of ethical leadership and competitive intensity. Journal of Business Research, 100, 111-121.

Fornell, C., Larker, D., (1981). Structural equation modeling and regression: guidelines for research practice. Journal of Marketing Research, 18, 39-50.

Garcia-Torres, S., Albareda, L., Rey-Garcia, M., Seuring, S., (2019). Traceability for sustainability – literature review and conceptual framework. Supply Chain Management: An International Journal, 24, 85-106.

Gillani, F., Chatha, K.A., Jajja, M.S.S., Farooq, S., (2020). Implementation of digital manufacturing technologies: antecedents and consequences. International Journal of Production Economics, 229, 107748.

Grant, R.M., (1996). Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. Organization Science, 7, 375-387.

Gunday, G., Ulusoy, G., Kilic, K., Alpkan, L., (2011). Effects of innovation types on firm performance. International Journal of Production Economics, 133, 662-676.

Hastig, G.M., Sodhi, M.S., (2020). Blockchain for supply chain traceability: Business requirements and critical success factors. Production And Operations Management, 29, 935-954.

Hervas-Oliver, J.-L., Sempere-Ripoll, F., Boronat-Moll, C., (2021). Technological innovation typologies and open innovation in SMEs: Beyond internal and external sources of knowledge. Technological Forecasting and Social Change, 162, 120338.

Hew, J.-J., Wong, L.-W., Tan, G.W.-H., Ooi, K.-B., Lin, B., (2020). The blockchain-based Halal traceability systems: a hype or reality? Supply Chain Management: An International Journal, 25, 863-879.

Hong, J., Liao, Y., Zhang, Y., Yu, Z., (2019). The effect of supply chain quality management practices and capabilities on operational and innovation performance: Evidence from Chinese manufacturers. International Journal of Production Economics, 212, 227-235.

Jansen, J.J.P., Bosch, F.A.J.V.D., Volberda, H.W., (2006). Exploratory innovation, exploitative innovation, and performance: Effects of organizational antecedents and environmental moderators. Management Science, 52, 1661-1674.

Kamasak, R., Yozgat, U., Yavuz, M., (2017). Knowledge process capabilities and innovation: Testing the moderating effects of environmental dynamism and strategic flexibility. Knowledge Management Research & Practice, 15, 356-368.

Kavalić, M., Nikolić, M., Radosav, D., Stanisavljev, S., Pečujlija, M., (2021). Influencing factors on knowledge management for organizational sustainability. Sustainability (Switzerland), 13(3), 1-18.

Khan, A., Tao, M., Li, C., (2022). Knowledge absorption capacity's efficacy to enhance innovation performance through big data analytics and digital platform capability. Journal of Innovation & Knowledge, 7, 100201.

Kim, D.-Y., Kumar, V., Kumar, U., (2012). Relationship between quality management practices and innovation. Journal of Operations Management, 30, 295-315.

Lane, P.J., Koka, B.R., Pathak, S., (2006). The reification of absorptive capacity: A critical review and rejuvenation of the construct. Academy of Management Review, 31, 833-863.

Lee, J.-N., Choi, B., (2009). Determinants of knowledge management assimilation: An empirical investigation. IEEE Transactions on Engineering Management, 57, 430-449.

Lee, V.-H., Foo, P.-Y., Tan, G.W.-H., Ooi, K.-B., Sohal, A., (2021). Supply chain quality management for product innovation performance: insights from small and medium-sized manufacturing enterprises. Industrial Management & Data Systems, 121, 2118-2142.

Lee, V.-H., Ooi, K.-B., Chong, A.Y.-L., Sohal, A., (2018). The effects of supply chain management on technological innovation: The mediating role of guanxi. International Journal of Production Economics, 205, 15-29.

Liu, J., Chang, H., Forrest, J.Y.-L., Yang, B., (2020). Influence of artificial intelligence on technological innovation: Evidence from the panel data of china's manufacturing sectors. Technological Forecasting and Social Change, 158, 120142.

Malhotra, A., Gosain, S., Sawy, O.A.E., (2005. Absorptive capacity configurations in supply chains: Gearing for partner-enabled market knowledge creation. MIS Quarterly, 145-187.

Matta, N., Ducellier, G., Djaiz, C., (2013). Traceability and structuring of cooperative knowledge in design using PLM. Knowledge Management Research and Practice, 11, 53-61.

Moe, T., (1998). Perspectives on traceability in food manufacture. Trends in Food Science & Technology, 9, 211-214.

Nunnally, J.C., Bernstein, I.H., Berge, J.M.F., (1978. Psychometric Theory (2nd eds). New York: McGraw-Hill.

Olsen, P., Borit, M., (2013). How to define traceability. Trends in Food Science & Technology, 29, 142-150.

Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y., Podsakoff, N.P., (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology, 88, 879-903.

Purvis, R.L., Sambamurthy, V., Zmud, R.W., (2001). The assimilation of knowledge platforms in organizations: An empirical investigation. Organization science, 12, 117-135.

Santos-Vijande, M.L., Álvarez-González, L.I., (2007). Innovativeness and organizational innovation in total quality oriented firms: The moderating role of market turbulence. Technovation, 27, 514-532.

Schilling, M., (2005). Strategic Management of Technological Innovation. McGraw-Hill Irwin, New York.

Shahzad, M., Qu, Y., Zafar, A.U., Rehman, S.U., Islam, T., (2020). Exploring the influence of knowledge management process on corporate sustainable performance through green innovation. Journal of Knowledge Management, 24, 2079-2106.

Shou, Y., Zhao, X., Dai, J., Xu, D., (2021). Matching traceability and supply chain coordination: Achieving operational innovation for superior performance. Transportation Research Part E: Logistics and Transportation Review, 145, 102181.

Sjödin, D., Frishammar, J., Thorgren, S., (2019). How individuals engage in the absorption of new external knowledge: A process model of absorptive capacity. Journal of Product Innovation Management, 36, 356-380.

Skilton, P.F., Robinson, J.L., (2009). Traceability and normal accident theory: how does supply network complexity influence the traceability of adverse events? Journal of Suppy Chain Management, 45, 40-53.

Smuts, H., Van der Merwe, A., (2022). Knowledge Management in Society 5.0: A Sustainability Perspective. Sustainability (Switzerland), 14, 6878.

Song, M.X., Yang, M.X., (2019). Leveraging core capabilities and environmental dynamism for food traceability and firm performance in a food supply chain: A moderated mediation model. Journal of Integrative Agriculture, 18, 1820-1837.

Soto-Acosta, P., Popa, S., Martinez-Conesa, I., (2018). Information technology, knowledge management and environmental dynamism as drivers of innovation ambidexterity: a study in SMEs. Journal of Knowledge Management, 22, 824-849.

Sunny, J., Undralla, N., Pillai, V.M., (2020). Supply chain transparency through blockchain-based traceability: An overview with demonstration. Computers & Industrial Engineering, 150, 106895.

Tsai, K.-H., Yang, S.-Y., (2013). Firm innovativeness and business performance: The joint moderating effects of market turbulence and competition. Industrial Marketing Management 42, 1279-1294.

Wowak, K.D., Craighead, C.W., Ketchen, D.J., Jr., (2016). Tracing bad products in supply chains: The roles of temporality, supply chain permeation, and product information ambiguity. Journal of Business Logistics, 37, 132-151.

Yam, R.C., Lo, W., Tang, E.P., Lau, A.K., (2011). Analysis of sources of innovation, technological innovation capabilities, and performance: An empirical study of Hong Kong manufacturing industries. Research policy, 40, 391-402.

Yi, Y., Bremer, P., Mather, D., Mirosa, M., (2022). Factors affecting the diffusion of traceability practices in an imported fresh produce supply chain in China. British Food Journal, 124, 1350-1364.

Zahra, S.A., George, G., (2002). Absorptive capacity: A review, reconceptualization, and extension. Academy of Management Review, 27, 185-203.

Zeng, J., Phan, C.A., Matsui, Y., (2015). The impact of hard and soft quality management on quality and innovation performance: An empirical study. International Journal of Production Economics, 162, 216-226.

Zhou, X., Pullman, M., Xu, Z., (2022). The impact of food supply chain traceability on sustainability performance. Operational Management Research, 15, 93-115.

Zhou, X., Zhu, Q., Xu, Z., (2023). The role of contractual and relational governance for the success of digital traceability: Evidence from Chinese food producers. International Journal of Production Economics, 255, 108659.

Embodied Communication via Virtual Reality Devices: The Visual-Body Language of Avatar in Metaverse

Shiyi Zhou (Durham university), Efpraxia Zamani (Durham University) and Zsofia Toth (Durham University)

Abstract

In Neal Stephenson's visionary novel Snow Crash, the "Metaverse" presents a virtual utopia where cyber avatars transcend physical limitations, engaging in vibrant interaction and creation. In 2003, Second Life emerged as a pioneering glimpse of this digital frontier, allowing users to connect globally through an immersive 3D world. Today, platforms like VRChat and Decentraland have evolved beyond screens, utilizing advances in head-mounted displays and body-tracking movements to enable rich, embodied interactions. These technological leaps invite exploration of how VRChat users communicate through a visual-body language that reshapes our understanding of mediated senses and social interactions in the virtual realm.

Keywords: Virtual reality; VRChat; Avatar; Social interactions; Touching

1 Introduction

Neal Stephenson, in his book *Snow Crash* (2003), describes a virtual utopia parallel to the real world where everyone has a cyber avatar and can interact, commune, create, and own property within the digitized sphere – called the Metaverse. A few years later, in 2006, Second Life (SL)[1] became an initial realization of such metaverse (Parkin, 2023), as a free 3D virtual world where users can create, connect, and chat with others from around the world using voice and text.

Today, platforms like VRChat and Decentraland have evolved beyond screens and keyboards and provide opportunities for immersive experiences and communications within virtual worlds, thanks to technological advances in head-mounted displays (HMD) and virtual reality (VR) devices.

Unlike the internet, the metaverse distinguishes itself by replicating real-world experiences through technologies, enabling direct interaction and heightened presence (Oleksy et al., 2023).

The immersive experience remains the fundamental feature of the metaverse, with various technologies such as VR, AR, MR, XR, and the latest spatial computing, depicting different levels of immersion. However, VR equipment serves as the essential hardware support for fully immersive experiences by creating and displaying stereoscopic images that mimic the natural

depth perception of human vision (Moreau, 2013), thus they enable highly interactive and immersive experiences.

[1] https://secondlife.com

The use of devices and media can bring about a reconfiguration so extensive so as to change the nature of a person (McLuhan, 1994). The digital body and stimulated vision in virtual worlds challenge us to redefine human existence beyond the physical realm. There's a growing body of literature and theory addressing computer-mediated human interaction, where imagination and sensation extend beyond the confines of the human body (Hayles, 1999; Whitehead, 2009). People may establish their sensory experiences and interact with others in novel and unexpected ways, influenced by VR devices.

This raises intriguing questions: How does the VR device mediate our sensations? How do users perceive their surroundings despite the sensory limitations in VR? When represented as virtual avatars, how do users translate sensory experiences into social interactions with others? To explore these questions, this article presents fieldwork conducted by the first author on the social-metaverse platform VRChat. Together with co-authors, we analyse how VR devices shape vision and senses in virtual social settings, emphasizing the importance of sensory experiences in metaverse studies.

2 Background

2.1 Sensory Experiences and Bodily Interaction via VR Devices

The history of such immersive devices roots in the 1930s: the flight simulator was used for pilot training which is now considered the early example of virtual reality. By the 1960s, virtual simulators were already utilized in the training astronauts, soldiers, and surgeons. However, contemporary VR devices are quite different from these earlier versions, Sutherland, in 1965, created the first HMD to incorporate computer technology to mediate a VR system, which was the first time that computers were used to display a real-world environment (Bown et al., 2017). VR devices generate a simulation of an avatar's body in the 3D space, allowing users to temporarily embody a different persona and undergo a psychologically real experience (Wiederhold, 2020). These VR sets typically include a pair of goggles and two hand controllers with location sensors, providing instant movement responses for the avatar's upper limbs and body (Bailenson, 2018). Advanced users may even utilize full-body trackers for more precise responses.

This level of immersion creates a sense of presence, allowing users to interact with their surroundings and each other, receiving real-time feedback through visual effects and haptic responses. These new forms of experiences directly relate to perception as an active and creative process that engages the entire body and its sensory modalities (Weiss & Haber, 2002).

According to Ingold (2000), the creative interweaving of experience in discourse and resulting discursive constructions affect people's perceptions of the world around them. And indeed, virtual worlds have been described as a form of techne, whereby virtual selfhood is predicated on the idea that people can intentionally craft their lifeworlds through creativity (Boellstorff, 2008). In doing so, the perceiver, i.e., in the case of virtual worlds, the user, continuously moves, learns and develops in the forcefields of their environment, with the (virtual) world here can be perceived as a complex meshwork that interweaves life, growth, and movement of all the beings and things present there (Ingold, 2011).

Within this complex meshwork, perception, viewed as the inner experience of the present moment, becomes intricately intertwined with intentionality and experiential direction within the virtual realm (Merleau-Ponty, 2004). One's inner experience is closely linked to body-machine interactions and engages their sensory organs, enabling a fusion of desires and bodily movements that results in an immersive sense of embodiment (Csordas, 2002). This further suggests that interactions between individuals in the virtual world involve the consciousness of presence, exchange of body language and movement, which fosters supra-linguistic connections. In addition, it also suggests that consciousness extends beyond the confines of the individual, and spills over into the environment through sensory participation (Bohemia et al., 2019). It is thus essential to shift focus from traditional sensory categories to embodied sensory knowing and learning, where we can explore the ongoing, dynamic processes through which individuals engage with, experience, remember, and imagine their everyday surroundings. By moving beyond a static analysis of sensory data to focus more on fluid and evolving nature of human-environment interactions (Pink, 2014), we can gain a deeper understanding of the sensory experiences and bodily interactions enabled by VR devices and their impact on user experiences in virtual worlds.

2.2 Examining VRChat as a Form of Virtual World

The virtual world is a self-evident social construction with three defining components: it exists as a place, is inhabited by people interacting with each other and the environment, and is enabled by online technologies (Boellstorff, 2008). VRChat is one of virtual world platforms and now it has around in total 9.4 million players[2], with the majority being Generation Z (Gen Z). Born around the 2000s, Gen Z was raised in the era of the internet and smartphones, with a particular passion for technology and games. Socializing and community-building are top priorities for Gen Z gamers, who have turned games such as Minecraft and Fortnite into hubs for staying connected and spending time with their friends (Lamba & Malik, 2022). Therefore,

[2] "VRC Active Player Count & Population." 2024. MMO Stats. November 2024. https://mmostats.com/game/vrchat.

VRChat can be considered as a platform for socializing online, much like the previous generation did with Second Life.

Indeed, compared to Second Life, VRChat as a metaverse social platform is primarily designed for VR headset users, and it adds a layer of interaction that allows users to connect not just verbally and textually, but also physically. Without speaking or typing, users can establish bodily interactions with others through their avatars. As such, being physically present in a virtual space can evoke a sense of presence and engagement that is distinct from other online platforms. These features make VRChat as a main research field for examining how VR devices can affect social interaction and mediate the senses.

A game journalist (King, 2022), has described VRChat as a wild experience that can easily swallow newcomers with its many memes, custom servers, and approachable vibe:

> *"As the name suggests, this is a chat platform, but one that has outgrown its namesake to become a living, breathing, ecosystem… major companies who are either aware of its success and seek to replicate it or instead wish to avoid VRC's obvious influence because it is viewed as eccentric or unappealing to the average consumer. Sure, it's weird, but that's the point. People are inherently weird, and embracing that uniqueness in online spaces through our own avatars and mannerisms is the way forward we should be embracing."*

Despite VRChat remaining a niche culture and a small community, it offers a unique field for rethinking interpersonal interactions at the intersection of the metaverse and immersive technology. As researchers immersed in this virtual world, we are able to observe and interpret the personal and contextual aspects of participants' sensory practices, and understand the cultural factors that shape their perceptions and interpretations (Pink, 2015). The interaction and communication in VRChat are constantly switching between sound, text, subtle body movements, and ambient objects, creating multi levels of sensory experiences. Therefore, examining VRChat as a form of a virtual world is valuable. It will uncover how digital presence evolves based on the idea that people can intentionally craft themselves through creativity and how the metaverse is predicated as a virtual extension of selfhood.

2.3 Phenomenological Perspectives on Perception and Intentionality

Merleau-Ponty (2012) introduced the term 'motor intentionality' to describe intentional activities that inherently involve our bodily understanding of space and spatial features, such as the act of grasping. Our gaze, touch or body movement arouses a certain motor intention which at the things from which they are, as it were suspended. This embodied experience provides us with a way of access to the world and the object, and an experience of our body with a meaning that enriches verbal orders (Welton, 1999). Phenomenology of perception posits that in skilled, unreflective bodily actions, the body presents itself as an orientation towards a particular task, with its spatiality not merely defined by position but by situational context (Merleau-Ponty, 2004).

Verbeek (2008) extends the concept of intentionality beyond the realm of the exclusively human to include technology. He suggests that intentionality needs to encompass human-technology amalgams, contributing to the understanding of posthuman or transhuman beings. He introduces the category of 'composite intentionality', which results from the fusion of technological and human intentionality (Verbeek, 2008). There are different technological intentionalities: some aim to generate a representation of the world, like the thermometers, spectrographs, or sonograms; some are geared towards constructing reality, for instance, radio telescopes construct reality by rendering non-visible radiation from stars into visible images (Verbeek, 2008).

According to Ihde (1990, p. 29), there are two modes of perceptions, one is microperception that mediates the sensory perception which focuses on bodily actual seeing and hearing; the other is macroperception, which focuses on cultural hermeneutic perception. The digital world which we retrieve, communicate, entertain, and work, mediates the hermeneutic relation in the cultural macroperception, where the world represents itself in the large, connected web. Interaction tools as such VR devices are how we access the digital world, which mediates the embodied experiences in microperception. The various interface tools and the large digital sphere are interweaved together and shape how we perceive the lifeworld.

However, from the user's perspective, there is no single internet or digital world; rather, there is a series of devices through which we interact and contextualize our actions. In other words, VR devices serve as the tools to craft a virtual world that is uniquely accessible through technology, transforming technological intentionality into human intentionality. "Technologies provide a framework for human actions, they have a certain influence on those actions" (Verbeek, 2001, p.10). Such visual-audio stimuli intentionally construct another reality for users, and the boundaries between the digital and physical become blurred. Within the context of virtual worlds, the intentionality of the technology becomes even more evident as it actively constructs reality akin to how artists use brushes to paint (Verbeek, 2008).

3 Autoethnography and Participant Observation

According to Pink (2015), awareness of sensory intersubjectivity is crucial in understanding multisensorial encounters between individuals. Through participation in social and material environments, our sensory practices and identities are lived out. Autoethnography is crucial for exploring embodied experiences that empowers the researcher to explore personal experiences and portray individuals in the process of understanding and navigating their lives. It focuses on the researcher's relationships with others, and employs reflexivity to explore intersections between self and society (Adams et al., 2015). Since this study focuses on the senses and social interactions, we adopt this method to better understand the above from an intersubjectivity perspective, whereby the first author reports on their experiences in VRChat, coupled with their observations, a typical methodological choice for research in virtual worlds (Boellstorff, 2008). This approach enables researchers to deeply immerse in native practices (Pearce, 2009), which is "the centrepiece of any truly ethnographic approach" (Boellstorff, 2008, p. 69).

Ethnographers must be aware that they are part of what they study and should demonstrate how they are shaped and affected by their fieldwork experiences (Adams et al., 2015). To ensure effective fieldwork, the first author documented interactions and personal feelings through diary entries and fieldnotes each time they logged into VRChat. In addition, VRChat has a feature that allows users to activate a virtual camera, following them or recording their first-person perspective inside the virtual world. This makes self-recorded audiovisual footage the most direct way to authentically represent reality and sensually connect with other users. It became particularly valuable as it not only documented the author's experiences in detail but also allowed other authors to relive them vividly when analysing the episodes and context.

This study reported here draws from the above two methods and it represents the first author's fieldwork in VRChat. The role of the co-authors in the write-up of this journey is to support exploring and iteratively interpreting this journey and its findings through the perspectives of others, and the vantage points of different experiences. In what follows, the first author provides their account, organised in vignettes and structured around the different phases of their journey, which is then further analysed and interpreted through a collective analytic process on the basis of conversations between all authors and their subjective experiences (Tekeste et al., 2024).

4 Immersive Experiences at the Intersection of Avatar Dynamics and Social Norms

4.1 Entering the field

I created a VRChat account named 'Eleva' and logged in through both my PC and my VR gear at the beginning. As my VR journey was part of my research journey, I stated in my profile that my purpose was to conduct research and I also encouraged users to talk with me, making my status visible to everyone (users can choose to be invisible to others). Firstly, I have to admit it was somewhat difficult to understand how the platform works. On the one hand, there are many different technological barriers. When I first logged into VRChat, despite having experience with other VR games, I still struggled to navigate the different functions. Fortunately, I met many helpful people who taught me about VRChat and visited a world designed for new users to learn its different functions. On the other hand, finding a proper way to be part of this virtual world is also tricky. Many new users usually feel bored after cruising around and experiencing some visually stunning effects or realistic nature scenes. It seemed like other users also had their own groups and things to do, while others were just wandering around like ghosts, exploring different worlds without engaging. I had the same feelings at the beginning, but VRChat showed its appeal soon after I met a kind and helpful friend. He explained to me the basic social rules in VRChat and introduced me to his friend gang. That was the moment I truly connected with other users. VRChat offers a virtual experience but with real people, which is what makes it so special.

The diverse array of avatar options also posed a challenge to a researcher in virtual world. VRChat encourages users to build up models and provides detailed instructions on the avatar and world creation, uploading through its own SDK[3] and 3D modelling tools. Since I learnt a bit about 3D modelling skills and wanted to experience the whole process of uploading model, I made and uploaded my own avatar using Blender[4] and Unity.[5] And it became my regular avatar in VRChat. Users can customize avatars' height, body gestures, and motions and send memes, through the left controller menu. Some self-uploaded avatars have more functions, such as changing multiple appearances and clothes, holding different objects, or some special visual effects. VRChat settings menu also provides various functionalities for utilizing avatars in social interactions. By selecting another's avatar, a blue bubble pops up containing the user's

[3] An SDK is a toolkit of programs needed to build on a specific platform. https://creators.vrchat.com/sdk/

[4] Blender is a free and open-source 3D computer graphics software tool set used for creating animated films, visual effects, art, 3D-printed models, motion graphics, interactive 3D applications, virtual reality, and, formerly, video games. https://en.wikipedia.org/wiki/Blender_(software)

[5] Unity is a cross-platform game engine developed by Unity Technologies, The engine can be used to create three-dimensional (3D) and two-dimensional (2D) games, as well as interactive simulations and other experiences. https://en.wikipedia.org/wiki/Unity_(game_engine)

profile. There are several options, including “clone”, “favourite” and “change to the same height as the other’s avatar”. Additionally, users can configure avatar security measures within personal settings to prevent cloning and safeguard against clipping through the avatar model. This ensures that if other avatars come too close, one’s avatar can be protected by becoming invisible to others.

Upon entering this virtual world, I was initially more attracted to immersive experiences, which felt like a visual culture shock. As I became more involved in interacting with others and my surroundings, I discovered some uniquely perceived social rules and boundaries, as well as psychological feelings, shaped by both individuals and devices. I began being curious about whether VRChat users establish sensory experiences differently, and particularly how they make sense of their surroundings through partially sensory-deprived VR devices, and how they develop and experience social interactions through their avatars. VR serves as a sensory interactive tool that shapes social reactions and cultural norms within the virtual world. To begin understanding these, I followed the path of sensory connection with others.

4.2 Immersive Vision

Computer graphics techniques have brought about a sweeping reconfiguration of relations between an observing subject and modes of representation. Visual experiences are now highly mobile and exchangeable, abstracted from any founding site or referent (Crary, 1988). In VRChat, there is not one, but multiple parallel worlds initially presented as a 2D image on the menu list. The users have the power to set, copy, and even place a world inside of another world as they wish. The world is not only surroundings for users but also visual elements that users can change, trade, and create. The meaning of the world is on the surface, represented by its ‘texture’. For instance, when a user holds a pizza or stands below a starry sky, they are aware that the pizza is only a triangular box covered with the texture of a pizza top, and the sky is a curved large sheet of painted white stars on the dark-blue texture. However, this does not affect whether they are talking about how pretty the milky way is or enjoying the pleasure of eating the tasteless pizza. In VRChat, seeing and looking are not merely passive acts of perception but active processes in which users acknowledge their presence, and shape each other’s understanding, presence, and interactions within the virtual world.

Ingold (2000b) presents a radical way of how we perceive the world, where our senses work together to create a participatory engagement with our environment. For example, he argues, active looking and watching make the vision from ordinary sight, where we detect patterns in the ambient light reflected off surfaces, to revelatory sight, where the world seems to open up to the perceiver in the movement of its birth. Vision not only provides us with a fundamental way to the virtual world but also allows us access to ourselves. Mirrors can be found almost everywhere. They come in different sizes and serve various purposes, but their primary function is to provide users with a full-body portrait of their avatars. One reason for this is that mirrors allow users to capture body movements and facial changes that are not visible through a first-person perspective, giving them an overview of their environment and interactions.

I discovered another reason for the presence of mirrors through my own experience. When I finally managed to upload my self-made avatar, I couldn't wait to see it in VRChat. As I stood in front of the mirror, I saw my blue and starry eyes looking back at me, and my mouth opened and closed as I spoke. My arms waved and touched my face. As I immersed myself in this virtual body, I looked forward to embodying it. Before, it was a still 3D model – an object on the screen, distant from me. However, after I immersed myself in the virtual environment and looked at the reflection of this vivid body in the mirror, something changed. By seeing myself in the avatar, I confirmed my own existence and the subjectification of the avatar. When I see my mirrored image in the virtual world, I am reminded that I exist here as a visible entity. This is precisely because I, too, am the object that is seen by my own eyes.

Our seeing returns upon us through the things we see and ultimately leads to our self-recognition as both seeing subjects and objects of seeing (Grasseni, 2007).

4.3 Saying Hello = 'Head Patting'

When the vision in VRChat goes beyond just seeing, i.e., when it becomes an intentional act of making sense of our virtual selves and the world around us, it merges into the user's intention. But what happens when we start interacting with others in VRChat? Our experience of being there isn't just about the self anymore; it is also driven by our desire to connect with others. Touching and body movement work together in the context of social interactions and reconfigure our sensory experience when building relationships and being intimate with others.

I first met Neku at the entrance of the world 'Chinese Bar' while they were chatting with a friend. Neku's avatar is a cute cat girl who referred to himself as a "male girl" in his profile. With a friendly wave, I signalled my approach, and Neku promptly returned a gesture of waving back, inviting me to join their little group. We were both using VR gear, and as we stood close to each other in VRChat, our interactions took on a bodily dimension. After I joined, Neku's hands came to playfully pat my head and touch my rabbit ears. And I reciprocated by gently rubbing his plush cat ears and head in return. Our spontaneous exchange of touching each *other left Neku's friend somewhat sidelined. Neku's friend was limited to the PC terminal, so he could not partake in this bodily communication and interaction. After our initial patting, the three of us engaged in a conversation about the non-binary gender of VRChat users. Throughout the conversation, me and Neku never stopped touching and playing with each other's ears and hands, further deepening our connection. Later, Neku expressed an interest in adding me as a friend, despite stating in their profile that they didn't add users below Green*[6]*. Their friend was surprised and asked them why they wanted this 'newbie' (I was 'new user' at the time) on their friend list. Neku replied "We are already friends". This kind of interaction allowed Neku and myself to develop a temporary but unique intimate moment.*

This scenario juxtaposes users from two different log-in ends. In comparison to computer users, VR users benefit from the mobility afforded by upper body and arm movements facilitated by VR controllers. This additional layer of body-movement interaction enriches greetings and social encounters. The act of touching heads, faces, and hands establishes a channel for sensory communication. It is not an isolated situation: there are many YouTube videos were 'head patting' is depicted as the prevalent method of saying hello among VR users in VRChat. Besides, I witnessed this gesture many times when strangers in VRChat signal interest and initiate conversations with others through such gestures.

Sensory experiences are not merely biological processes but also involve recognizing the social significance of sensory cues, understanding the cultural norms associated with sensory behaviours, and participating in social activities that are mediated by sensory engagement (Hsu, 2008). Indeed, touching emerges as the primary way of interaction for many VR users in social settings, possibly because through this visible touch, they are rewarded with tangible feelings of acceptance and the gradual building of intimacy. "Sensory experience is socially made and mediated" (Hsu, 2008, p. 433): the act of touch can create deep connections, fostering closeness, familiarity, trust, and friendship. These interactions intertwine emotion, meaning, and memory, creating a holistic sensory experience.

4.4 When Avatar Dynamics Meet Social Norms

In VRChat, most humanoid avatars possess properties such as hair, ears, and clothing, they can interact and move in real-time with physics-aware. This feature is known as 'avatar dynamics',

[6] VRChat has a colour coded 'Trust Rank', which is computed based on how much time a user has spent in VRChat, how much content they've contributed, the friends they've made, and many other factors. These ranks are as follows: Visitor (white), New User (Blue), User (Green), Known User (Orange), Trusted User (Purple).

which refers to the way an avatar’s body motion and reactions are designed to make users feel more physically present in the virtual world.

Since any movable part of the avatar can become an interactive component that is available to everyone, VRChat offers a feature that allows users to control whether their avatar reacts to other users’ touch. Through this, users can also limit body touching to only friends, which can help prevent harassment or unpleasant interactions. This feature is always displayed above the avatar’s head, thus allowing users to discern whether others accept avatar interactions. It is important to note that some dynamic interactions include common areas like ears, hair, and tail, as well as unusual areas like the chest or genitals.

The physics-aware avatars offer opportunities to experience tactile sensations with non-existent objects, such as furry ears and tails on human bodies or exaggerated features, like large anime-style breasts. Interestingly, there is a large proportion of male users in VRC who employ female avatars equipped with animated breasts (mostly you can figure out their gender through the user’s voice). Seeing and experiencing these reaction effects provides both interacting parties with a sense of social presence and emotional resonance (Bissonnette, 2019).

Touching others’ breasts is considered inappropriate in many contexts and can seem quite unconventional in the context of friendly interactions. When asked about such touching, however, several male users indicated that, while they found it weird, it didn’t necessarily make them uncomfortable and emphasized that they didn’t mind such actions if others were acting out of curiosity or friendly touching.

Myself, I had an awkward moment involving virtual genital touching. One day, I entered a public avatar world with two friends, intending to find new avatars for fun. I was interested in experiencing being in a male body and designed a new avatar. The three of us stood facing a wall mirror, examining this new avatar - a muscular Uncle Mario. Initially, I was indulging in the humorous ambience of being a male in a virtual world, trying to act like a muscular guy by showing off fitness moves and waving my arms as I walked around. Soon, I found a function in my left controller that allowed me to make clothes disappear, leaving me only in my underwear. We immediately noticed the genital area of this avatar to be particularly ridiculous - a strip wrapped in panties hanging down in front. My two friends (using female avatars with male voices) burst into laughter and quickly came over to touch and play with this virtual penis, which could move and even stretch to a comically long size when pulled.

At first, I found the interaction hilarious and never thought to stop them. However, when I looked in the mirror and saw my friends crouching down in front of me, staring and holding onto that virtual penis, their sexual-like actions and inappropriate touch became apparent, transitioning from a funny virtual penis to the awareness that it was me standing there. I suddenly realized the connection between this virtual penis and my real body and identity and how inappropriate it was to be almost naked in front of two males and allow them to touch the virtual genitalia in public space. I almost broke down and quickly stepped away from them. I then quickly changed back to my original cat-girl avatar and suggested we visit another world.

Within virtual environments, the materiality of touch is challenged, with notions of touch folded into actions and gestures. VR controllers enable actions or gestures that mimic tactile interactions without physical contact with objects (Price et al., 2021). Although users can't (always) physically feel these movements, the avatar body itself serves as an active vehicle, allowing users to manage the accessibility of their bodily reactions and assert the right of body control. In this context, bodily gestures and actions become a tool for interaction, autonomously responding to touch from any source.

However, this does not mean that there are no limitations on touching and tactile interactions. On the one hand, the body parts that are available to touch depend on the interactivity of the avatar's design, with the avatar body becoming an agency for users to feel the sensation of being touched. On the other hand, these interactions create a distinct social framework for bodily contact in the virtual world, where the boundaries of touchable areas become blurred and optional. When bodily interactivity translates into social interaction within a specific context, real-world social norms may assert themselves. In my story, my awareness of bodily presence became the breaking point that brought our actions into the context of general social ethics (Sharma et al., 2015). As I began to feel uncomfortable with others touching my avatar, morals, or individual ethics shifted and activated the boundaries between virtual interaction and tangible experiences, as governed by certain social norms.

5 The Avatar Body as a Visual Language

In the preceding section, we presented how the intersection of avatar dynamics and social norms in VRChat creates a unique social environment where users navigate the boundaries of personal space and interpersonal interactions. Building upon this foundation, we can now discuss how the avatar's body becomes a visual language, capable of expressing emotions and intention traits.

First, in VRChat, it is common to see people sitting closely to or reclining against each other, sometimes in ways that resemble hugging or adopting sexual-like positions. One of the features that afford this is the spatial audio effect, where the incorporation of whisper voice quality serves as a vocal expression of the intimacy conveyed through enveloping arms and close body contact (Haddington et al., 2023). In addition, their visible positioning establishes a binary social relation. The two users engaged in a conversation using, besides voice, hugging and touching gestures through their avatars, steering their dialogue towards intimacy. These embraces created a small, private space between them within a public setting, signifying the occurrence of a unique conversation space. Despite the presence of others around them, bystanders refrained from interrupting, respecting the intimate boundary constructed by the two avatars. Intracorporeal intertwining created a moment of closeness which was exclusive to the dyadic unit. Participants used gestures and body language to convey one form of engagement while simultaneously talking and gazing at each other to communicate. These relations were not entirely exclusive and private, however, despite existing within the dyad. Silent and non-silent observers could in principle had joined this dyadic unit. Yet, this did not happen, due to the social norms that exist in terms of avatars interactions, which function as rules of conduct in VRChat.

Second, the form and positioning (distance/proximity) of the avatar indicates one's intention regarding possible social interaction, and this applies to both multiparty participation and dyadic frameworks.

When I was transported to Johnson's current world, he was sitting at the bar of a pub, with other avatars chatting around him. He moved to my side and stood in front of me, and after a while, he moved towards the bathroom next to the bar counter. I followed him, realizing his intention to chat privately in an enclosed, soundproof space without background noises. My avatar was that of a short rabbit girl, while Johnson's was a tall, all-black human figure. At first, I tilted my head up to look at him while speaking, but he quickly adjusted his posture to sit on his knees so that we were at the same height and talked face to face. He mentioned that he had been drinking, and his speech was slightly slurred, making it challenging for me to understand him at times. We stood facing each other in silence for a while until he abruptly switched to a tiny black cat avatar and began moving backwards out of the bathroom. I moved after him and once again returned to where we had met, where the crowd had gathered. At this point, I understood that he had ended his pleasantries with me and continued into his drunken group, so I logged out of the world. In this small plot, Johnson expressed a desire to communicate with me by moving forward to me and staying at the same height, which implies a respectful and engaged conversation, then moving back to signal the end of the meeting.

The transformation in the avatar body is like talking in different volumes. It can take on various postures, such as standing, sitting, or lying down, and can be customized with movements, height adjustments, and special effects. When chatting, people will adopt a preferred posture and body language, whereby these will communicate and be expressions of their attitudes (positive or negative) (Mehrabian, 2008). Some users, therefore, will prefer to maintain a consistent posture or height and maintain their proximity to the avatar they are interested in. In other words, changing or adapting the way an avatar looks or behaves can change the way it is perceived, as well as communicate different intentionalities, moods and attitudes. These actions serve as a supplement for social interactions and provide additional context.

Finally, the avatar can serve as an expression woven into certain communication contexts. Many users include labels in their profiles that communicate to others certain aspects of their personality, such as being introverts or even having social anxiety disorder. An avatar body can avoid direct communication, and language expression, but can still allow the user to connect with others through bodily interactions.

Igg, a VRChat user, and I engaged in a unique form of communication: texting each other rather than talking to each other, by holding our pens in the air and writing up our thoughts. Igg wrote: "IDK WHAT'S GOING ON BUT IT'S COOL HERE." Then, pointing the pen towards me and another avatar, he wrote "FRIEND", expressing how he was perceiving us. He then jokingly questioned his spelling, writing: "I spelled that wrong", pointing at the word "FRIEND." We both quickly reassured him that it was indeed spelled correctly. In turn, he responded: "AM I STUPID?" I vigorously waved my hands to deny it and opened my arms for him. Without hesitation, he approached, and we shared a brief, spontaneous hug.

As highlighted by Goodwin (2020), hugs, rooted not solely in verbal discourse, constitute intracorporeal activities that unfold step by step. Here, language, movable avatars, and bodily senses work together to integrate the person into a social discourse. By engaging in this hug activity, Igg and I embodied an expression of camaraderie and understanding in our social relations.

6 Conclusion

In this paper, we adopted a phenomenological lens for the presentation and analysis of an auto-ethnography account produced by the first author. Our aim was to investigate the social interactions that take place within the virtual world, and specifically explore the role of VR devices within the context of sensory perception and the technology-mediated body senses, as positioned between the virtual and the real. Vision has been viewed as a means of objectifying the outside world and the 'Other,' a perspective that has been criticized in the West (Ingold, 2021, p. 247). However, in the virtual world, technological mediators may re-configure this subject-object relationship: VR devices can change the way we position and perceive ourselves and others, and the proximity between ours and others' avatars can create perceptions of being physically and mentally close or far apart. In other words, avatar bodies serve as a powerful language, enabling users to express emotions, intentions, and social cues through bodily movements and interactions, and their spatial positioning. This visual language complements and enhances verbal communication, creating a rich and immersive social experience. As users continue to use and appropriate immersive tools, the potential for avatar bodies in the virtual world opens up new possibilities for social interaction and understanding.

7 References

Adams, T. E., Jones, S. L. H., & Ellis, C. (2015). *Autoethnography*. Oxford University Press.

Bailenson, J. (2018). *Experience on Demand: What Virtual Reality Is, How It Works, and What It Can Do*. W. W. Norton & Company.

Bissonnette, S. (2019). *Affect and Embodied Meaning in Animation: Becoming-Animated*. Routledge. https://doi.org/10.4324/9781351054461

Boellstorff, T. (2008). *Coming of age in Second Life: An anthropologist explores the virtually human*. Princeton University Press.

Bohemia, E., Gemser, G., de Bont, C., Fain, N., & Almendra, A. (2019). Conference proceedings of the Academy for Design Innovation Management. Research Perspectives In the era of Transformations. *E Academy for Design Innovation Management*.

Bown, J., White, E., & Boopalan, A. (2017). Chapter 12 - Looking for the Ultimate Display: A Brief History of Virtual Reality. In J. Gackenbach & J. Bown (Eds.), *Boundaries of Self and Reality Online* (pp. 239–259). Academic Press. https://doi.org/10.1016/B978-0-12-804157-4.00012-8

Crary, J. (1988). Techniques of the Observer. *October*, *45*, 3. https://doi.org/10.2307/779041

Csordas, T. J. (2002). Embodiment as a Paradigm for Anthropology. In T. J. Csordas (Ed.), *Body/Meaning/Healing* (pp. 58–87). Palgrave Macmillan US. https://doi.org/10.1007/978-1-137-08286-2_3

Goodwin, M. H. (2020). *'The interactive construction of a hug sequence.' In Touch in social interaction, pp. 27-53. Routledge.* https://www.taylorfrancis.com/chapters/edit/10.4324/9781003026631-2/interactive-construction-hug-sequence-marjorie-harness-goodwin

Grasseni, C. (2007). *Skilled Visions: Between Apprenticeship and Standards*. Berghahn Books.

Haddington, P., Kohonen-Aho, L., Tuncer, S., & Spets, H. (2023). Openings of Interactions in Immersive Virtual Reality: Identifying and Recognising Prospective Participants. In P. Haddington, T. Eilittä, A. Kamunen, L. Kohonen-Aho, I. Rautiainen, & A. Vatanen (Eds.), *Complexity of Interaction: Studies in Multimodal Conversation Analysis* (pp. 423–456). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-30727-0_12

Hayles, N. K. (1999). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. University of Chicago Press.

Hsu, E. (2008). The Senses and the Social: An Introduction. *Ethnos*, *73*(4), 433–443. https://doi.org/10.1080/00141840802563907

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth*. Indiana University Press. https://philarchive.org/rec/IHDTAT-3

Ingold, T. (2000a). Stop, look and listen!: Vision, hearing and human movement. In *The Perception of the Environment*. Routledge.

Ingold, T. (2000b). *The Perception of the Environment: Essays on Livelihood, Dwelling and Skill*. Psychology Press.

Ingold, T. (2011). Worlds of sense and sensing the world: A response to Sarah Pink and David Howes. *Social Anthropology*, *19*(3), 313–317. https://doi.org/10.1111/j.1469-8676.2011.00163.x

Lamba, S. S., & Malik, R. (2022). Into the Metaverse: Marketing to Gen Z Consumers. In *Applying Metalytics to Measure Customer Experience in the Metaverse* (pp. 92–98). IGI Global. https://doi.org/10.4018/978-1-6684-6133-4.ch008

McLuhan, M. (1994). *Understanding media: The extensions of man*. MIT press.

Mehrabian, A. (2008). Communication Without Words. In *Communication Theory* (2nd ed.). Routledge.

Merleau-Ponty, M. (2004). *The world of perception*. Routledge.

Merleau-Ponty, M. (with Landes, D. A.). (2012). *Phenomenology of perception*. Routledge. https://doi.org/10.4324/9780203720714

Moreau, G. (2013). Visual Immersion Issues in Virtual Reality: A Survey. *2013 26th Conference on Graphics, Patterns and Images Tutorials*, 6–14. https://doi.org/10.1109/SIBGRAPI-T.2013.9

Oleksy, T., Wnuk, A., & Piskorska, M. (2023). Migration to the metaverse and its predictors: Attachment to virtual places and metaverse-related threat. *Computers in Human Behavior*, *141*, 107642. https://doi.org/10.1016/j.chb.2022.107642

Pearce, W. B. (2009). *Making Social Worlds: A Communication Perspective*. John Wiley & Sons.

Pink, S. (2014). Digital–visual–sensory-design anthropology: Ethnography, imagination and intervention. *Arts and Humanities in Higher Education*, *13*(4), 412–427. https://doi.org/10.1177/1474022214542353

Pink, S. (2015). *Doing Sensory Ethnography*. SAGE.

Price, S., Jewitt, C., & Yiannoutsou, N. (2021). Conceptualising touch in VR. *Virtual Reality*, *25*(3), 863–877. https://doi.org/10.1007/s10055-020-00494-y

Sharma, S., Lomash, H., & Bawa, S. (2015). Who Regulates Ethics in the Virtual World? *Science and Engineering Ethics*, *21*(1), 19–28. https://doi.org/10.1007/s11948-014-9516-1

Stephenson, N. (2003). *Snow Crash*. Random House Worlds.

Tekeste, M., Zakariah, A., Azer, E., & Salahuddin, S. (2024). Surviving academia: Narratives on identity work and intersectionality. *Gender, Work & Organization*. https://doi.org/10.1111/gwao.13209

Verbeek, P.-P. (2008). Cyborg intentionality: Rethinking the phenomenology of human–technology relations. *Phenomenology and the Cognitive Sciences*, *7*(3), 387–395. https://doi.org/10.1007/s11097-008-9099-x

Welton, D. (Ed.). (1999). *The Body: Classic and Contemporary Readings*. Wiley-Blackwell.

Whitehead, N. L. (2009). Post-Human Anthropology. *Identities*, *16*(1), 1–32. https://doi.org/10.1080/10702890802605596

Wiederhold, B. K. (2020). Embodiment Empowers Empathy in Virtual Reality. *Cyberpsychology, Behavior, and Social Networking*, *23*(11), 725–726. https://doi.org/10.1089/cyber.2020.29199.editorial

Deciphering And Evolution of AI Practices

Boineelo R Nthubu
Lancaster University, b.nthubu1@lancaster.ac.uk

Hassan Raza
Lancaster University, h.raza7@lancaster.ac.uk

Research In progress

Abstract

There is a rapid increase in AI-based systems using NLP-based conversational assistants and AI agents for task augmentation and automation. However, despite the advancements in AI technology, challenges persist regarding the different interpretations of Responsible, Trustworthy and Explainable AI, as well as understanding the evolution of their practices. Using a systematic literature review, we will explore the relationships between responsible, trustworthy, and explainable AI in the organisational context. Furthermore, we will use a practice-based approach to identify the best practices for AI implementations and investigate how these practices evolve when they are situated in different organisational contexts. We will use a qualitative multiple case study design with semi-structured interviews in two organisations: a recruitment company with an AI platform for customers and a telecommunications company using AI for customer services. This research will contribute to IS literature by identifying the different types of responsible, trustworthy, and explainable AI best practices and how they evolve.

Keywords: Responsible AI, Explainable AI, XAI, Trustworthy AI, Practices.

1.0 Introduction

Generative AI (GenAI) is a transformative technology, and we are seeing the positive impact across different types of organisations including businesses and government organisations. However, the increased use of GenAI comes with risks which have been formally captured in AI regulations. The European Parliament is one of several other legislative organisations that have drafted legislation that can be used to identify the different levels of strategic risk associated with AI (Kalodinis et al, 2024). The four types of risks are: Unacceptable, High, Limited and Minimal (Schuet. 2024). Examples of unacceptable risk include the use of AI for technology used for social scoring and facial recognition, that can have an adverse impact on society. High risk involves the use of AI in critical functions such as transport, safety and security. Most AI based business systems used by organisations fall into the category of limited and minimal risk.

The challenge with current AI legislation and extant IS literature is that it does not provide compliance guidelines. Therefore, in the absence of guard rails, the interpretation of AI risk with AI technology and the wider socio-technical context is open to interpretation (Michael et al, 2024). There are three properties commonly associated with AI: Responsible, Trustworthy and Explainable AI. These terms have evolved with the AI discourse and by identifying how these terms can be used in practice, we will have a better of understanding AI risks. In this research, we investigate the mapping of AI properties to best practices and study the evolution of AI practices. Following this aim, the research questions for the study are:

RQ1 What are the best socio-technical practices for Responsible, Trustworthy and Explainable AI?

RQ2 How do socio-technical AI practices evolve in organisations?

2.0 Literature Review

We conducted an initial literature review on Responsible, Trustworthy and Explainable AI practices and definitions using the Association of Information Systems elibrary, following vom Brocke et al. (2009). This database was chosen to explore how these terms and their associated practices are defined in the IS literature. Using the keywords ("Responsible AI" OR "Responsible Artificial Intelligence") AND ("Trustworthy AI" OR "Trustworthy Artificial Intelligence") AND ("Explainable AI" OR "Explainable Artificial Intelligence"), we identified 377 studies. After a title analysis, 119 studies remained. We then screened abstracts for mentions of "practices," excluding irrelevant studies, leaving 10. Full-text review further excluded studies not directly addressing practices, resulting in 7 studies. Next, we plan to conduct a backward and forward search.

2.1 Responsible, Explainable & Trustworthy AI Definitions

Our initial review highlights varying interpretations of RAI, TAI, and XAI in the literature (Table 1). For example, RAI is interpreted from a practice-oriented perspective by some scholars (Vassilakopoulou et al., 2022; Shollo & Vassilakopoulou, 2024; Ghatar et al., 2023), while others adopt a governance-oriented view (Wang et al., 2020). The practice-oriented perspective defines RAI as a practice, whereas the governance perspective frames it as a comprehensive governance framework emphasising ethical, transparent, and accountable, and trustworthy AI (Wang et al., 2020).

AI Property	Definition	Encompasses	References
Responsible	Viewed as a practice	Trustworthy AI, Ethical AI, Human centred AI, Human well-being, Fairness, Safety, Transparency, Explainability, Accountability.	Vassilakopoulou et al. 2022; Shollo, and Vassilakopoulou, 2024; Ghatar et al. 2023.
	Viewed as governance	Ethical, Transparent, Accountable AI, Trustworthy AI	Wang et al. 2020
Explainable	Viewed as multidisciplinary, layered, and dynamic		Vatn and Mikalef, 2024
	Governance	Trustworthy, Ethical AI	Figueras et al. 2022
Trustworthy	Viewed from an organisational perspective		Schaschek, and Engel, 2023
	Viewed from a user-centric perspective		Vassilakopoulou et al. 2022

Table 2. Responsible, Trustworthy and Explainable AI Definitions

The lack of consensus on what each property encompasses underscore the need to provide synergies of how these terms are linked. This synergy will allow for a further exploration of practices through a uniform understanding.

2.2 Responsible, Explainable & Trustworthy AI Practices

Table 2 presents Responsible and Trustworthy AI practices identified from the literature, demonstrating that these practices have been derived from various sources, namely; academic literature, industry reports, and expert insights, rather than being guided by regulatory standards. An exception is noted in the work of Figueras et al. (2022), who draw upon principles from the European Union’s 2019 Ethics Guidelines for Trustworthy AI. However, explainable AI (XAI) and these related properties are dynamic (Vatn and Mikalef, 2024), with their interpretations and applications subject to change due to technological advancements and evolving regulatory frameworks. Consequently, there is a need to consider practices informed by emerging regulations.

This approach would allow for an analysis that acknowledges the dynamic nature of these practices and captures their evolution over time in response to regulatory and technological developments.

AI Property	Practice	Authors	Gap
Responsible	Data governance, Ethically designed, Human-centric surveillance/risk control, Training and Education. Mapped from industry reports.	Wang et al. 2020	The authors call for further research to identify responsible AI practices using primary data at an organisational level.
	Mapped 10 practices on Explainability, Fairness, and Inclusiveness, from experts across different companies.	Ghatar et al. 2023	Calls further research to be conducted through case studies focusing on specific domain.
	Transparency, Fairness, Stakeholder consideration and involvement. Mapped from the EU "Ethics guidelines for Trustworthy AI" (AI-HLEG, 2019).	Figueras et al. 2022	Calls for future research to explore how ethical principles in AI can be diversely interpreted and enacted in practice.
	None	Vassilakopoulou et al. 2022	The authors call for more research to be conducted on situated and contextual aspects of AI technology use.
Trustworthy	Responsibility by design, Explainability, Fairness, Simplest possible solution, Human oversight, Testing, Governance, Ethical design, Risk control. These are based on literature.	Schaschek, and Engel, 2023	How these practices are enacted requires further empirical work.

Table 2. Table 2 RAI, TAI and XAI practices identified in the Literature

.

Theory

This research draws on practice theory as a theoretical lense. We use the definition of practices as. "structured spatial-temporal manifolds of action.." (Schatzki, 2006 pg.

1864). The spatial dimension practices enable us to understand how practices are situated in a context. The temporal dimension helps us understand the evolution of the practice. Practices are made up of structure and action and the structure of practices can be explained with the “why” and “how” (Schatzki, 2006).

The “why” addresses the reason for the action to take place and takes into consideration perceived benefit for the organisation. The “how” focuses on the specific actions that need to occur to achieve the goal and therein lies the challenge as responsible, trustworthy and explainable AI practices are not clearly defined. The technical practices are well-defined and follow a pre-defined process logic such as the training of a LLM or the interaction between the user and the transformer. Most organisation rely on bricolage or improvisation (Verjans, 2005) and sense making (Weick et al. 2009) to determine the socio-technical practices.

The overcome the challenge of various interpretations of RAI, TAI & XAI, we theorise that XAI is a governance function and RAI and TAI are practical components of XAI (Figure 1). Moreover, we argue that RAI is related to the practices associated with the secure design and implementation of AI. Furthermore, TAI is related to practices associated with training the AI models to ensure that the information provided is accurate. We also argue that RAI and TAI are interdependent because the ability to get an accurate output from the Gen AI application is reliant on a well developed AI Model.

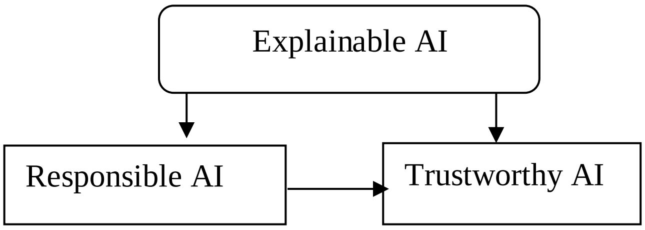


Figure 1. Interconnectedness Among XAI, RAI & TAI.

3.0 Methods

The research study uses the constructivist paradigm to understand how practices are constituted and implemented. Moreover, we rely on the interpretation of these practices in different settings. Using a qualitative multiple case study research design (Yin, 2009), we will investigate two organisations through semi structured interviews.

Using a critical lens, we will identify best practices based on the ISO 42001 AI management standard and map them to our conceptual definitions of responsible, trustworthy and explainable AI from the literature. We will then use these best practices as a basis for semi-structured interviews with organisational stakeholders responsible for designing and implementing AI practices in organisations. The first aim is to understand the deviation from best AI management practices in organisations. The second aim is to study the evolution of practice over a longitudinal period. Sample questions include: *1. What is your understanding of RAI, EAI and TAI? 2. Please explain how you implement "Practice A"?*

We will be using qualitative data analysis techniques to examine the data and find themes using the Gioia methodology (Goia et al., 2013).

4.0 Discussion

This research will demonstrate the effectiveness of using practice theory to study AI practices in organisations. The rapid advancement in AI technology continues to disrupt AI practices in organisations. Our approach traces the evolution of AI best practices to emergent practices in different empirical contexts.

4.1 Expected Contributions

We expect the findings of this study to contribute to IS literature on Responsible, Trustworthy and Explainable AI. First, it is anticipated that our work will demonstrate how ISO 42001 standard practices can be mapped to XAI, RAI & XAI. Second, we expect the study will show the dynamic nature of these practices as they evolve over time. Finally, this research study will demonstrate how the emergent nature of AI practices leads to changes in IS security risk in organisations.

5.0 Future Research Direction

This research in progress study is still in the early ages. We plan to continue with the literature review and primary data collection from the two identified case studies.

Acknowledgements

This study has been undertaken as part of UKAIS research grant.

References

Akbarighatar, P., Pappas, I., and Vassilakopoulou, P. (2023). Practices for responsible AI: Findings from interviews with experts (2023). *AMCIS 2023 Proceedings*. 4.

Figueras, C., Verhagen, H., and Cerratto Pargman, T. (2022). Exploring tensions in Responsible AI in practice. ," *Scandinavian Journal of Information Systems*: Vol. 34: Iss. 2, Article 6.

Gioia, D. A., Corley, K. G., and Hamilton, A. L. (2013). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational Research Methods*, 16(1), 15–31

Kalodanis, K., Rizomiliotis, P., and Anagnostopoulos, D. (2024). European Artificial Intelligence Act: an AI security approach. *Information & Computer Security*, *32*(3), 265-281.

Michael, K., Vogel, K. M., Pitt, J., and Zafeirakopoulos, M. (2024). Artificial Intelligence in Cybersecurity: A Socio-Technical Framing. *IEEE Transactions on Technology and Society*.

Schaschek, M., and Engel, S. (2023). Measuring Trustworthiness of AI Systems: A Holistic Maturity Model. *ICIS Proceedings*. 7.

Schatzki, T. R. (2006). On organisations as they happen. *Organization studies*, *27*(12), 1863-1873.

Schuett, J. (2024). Risk management in the artificial intelligence act. *European Journal of Risk Regulation*, *15*(2), 367-385.

Sen, R., Heim, G., and Zhu, Q. (2022). Artificial Intelligence and Machine Learning in Cybersecurity: Applications, Challenges, and Opportunities for MIS Academics. *Communications of the Association for Information Systems*, 51, pp-pp.

Shollo, A., and Vassilakopoulou, P. (2024). Beyond Risk Mitigation: Practitioner Insights on Responsible AI as Value Creation.*ECIS Proceedings*. 6.

Vassilakopoulou, P., Parmiggiani, E., Shollo, A., & Grisot, M. (2022). Responsible AI: Concepts, critical perspectives and an Information Systems research agenda. *Scandinavian Journal of Information Systems*: Vol. 34: Iss. 2, 3.

Vatn, D. M. K., and Mikalef, P. (2024). Theorizing XAI-a layered concept being multidisciplinary at its core. *AMCIS 2024 Proceedings*

Verjans, S. (2005). Bricolage as a way of life: Improvisation and irony in information systems. *European Journal of Information Systems*, 14(5), 504–506

vom Brocke, J., Simons, A., Niehaves, B., Reimer, K., Plattfaut, R., and Cleven, A. (2009). Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process. *ECIS Proceedings*, 2206–2217

Wang, Y., Xiong, M., and Olya, H. (2020). Toward an understanding of responsible artificial intelligence practices. In *HICCS* (pp. 4962-4971).

Weick, K. E., Sutcliffe, K. M., and Obstfeld, D. (2005). Organising and the process of sensemaking. *Organisation Science*, 16(4), 409–421.

Yin, R. K. (2009). *Case Study Research: Design and Methods* (4th ed.). Thousand Oaks, CA: Sage

Control Theory in the Age of Generative AI: The Case of Stack Overflow

Dewan Scholtz[*]**, Anastasia Griva and Kieran Conboy**

J.E. Cairnes School of Business and Economics, Lero—The Science Foundation Ireland Research Centre for Software, University of Galway, Galway, Ireland

Efpraxia Zamani

Durham University Business School, UK

* *Corresponding author: D.Scholtz1@universityofgalway.ie*

Developmental paper

Abstract

This paper aims to investigates the influence of Generative AI (GAI) on control modes, styles and purpose within open-knowledge networks, using Stack Overflow as a case study. Traditionally, control mechanisms on Stack Overflow are largely informal and rely on a merit-based reputation system that incentivizes high-quality, contributions from a community of over 23 million developers. However, the introduction of GAI disrupts these established control dynamics, challenging the integrity of knowledge validation processes and altering accountability structures. Analysis reveals that GAI's influence shifts the network from structured controls to a discourse-driven environment – which is not a static state. Insights on the role of GAI in these changes suggest that control structures, particularly driven by merit-based activities, are greatly under pressure as the human mimicking nature of this technology introduces various challenges.

Keywords: Control Theory, Generative Artificial Intelligence, Open-Knowledge Networks, Discourse-driven control

1. Introduction

Open-source has become increasingly popular, and with the rise of Generative AI (GAI), its impact on these environments is yet to be fully understood. Open-knowledge networks are collaborative platforms where users freely contribute, share, and refine information across various topics, fostering accessible and community-driven knowledge exchange (Sumanth & K, 2018) These networks introduce challenges and complexities in managing user contributions, quality control, and platform integrity in the age of GAI (Burtch et al., 2024). The human-mimicking nature of GAI brings new challenges to control settings. Merit and expertise are under threat as it is becoming more difficult to differentiate between experts and non-experts, especially in the realm of IS. Controls that rely on recognition of expertise and peer-regulation for maintaining platform integrity as battling to account for this phenomenon.

A well-known open-knowledge network at the forefront of this disruption is Stack Overflow, a collaborative question-and-answer platform with over 23 million contributors. Designed as an open-knowledge, coding platform, Stack Overflow is sustained by the contributions of developers and operates on a reputation-based system that incentivizes participation and attracts skilled users. However, with the advent of GAI, this platform is experiencing a profound shift, grappling with unforeseen challenges to its traditional meritocratic system of user-generated content and peer regulation (Jin et al., 2015; Vranić et al., 2023; Wang et al., 2021). With the introduction of GAI, several users have improved their scores and achieved high-reputation rankings. However, this does not necessarily reflect their expertise in their domain (Wang et al., 2021), raising questions about the relevance of merit-based incentives and the culture of community-driven moderation.

Despite the growing popularity of GAI, current control theories largely overlook its role in environments that rely on merit-imposed control mechanisms. Research has yet to fully explore how control modes (Choudhury & Sabherwal, 2003; Kirsch, 1996; Ouchi, 1979), styles (Adler & Borys, 1996; Heumann et al., 2015; Wiener et al., 2016) and purpose (Dekker, 2004; Gulati & Singh, 1998; Wiener et al., 2019) are or can be adapted when AI-generated contributions become indistinguishable from human input. It also remains unclear whether GAI should be treated merely as a tool or something with more control enabling characteristics.

This study aims to address these gaps by analysing Stack Overflow as a case study to reveal how GAI disrupts traditional control concepts in open-knowledge environments and to determine the emergent role of GAI in influencing and shaping platform control dynamics (Gleasure et al., 2019; Wiener et al., 2016). These control concepts will be used to examine how the current structures of control in Stack Overflow holds up against the disruption of GAI, by analysing how it changes as the events in the case progresses.". These events on Stack Overflow, instigated by GAI, paint a clear picture of its transformative nature. Through this exploration, the study seeks to provide insights looking into three major catalytic events in the case – the ban of GAI, the moderator strike and Stack Overflow's partnership with OpenAI. Through this process, the paper aims to address the de- or re-configuration of control purpose, modes and style, driven by the human-mimicking nature of GAI.

Research Questions (RQ):

RQ1: *How does control modes, styles and purpose manifest in open-knowledge networks in its indented control state?*

RQ2: *How has the introduction of GAI influenced control modes, styles and purpose in open-knowledge networks.*

This paper begins with a background section describing Stack Overflow's open-knowledge environment, current research on GAI in these spaces, and foundational control theory concepts. The methodology section outlines the case study approach, data collection, and analysis. In the findings, we explore the static control structure in Stack Overflow pre-GAI, the shifts in control modes, styles and purpose post-GAI introduction, and investigate GAI's role in control settings. The paper concludes with theoretical contributions, practical implications, and suggestions for future research.

2. Background

2.1. Stack Overflow as Open-Knowledge Network

Stack Overflow, part of the broader Stack Exchange network, is the largest open-knowledge platform for developers, at its peak hosting more than 24 million questions and 35 million answers. Among the 173 Stack Exchange communities, it stands out as a vast, community-driven database of programming knowledge, focusing on technical, rather than opinion-based, content. (Movshovitz-Attias et al., 2013; Sumanth & K, 2018).

To maintain content quality, Stack Overflow uses a reputation system—a merit-driven engagement model that rewards users for activities like asking relevant questions, providing accurate answers, and evaluating content through upvotes, downvotes, flagging, and commenting. High-reputation users gain influence and privileges, such as editing posts and enforcing guidelines. Some contributors even become moderators, ensuring adherence to the platform's rules and maintaining content standards (Sumanth & K, 2018). This reputation system enables employers to evaluate user profiles for expertise (Wang et al., 2021).

To encourage engagement, Stack Overflow employs various gamification and social engagement features. Studies show that these mechanisms significantly boost activity on the platform (Cavusoglu et al., 2015; Jin et al., 2015). However, this activity heavily relies on trust in the integrity and merit of the reputation system (Gallivan, 2001; Vranić et al., 2023).

By choosing Stack Overflow as a case study, the aim is to expand on the context of control in IS with more focus on the environment where stakeholders work independently on various different projects—motivated by the reputation system. Unlike the focus of current IS control literature (Gleasure et al., 2019; Remus & Wiener, 2012; Wiener et al., 2016), that refer to the IS project, rather than the IS platform, as further discussed in the control section.

2.2 Control in Open-Source Networks

Studies on control in IS typically focus on Information System Projects (ISP), where clear objectives, formal roles, and measurable tasks are present (Kirsch, 1996, 2004). *Control amount*, influenced by task complexity, project size, and formalization, explains the *control dynamics* (Kirsch & Cummings, 1996; Remus & Wiener, 2012). A foundational framework for digital control identifies two dimensions: *control configuration*, which refers to formal mechanisms like policies and rules, and *control enactment*, which is the actual governance of these mechanisms (Wiener et al., 2016). Expanding this framework, *control purpose* (Wiener et al., 2016, 2019) shifts the focus from merely enforcing behaviour to aligning decentralized actors towards shared goals, with *value appropriation* ensuring alignment with organizational objectives and *value creation* fostering collaboration (Wiener et al., 2016, 2019). Traditional control systems assume actors are trustworthy and pro-organizational (Davis et al., 1997; Gallivan, 2001; Vranić et al., 2023), aiming to mitigate agency concerns by emphasizing the 'big picture' (Heumann et al., 2015; Wiener et al., 2016).

Control mode (Choudhury & Sabherwal, 2003) distinguish *formal control* (input, behaviour, and output) from *informal control* (clan and self-control). This framework guides the paper's analysis of control mechanisms in Stack Overflow, exploring both formal and informal controls in IS development (O'dwyer et al., 2010). *Control style* can be *coercive* (enforcing strict rules) or *enabling* (supporting user initiative and growth), with the latter fostering collaboration (Adler & Borys, 1996; Heumann et al., 2015; Wiener et al., 2016). In crowdsourced environments, control shifts from formal mechanisms to informal, *discourse-driven* practices, where contributors shape outcomes through influence rather than direct intervention (Gleasure et al., 2019).

Technology-mediated control (TMC) complements collaboration in control settings, coordinating peer interactions without direct oversight (Cram & Wiener, 2020; Howison & Crowston, 2014). In these environments, control emerges through iterative knowledge-sharing processes, with control renegotiated as participants engage asynchronously, emphasizing collective contributions rather than top-down authority (Gallivan, 2001; Jarvenpaa & Majchrzak, 2011).

Concept	Sub concept	Definition	Reference
Control purpose (why)	Value appropriation	Controls for monitoring behaviour to protect resources and maintain standards by reducing misuse.	(Dekker, 2004; Gulati & Singh, 1998; Wiener et al., 2019)
	Value creation	Controls for promoting collaboration and contribution in order to enhance their application of knowledge and skills.	
Control modes (what)	Formal input, behaviour, and outcome control	Rules or guidelines directly governing resource allocation or processes. Awards or sanction users.	(Choudhury & Sabherwal, 2003; Kirsch, 1996; Ouchi, 1979)
	Informal clan and self-control	Norms or social rules encouraging adherence to standards or self-regulation	
Control style (how)	Coercive (or authoritative)	Controls that enforce compliance through strict rules.	(Adler & Borys, 1996; Heumann et al., 2015; Wiener et al., 2016)
	Enabling	Controls that facilitate compliance by enabling users to meet expectations and gain expertise.	
Technology-Mediated Control (TMC)		Using automated systems to enforce, monitor, and support control mechanisms.	(Cram & Wiener, 2020)

Table 1: Adapted from key concepts of control (Cram & Wiener, 2020)

2.3 Generative AI in Open-Knowledge Networks

The integration of GAI is disrupting traditional user participation and content creation patterns, with a decline in activity among newer users, suggesting a disconnect between GAI-driven efficiencies and community-building efforts (Brühl, 2023; Vranić et al., 2023). While GAI offers significant potential for improving decision-making and knowledge sharing (Prasad Agrawal, 2024), it complicates governance, as platforms must balance human and AI-generated contributions, considering technological, organizational, and social forces (Cram & Wiener, 2020; Khaw et al., 2023; Burtch et al., 2024). This dynamic requires a revaluation of control systems to maintain quality and consistency.

3. Research Method

This study adopts a case study methodology (Yin, 2003) to explore control in Stack Overflow, offering insights into a real-life, contemporary phenomenon. This approach enables an in-depth examination of new phenomena, especially when empirical substantiation is limited (Eisenhardt & Graebner, 2007), and is valuable for observing changes over time. The study progresses through three phases: *Phase 1* gathers data, selects cases, and reviews literature; *Phase 2* analyses content and trace data to apply control theories and frame Stack Overflow as a control case; *Phase 3* uses surveys and interviews to explore GAI's impact on control dynamics in open-knowledge networks (ongoing). A multi-method strategy is used for data triangulation purposes which include incorporating content analysis of Stack Overflow comments, trace data analysis to track shifts in user activity patterns, observations and analysis of news from Meta Stack Exchange and ongoing interviews and surveys with users.

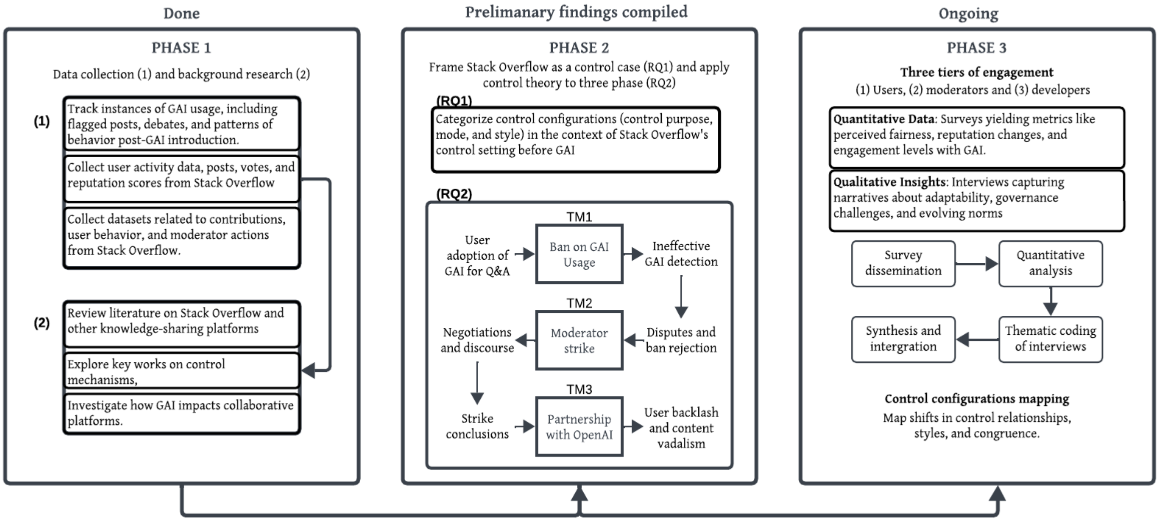

Done
Prelimanary findings compiled
Ongoing
PHASE 1
Data collection (1) and background research (2)
(1)
Track instances of GAI usage, including flagged posts, debates, and patterns of behavior post-GAI introduction.
Collect user activity data, posts, votes, and reputation scores from Stack Overflow
Collect datasets related to contributions, user behavior, and moderator actions from Stack Overflow.
(2)
Review literature on Stack Overflow and other knowledge-sharing platforms
Explore key works on control mechanisms,
Investigate how GAI impacts collaborative platforms.
PHASE 2
Frame Stack Overflow as a control case (RQ1) and apply control theory to three phase (RQ2)
(RQ1)
Categorize control configurations (control purpose, mode, and style) in the context of Stack Overflow's control setting before GAI
(RQ2)
TM1
User adoption of GAI for Q&A
Ban on GAI Usage
Ineffective GAI detection
TM2
Negotiations and discourse
Moderator strike
Disputes and ban rejection
TM3
Strike conclusions
Partnership with OpenAI
User backlash and content vadalism
PHASE 3
Three tiers of engagement
(1) Users, (2) moderators and (3) developers
Quantitative Data: Surveys yielding metrics like perceived fairness, reputation changes, and engagement levels with GAI.
Qualitative Insights: Interviews capturing narratives about adaptability, governance challenges, and evolving norms
Survey dissemination
Quantitative analysis
Thematic coding of interviews
Synthesis and intergration
Control configurations mapping
Map shifts in control relationships, styles, and congruence.

Figure 1. Research plan in three phases

The phased analysis, guided by process theory methodology (Burton-Jones et al., 2015), facilitates a structured examination of control adaptations across three pivotal events, referred to as ‘catalytic events” which mark significant shifts in Stack Overflow’s control structure. These events, treated as temporal markers, include: (TM1) the ban on GAI usage, (TM2) the moderator strike, and (TM3) the partnership with OpenAI. As illustrated in Figure 1 (Phase 2), these temporal markers delineate distinct moments in the platform’s evolution, enabling the systematic exploration of changes in control mechanisms as the platform navigates technological advancements and community challenges. This research is currently in Phase 2, while preparations are being made to start with Phase 3 (Figure 1).

4. Preliminary Findings

4.1 Framing Stack Overflow as a control case

By categorizing control purpose, modes, and style, we can understand how Stack Overflow’s reputation system restricts unwanted behaviour while encouraging self- and peer regulation to maintain quality, and adapt to community needs.

4.1.1 Control purpose, mode and style pre-GAI

Understanding how Stack Overflow through the lens of control theory has integrated systems of controls, in order to govern its community, is important in contrasting how these structures failed after the introduction of GAI. Controls are centralized around the reputation system – that mediates control purpose and modes, combining coercive moderation and enabling incentives to maintain platform integrity. These controls manifests and changes throughout the case. The table below shows how the controls look like in its static, intended state.

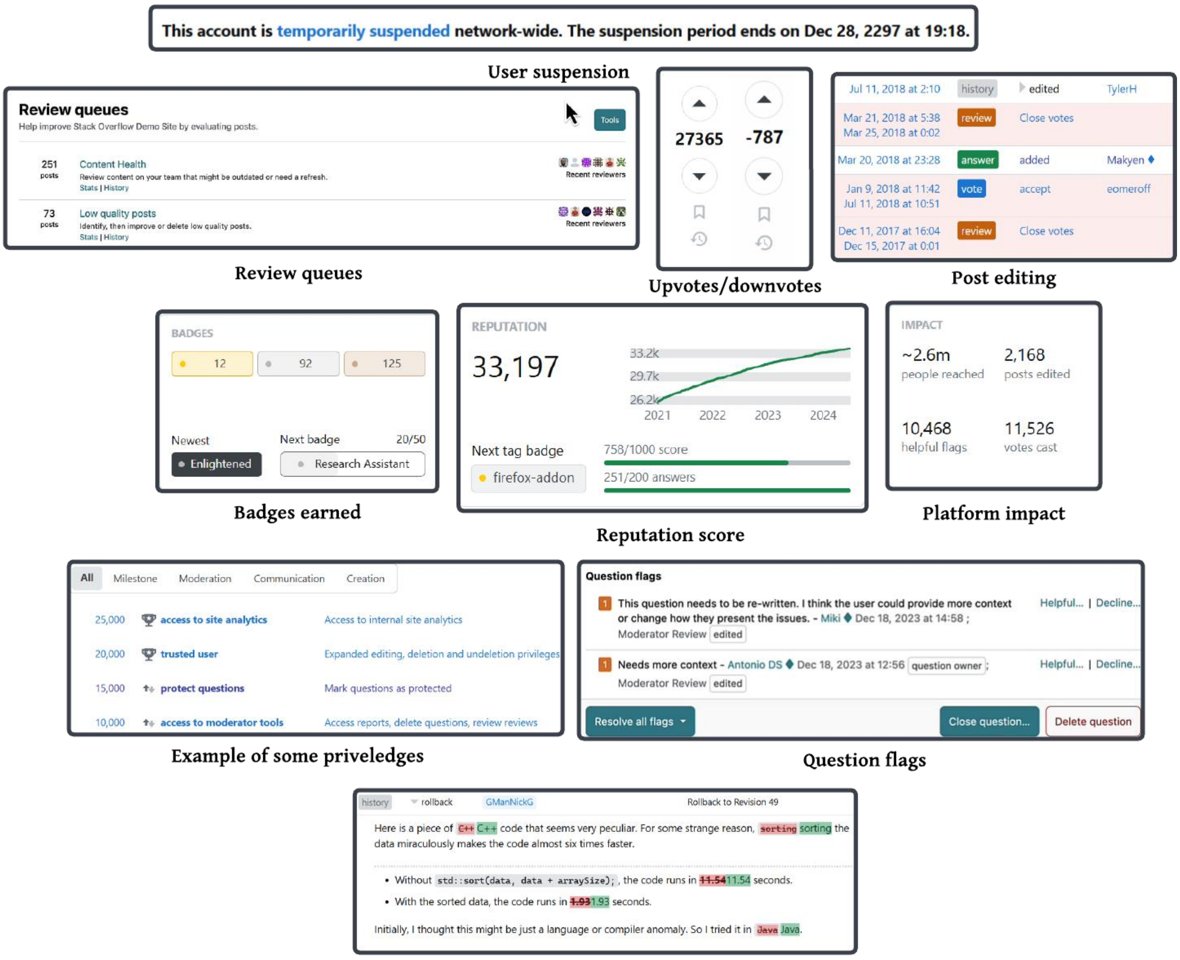
Concept	Sub concept	Application
Control purpose	Value appropriation	Reputation-based privileges ensure only experienced users can perform certain actions, protecting platform standards and resources (‘Example of some privileges’ in Figure 2).
	Value creation	Rewarding insightful answers enhances collective knowledge and individual users’ status, fostering high-quality contributions—without a pro-organization agenda (‘Upvotes/downvotes’ and ‘Platform impact’ in Figure 2).

Control modes	Formal	*Behaviour control:* Guidelines and moderation direct users to follow standards. (‘Review queues’ and ‘Question flags’ in Figure 2).
		Output control: Contributions are evaluated via upvotes, downvotes, and acceptance rates. (‘Upvotes/downvotes’ in Figure 2).
	Informal	*Self-control:* Users self-regulate to gain or maintain status and privileges (‘Example of content revision’ in Figure 2)
		Clan control: Community norms collaboratively uphold content quality
Control style	Coercive	Moderators enforce rules by deleting low-quality content and restricting accounts to maintain standards (‘User suspension’ and ‘Post editing’ in Figure 2)
	Enabling	The reputation system motivates users through rewards like badges, points, and privileges.

Table 2. Stack Overflow’s pre-GAI control environment

4.1.2 Control mediator: The reputation system

The reputation system mediates control by implicitly defining user behaviour expectations. Quality standards are reinforced through continuous feedback from upvotes, downvotes, and flags, aligning user actions with platform norms. Positive contributions earn rewards like points and privileges, while poor content results in penalties. Each contribution is collectively assessed, creating a self-sustaining ecosystem of governance and quality assurance (Figure 2).



This account is temporarily suspended network-wide. The suspension period ends on Dec 28, 2297 at 19:18.

User suspension

Review queues
Help improve Stack Overflow Demo Site by evaluating posts.

251 posts — Content Health
Review content on your team that might be outdated or need a refresh.
Stats | History

73 posts — Low quality posts
Identify, then improve or delete low quality posts.
Stats | History

Review queues

Upvotes/downvotes

Jul 11, 2018 at 2:10	history	edited	TylerH
Mar 21, 2018 at 5:38 Mar 25, 2018 at 0:02	review	Close votes	
Mar 20, 2018 at 23:28	answer	added	Makyen ♦
Jan 9, 2018 at 11:42 Jul 11, 2018 at 10:51	vote	accept	eomeroff
Dec 11, 2017 at 16:04 Dec 15, 2017 at 0:01	review	Close votes	

Post editing

Badges earned

Reputation score

Platform impact

25,000	access to site analytics	Access to internal site analytics
20,000	trusted user	Expanded editing, deletion and undeletion privileges
15,000	protect questions	Mark questions as protected
10,000	access to moderator tools	Access reports, delete questions, review reviews

Example of some priveledges

Question flags

1 This question needs to be re-written. I think the user could provide more context or change how they present the issues. - Miki ♦ Dec 18, 2023 at 14:58 ;
Moderator Review edited

1 Needs more context - Antonio DS ♦ Dec 18, 2023 at 12:56 question owner ;
Moderator Review edited

Question flags

history rollback GManNickG Rollback to Revision 49

Here is a piece of ~~C++~~ C++ code that seems very peculiar. For some strange reason, ~~`sorting`~~ sorting the data miraculously makes the code almost six times faster.

- Without `std::sort(data, data + arraySize);`, the code runs in ~~11.54~~11.54 seconds.
- With the sorted data, the code runs in ~~1.93~~1.93 seconds.

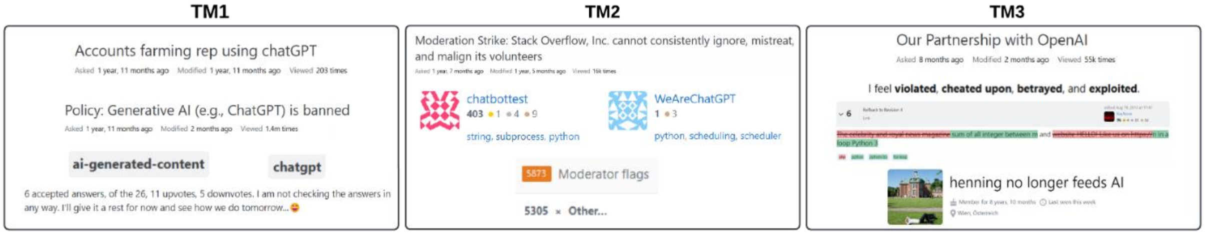
Initially, I thought this might be just a language or compiler anomaly. So I tried it in ~~`Java`~~ Java.

User self-revision

Figure 2: TMC/Reputation system

4.2 GAI's influence on control in Stack Overflow

The static state of Stack Overflow as seen in the previous section, was disrupted by the introduction of GAI. This section investigates how control changed in accordance to this disruption. Figure 3 below shows the phases that the study follows, highlighting the most important catalytic events, what triggered them and what the consequences was. By applying the lens of control theory, we are able to show the de-configuration of control modes, styles and purpose and the attempts to re-configuring it, showing the contrast between the intended state of control (in section 4.1) and how it looked after the GAI-disruption.



TM1

Accounts farming rep using chatGPT

Asked 1 year, 11 months ago Modified 1 year, 11 months ago Viewed 203 times

Policy: Generative AI (e.g., ChatGPT) is banned

Asked 1 year, 11 months ago Modified 2 months ago Viewed 1.4m times

ai-generated-content chatgpt

6 accepted answers, of the 26, 11 upvotes, 5 downvotes. I am not checking the answers in any way. I'll give it a rest for now and see how we do tomorrow...😍

TM2

Moderation Strike: Stack Overflow, Inc. cannot consistently ignore, mistreat, and malign its volunteers

Asked 1 year, 7 months ago Modified 1 year, 5 months ago Viewed 16k times

chatbottest 403 ● 1 ● 4 ● 9 string, subprocess, python

WeAreChatGPT 1 ● 3 python, scheduling, scheduler

5873 Moderator flags

5305 × Other...

TM3

Our Partnership with OpenAI

Asked 8 months ago Modified 2 months ago Viewed 55k times

I feel **violated**, **cheated upon**, **betrayed**, and **exploited**.

6

~~The celebrity and royal news magazine~~ sum of all integer between m and ~~website HELLO! Like us on https://~~n in a loop Python 3

henning no longer feeds AI

Member for 8 years, 10 months Last seen this week

Wien, Österreich

Figure 3: Ban on GAI (TM1), Moderator Strike (TM2) and Partnership with OpenAI (TM3) causes and effects

4.2.1 TM1: Ban on GAI usage

The first phase (TM1) marked the introduction of a GAI ban to counter "rep farming," where users exploited AI to gain reputation points. This overwhelmed moderators and exposed failures in self, behaviour, and output control. The platform shifted from enabling to coercive control to maintain content quality, but the ban sparked debate within the community. Divided into pro- and anti-GAI factions, users demonstrated a loss of clan cohesion as norms fragmented around tags like "AI-generated-content". During this phase we can see users starting to use GAI to appear as experts, through manipulating the systems of control that are supposed to ensure quality human contribution.

Concept	Description
Purpose	*Value Appropriation:* GAI-fuelled "rep farming" disrupted content quality, prompting a ban to regain control. *Value Creation:* AI-generated content provided little to no value, undermining knowledge quality.
Modes	*Behaviour:* Enforced GAI ban and user suspensions attempted to realign behaviour with quality standards. *Output:* Introduced new policies for GAI usage to maintain quality but faced enforcement challenges. *Self-Control*: Eroded as users exploited the system for reputation gains, weakening autonomous adherence to norms. *Clan:* Divisions between pro- and anti-GAI users created dissonance in community norms.
Styles	Shifts between *enabling* (e.g., flagging GAI content) and *coercive* (e.g., banning GAI and suspending users), with coercive control becoming dominant.

Table 3. How GAI changed control after the ban

4.2.2 TM2: The Moderator Strike

During the moderator strike (TM2), control weakened significantly as enforcement of the GAI ban shifted to the community. Reliance on AI detection exposed limitations in behaviour and output control, while staff intervention to reverse the ban led to widespread backlash. The strike disrupted value creation and appropriation, with content quality declining and norms fragmenting between pro- and anti-GAI factions. Conflict between moderators and staff members became more prevalent, with conflicting views on GAI – moderators wanting it banned, and staff members wanting it integrated. With the absence of coercive control enforced by moderators, the usage of GAI flooded the platform, with moderation backlogs stacking unmanageably high.

Concept	Description
Purpose	*Value Appropriation:* The strike halted moderation, enabling unchecked 'rep farming' and reducing content quality. *Value Creation:* Declined due to unregulated posts, fuelling concerns of platform decline.
Modes	*Behaviour*: GAI ban enforcement failed due to weak detection, leading to moderation strike and community frustration. *Output*: Reliance on the reputation system proved ineffective in curbing low-quality posts during the strike. *Self-Control*: Sentiment divides weakened user-driven adherence to standards. *Clan*: Community split into pro- and anti-GAI factions, disrupting shared norms and cohesion.
Styles	*Enabling:* Reverting the GAI ban attempted to restore enabling control but failed without adequate moderation. *Coercive:* Administrative intervention and ban retraction led to backlash and the strike.

Table 4. Moderator strike's effect on control in Stack Overflow

4.2.3 TM3: Partnership with OpenAI

TM3, Stack Overflow's partnership with OpenAI, marked the peak of control disruption. This partnership meant that OpenAI would have full access to the content on Stack Overflow to train their AI models, while Stack Overflow can deploy an AI assistant on the platform called OverflowAI. Users felt their contributions were appropriated for external gains without recognition or compensation, sparking backlash and eroding platform trust. Value creation declined as users defaced or deleted content, undermining the knowledge base. At this point, little to no form of the static state of control still existed, with the distrust in the platform at its peak. Controlling the crowd was almost impossible, and users left the platform, attempting to remove their contributions. As seen in Table 5 and Figure 3, almost no form of control existed in this phase.

Concept	Description
Purpose	*Value Appropriation:* User contributions repurposed for OpenAI's training of ChatGPT without recognition, sparking backlash. *Value Creation:* Defaced and deleted content undermined the platform's knowledge base.
Modes	*Behaviour*: Guidelines lost influence as users defied rules, and moderators struggled to enforce policies. *Output:* OverflowAI shifted focus from community knowledge to serving OpenAI, eroding content standards. *Self-Control:* Declined as users protested governance through non-compliance. *Clan:* Community norms fractured; users felt alienated by the OpenAI partnership.
Styles	*Enabling:* Reputation-based incentives lost effectiveness as platform goals diverged from community values. *Coercive:* OverflowAI launch disregarded user concerns, provoking defiance through content removal and account deletion.

Table 5. Control after OpenAI partnership announcement

5. Discussion and Conclusions

The goal of this study is to contribute to the understanding of how GAI disrupts controls in open-knowledge networks by mimicking human behaviour. It highlights GAI's challenge to the foundational human-centric mechanisms of platforms like Stack Overflow—peer governance, reputation, and trust—by enabling users to bypass these structures with AI-generated content. With these preliminary findings we are able to pinpoint critical issues of authorship forgery and misplaced credibility introduced by this technology. Control settings, within crowd-driven environments, need to rethink how effective their controls structures are. This applies across domains, as a new trend of non-experts using GAI to manipulate merit-based systems as seen in this case, are becoming more frequent. Research needs to further explore how to account for these issues. This research also emphasizes the need to expand control theory to account for human-like entities influencing human-driven systems and underscores the importance of adapting moderation practices and governance frameworks to maintain user trust and content standards in the face of AI-induced disruptions.

RQ3: What control relationship, style and congruence manifests through GAI usage?

Regarding our next steps apart from refining RQ1 and RQ2 with conducting interviews and surveys, our goal is to focus on RQ3. In more detail, our research will delve deeper into the dynamic of disruptive human-like technology, exploring how GAI functions within control relationships. Does it act as a stakeholder, a tool, or something entirely new? By examining its congruence with existing control systems and its role in shaping governance, our research aims to contribute to the understanding of control theory in the age of GAI.

Funding

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 13/RC/2094_2. Dewan Scholtz's PhD is funded by Hardiman PhD Scholarship, University of Galway.

References

Adler, P. S., & Borys, B. (1996). Two Types of Bureaucracy: Enabling and Coercive. *Administrative Science Quarterly*, *41*(1), 61.

Brühl, V. (2023). Generative Artificial Intelligence (GAI) – Foundations, Use Cases and Economic Potential. *SSRN Electronic Journal*.

Burtch, G., Lee, D., & Chen, Z. (2024). The consequences of generative AI for online knowledge communities. *Scientific Reports*, *14*(1), 10413.

Burton-Jones, A., McLean, E. R., & Monod, E. (2015). Theoretical perspectives in IS research: from variance and process to conceptual latitude and conceptual fit. *European Journal of Information Systems*, *24*(6), 664–679.

Cavusoglu, H., Li, Z., & Huang, K.-W. (2015). Can Gamification Motivate Voluntary Contributions? *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, 171–174.

Choudhury, V., & Sabherwal, R. (2003). Portfolios of Control in Outsourced Software Development Projects. *Information Systems Research*, *14*(3), 291–314.

Cram, W. A., & Wiener, M. (2020). Technology-mediated Control: Case Examples and Research Directions for the Future of Organizational Control. *Communications of the Association for Information Systems*, 70–91.

Davis, J. H., Schoorman, F. D., & Donaldson, L. (1997). Toward a Stewardship Theory of Management. *The Academy of Management Review*, *22*(1), 20.

Dekker, H. C. (2004). Control of inter-organizational relationships: evidence on appropriation concerns and coordination requirements. *Accounting, Organizations and Society*, *29*(1), 27–49.

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory Building From Cases: Opportunities And Challenges. *Academy of Management Journal*, *50*(1), 25–32.

Gallivan, M. J. (2001). Striking a balance between trust and control in a virtual organization: a content analysis of open source software case studies. *Information Systems Journal*, *11*(4), 277–304.

Gleasure, R., Conboy, K., & Morgan, L. (2019). Talking Up a Storm: How Backers Use Public Discourse to Exert Control in Crowdfunded Systems Development Projects. *Information Systems Research*, *30*(2), 447–465.

Gulati, R., & Singh, H. (1998). The Architecture of Cooperation: Managing Coordination Costs and Appropriation Concerns in Strategic Alliances. *Administrative Science Quarterly*, *43*(4), 781.

Heumann, J., Wiener, M., Remus, U., & Mähring, M. (2015). To Coerce or to Enable? Exercising Formal Control in a Large Information Systems Project. *Journal of Information Technology*, *30*(4), 337–351.

Howison, J., & Crowston, K. (2014). Collaboration Through Open Superposition: A Theory of the Open Source Way. *MIS Quarterly*, *38*(1), 29–50.

Jarvenpaa, S. L., & Majchrzak, A. (2011). Knowledge Collaboration in Online Communities. *Organization Science*, *5*(22), 12241239.

Jin, Y., Yang, X., Kula, R. G., Choi, E., Inoue, K., & Iida, H. (2015). Quick Trigger on Stack Overflow: A Study of Gamification-Influenced Member Tendencies. *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, 434–437.

Khaw, K. W., Alnoor, A., AL-Abrrow, H., Tiberius, V., Ganesan, Y., & Atshan, N. A. (2023). Reactions towards organizational change: a systematic literature review. *Current Psychology*, *42*(22), 19137–19160.

Kirsch, L. J. (1996). The Management of Complex Tasks in Organizations: Controlling the Systems Development Process. *Organization Science*, *7*(1), 1–21.

Kirsch, L. J. (2004). Deploying Common Systems Globally: The Dynamics of Control. *Information Systems Research*, *15*(4), 374–395.

Kirsch, L. J., & Cummings, L. L. (1996). Contextual influences on self-control of is professionals engaged in systems development. *Accounting, Management and Information Technologies*, *6*(3), 191–219.

Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., & Faloutsos, C. (2013). Analysis of the reputation system and user contributions on a question answering website. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 886–893.

O'dwyer, O., Conboy, K., & Scott, M. (2010). Control in E-Government Projects: An Exploratory Study. *Proceedings of the Sixteenth Americas Conference on Information Systems*, 1–10.

Ouchi, W. G. (1979). A Conceptual Framework for the Design of Organizational Control Mechanisms. *Management Science*, *25*(9), 833–848.

Prasad Agrawal, K. (2024). Towards Adoption of Generative AI in Organizational Settings. *Journal of Computer Information Systems*, *64*(5), 636–651.

Remus, U., & Wiener, M. (2012). The Amount of Control in Offshore Software Development Projects. *Journal of Global Information Management*, *20*(4), 1–26.

Sumanth, P., & K, R. (2018). Discovering Top Experts for Trending Domains on Stack Overflow. *Procedia Computer Science*, *143*, 333–340.

Vranić, A., Tomašević, A., Alorić, A., & Mitrović Dankulov, M. (2023). Sustainability of Stack Exchange Q&amp;A communities: the role of trust. *EPJ Data Science*, *12*(1), 4.

Wang, S., German, D. M., Chen, T.-H., Tian, Y., & Hassan, A. E. (2021). Is reputation on Stack Overflow always a good indicator for users' expertise? No! *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 614–618.

Wiener, M., Mähring, M., Remus, U., & Franke, W. A. (2016). Wiener et al./Control Configuration & Control Enactment in IS Projects. *MIS Quaterly*, *40*(3), 741–774.

Wiener, M., Mähring, M., Remus, U., Saunders, C., & Cram, W. A. (2019). Moving IS Project Control Research into the Digital Era: The "Why" of Control and the Concept of Control Purpose. *Information Systems Research*, *30*(4), 1387–1401.

Yin, R. K. (2003). Case Study Research: Design and Methods. *SAGE*, *5*(3).

Migraine Classification Using Machine Learning and Deep Learning in Low-Resource Healthcare Settings

Anithamol Ashokan
School of Computing and Engineering
University of West London
London W5 5RF, UK
anithamolashokan.9495@gmail.com

Dr. Ikram Ur Rehman
School of Computing and Engineering
University of West London
London W5 5RF, UK
ikram.rehman@uwl.ac.uk

Dr. Parisa Saadati
School of Computing and Engineering
University of West London
London W5 5RF, UK
parisa.saadati@uwl.ac.uk

Abstract

Migraine is a neurological condition that impairs quality of life, with diagnostic challenges, especially in resource-limited settings lacking specialised tools and expertise. While AI models for migraine classification have been explored in standard healthcare, limited research focuses on low-resource environments. To address this, we evaluate the efficacy of Machine Learning and Deep Learning models (SVM, KNN, DT, RF, and TabNet) for migraine classification, with a focus on computational efficiency and interpretability. Among the models, RF emerged as the best model, achieving 95.8% accuracy, precision, recall, and F1 score, while TabNet achieved slightly lower performance 91.1%, 91.8%, 91.1%, 90.7% respectively. RF demonstrated enhanced computational efficiency, with a training time of 0.9s and memory usage of 0.14 MB, compared to TabNet's 10.8s and higher memory usage. Furthermore, SHAP analysis supported RF's interpretability, and we propose RF as a cost-effective, AI-driven diagnostic tool for migraine classification, improving access to healthcare in resource-limited regions.

Keywords: Machine Learning, Deep Learning, Migraine, Computational Efficiency, SHAP, SMOTE, Random Forest, TabNet.

1.0 Introduction

Migraine, a complex and intense neurological condition, affect approximately 1 billion individuals worldwide (Pradeep et al., 2020), ranking as the second leading cause of disability according to the World Health Organisation. Though non-life-threatening, migraines severely impact work productivity, physical health, and emotional well-being (Khan et al., 2024). Triggered by factors like sensitivity to light, irregular sleep, and skipped meals, migraines can be challenging to diagnose accurately due to their symptomatic overlap with other headache types (Migraine Trust, n.d.). According to the study conducted by Ge and Chang (2023), migraine poses significant and persistent concerns especially in non-high-income East and Southeast Asia. Traditional diagnostic tools such as Magnetic Resonance Imaging, Computed Tomography, and Positron Emission Tomography scans are often employed to diagnose migraines, but these methods are costly and typically require access to highly skilled neurologists (Khan et al., 2024). This presents a significant barrier in low-resource settings where access to imaging facilities and neurology professionals is limited (Mortel *et al.*, 2022).

Moreover, in developing nations, where personal health insurance is often lacking, especially for the poorest populations, access to costly neuroimaging testing might be difficult. For example, almost 40% of people in Cameroon live under the poverty line and 96–98% lack financial support for medical bills (Mortel *et al.*, 2022). Consequently, expenses out of pocket for CT imaging can adversely affect patients and their families financially. These circumstances emphasise the crucial need for early disease intervention to decrease the impact of chronic illnesses on patient’s lives and a country's socioeconomic conditions, especially in areas with limited access to neurologists.

The rapid advancement of Artificial Intelligence (AI), particularly in Machine Learning (ML) and Deep Learning (DL), offers promising solutions for healthcare diagnostics by extracting patterns from complex clinical data (Rathore and Mannepalli, 2021). ML/DL models have been effectively used in disease classification, including patient clustering and diagnostic support (Torrente et al., 2024), which could aid in diagnosing migraine subtypes more affordably and efficiently. However, ML/DL model adoption in clinical settings has been limited by challenges in interpretability and the high computational demands of some models, which may be impractical in low-resource environments (Habehh and Gohel, 2021). Thus, developing computationally efficient and interpretable models is crucial (Rundel et al., 2024) for enabling broader use of AI-driven diagnostic tools in underserved regions.

To the best of our knowledge, this study is the first to propose a computationally efficient and interpretable model for migraine classification specifically trained for low-resource healthcare settings, while assessing key performance metrics accuracy, precision, recall, F1 score, Area under the Receiver Operating Characteristic Curve (AUC-ROC) and Mathew's correlation coefficient (MCC). In this research we compared the performance of ML models—Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Decision Tree (DT), and Random Forest (RF) as well as the DL model TabNet. The models were evaluated on a secondary migraine dataset, with performance metrics assessed both prior to and following Hyper-Parameter Tuning (HPT) to identify the most effective approach for the classification of migraine subtype.

In addition, computational efficiency was assessed based on training and prediction time, memory usage during these process, and overall size of the model. Interpretability, a key requirement for clinical adoption was addressed by analysing the effective model output through SHapley Additive exPlanations (SHAP) summary and waterfall plots, which highlight each feature's contribution to predictions, helping healthcare professionals understand and trust the model decision. By focusing on performance, computational efficiency, and interpretability, we aim to establish an AI-driven diagnostic approach for migraine classification that is both practical and clinically acceptable, thereby reducing reliance on expensive diagnostics and facilitating early intervention in underserved areas.

This paper is structured into five sections, beginning with the introduction following the abstract. The second section presents a literature review, examining existing studies relevant to our research and identifying their limitations. The third section outlines the methodology applied in this research, while the fourth section discusses the results and analysis. The fifth and final section offers the conclusion.

2.0 Literature Review

AI poses several noteworthy applications in headache diagnosis. A Computer-based Diagnostic Engine is one of the tools designed to diagnose migraine. The engine employs a rule set that is derived from the International Classification of Headache Disorder-3 criteria for primary headaches, in addition to evaluating secondary headaches and medication overuse headaches (Cowan *et al.,* 2022).

Chiang et al. (2024) proposed a novel natural language processing technique which was recently posted on GitHub to reliably assess headache frequency from free-text clinical notes. This technology attempts to help individuals determine the proper degree of treatment depending on their headache frequency, demonstrating another unique application of AI in headache management.

Moreover, AI offers innovative methods for diagnosing and categorising migraines using a range of ML and DL algorithms. Many studies have been proposed ML/DL models that can be used for migraine diagnosis and classification.

SVM is one of the commonly utilised algorithms in the studies, especially for classification problems. For instance, the study conducted by Hsiao et al. (2023) utilised SVM with different types of kernels and discovered that a median Gaussian kernel yielded the highest level of accuracy when distinguishing chronic migraines from other types of headaches with accuracy 92.6%. SVM's adaptability, together with its efficacy in managing high-dimensional data, has made it a widely favoured option (Kazemi and Katibeh, 2018).

Another efficient algorithm identified is the RF model and its variations, such as Extremely Randomised Trees, were extensively employed, especially in ensemble techniques aimed at enhancing classification accuracy. Sasaki et al. (2023) utilised these algorithms alongside additional methods such as XGBoost, showcasing their effectiveness in accurately identifying migraine headaches with accuracy noted 94.50%. ORHANBULUCU and LATİFOĞLU (2024) employed the Rotation Forest ensemble method to increase the resilience of the model by creating a variety of decision trees, resulting in an improvement in the overall accuracy of diagnosis by 95.14%. Utilising the EEG signal data Subasi, Ahmed and Alickovic (2018) showed that RF achieved the performance of 85.18% using flash simulation.

Alternatively, DL models have had a substantial impact on studies, especially in dealing with intricate data like MRI and MEG scans. The study conducted by Khan et al. (2024) emphasised the effectiveness of Deep Neural Network (DNN) in accurately categorising seven types of migraines. The DNN attained an exceptional accuracy rate of 99.66% by utilising augmented tabular data. Moreover, the researchers Rahman Siddiquee et al. (2023) successfully employed the advanced deep learning model 3D ResNet-18 to accurately diagnose migraines and post-traumatic headaches. The model demonstrated its capability to analyse high-dimensional imaging data, achieving accuracies between 75% in differentiating migraineurs from healthy controls.

Another noteworthy DL model is TabNet employed in research conducted by (S. N. Mudassir and R. M, 2024). Due to its capacity to capture intricate patterns in tabular data while preserving interpretability, authors were able to effectively capture complex linkages within the data, leading to enhanced diagnostic outcomes noted as 98%.

Additionally, Artificial Neural Network (ANN) have been utilised in several studies to categorise different forms of headaches. Despite being less sophisticated than newer deep learning architectures, these models have demonstrated their effectiveness in research including structured or less complex data. A study showed the performance of ANN as 98% accuracy in classifying migraine subtypes (Sanchez-Sanchez, García-González and Ascar, 2020). The adaptability of ANN to different headache categorisation problems has made it a highly significant tool in this field of research (Taufique et al., 2021).

Despite being a relatively simple algorithm, KNN’s effectiveness in classifying migraine achieved 85% demonstrated its continued relevance in ML research (Romould et al., 2024).

In the research conducted by Mitrović et al. (2023), Logistic Regression and Linear Discriminant Analysis (LDA) models were used for classifying migraine categories and healthy controls, with LDA in particular achieved high accuracy of 98% when paired with feature selection methods.

Moreover, DT models were also examined, frequently as components of ensemble methods or in conjunction with other algorithms to enhance performance, to improve the accuracy and robustness of classification models which is analysed in a study utilising EEG data of patients, DT achieved performance level of 87.5% (Hsiao et al., 2023).

2.1 Limitation

While many studies report high predictive accuracy, particularly for DL models such as DNNs and 3D ResNet-18, few studies focus on the computational efficiency required for these models to run in resource-constrained environments. For instance, sophisticated DL models, although highly accurate, often demand substantial computational power and expertise, limiting their feasibility for use in low-resource clinical settings. For example, Study conducted by Rahman Siddiquee et al. (2023) highlighted the effectiveness of these models in handling data like MRI or MEG scans, but the practical applicability of these models in developing countries, where computational infrastructure is limited, is largely unexplored.

Furthermore, model interpretability is another critical factor often overlooked. Many of the algorithms, especially DL models, operate as "black box" systems, making it difficult for healthcare professionals to trust or interpret the outputs without a clear understanding of the underlying processes (Miotto *et al.*, 2018). While models like TabNet offer some degree of interpretability, the balance between performance, interpretability, and computational demands has not been comprehensively evaluated. Existing research primarily focuses on limited performance evaluation metrics rather than comparing a range of metrics also with computational efficiency and interpretability analysis (Rundel et al., 2024).

We aim to address these gaps by conducting a comprehensive comparison of ML/DL algorithms, focusing not only on predictive accuracy but also on their computational efficiency and interpretability, which are critical for practical application in resource-constrained environments. By working with tabular data instead of imaging data, we provide an alternative approach that minimises costs while ensuring that the algorithms are accessible and usable in low-resource settings. In addition, we will contribute to the literature by evaluating models using a wider range of metrics in a single study, providing valuable insights into which models are most suitable for classifying migraine subtypes, particularly in underdeveloped regions where computational and financial resources are limited.

3.0 Methodology

This section describes the methods used to identify the effective algorithm in migraine classification. Firstly, the secondary dataset was obtained followed by Data preprocessing, Exploratory Data Analysis (EDA), Feature selection, and Data augmentation technique. Secondly, the model selection process for migraine classification was conducted, followed by model training and testing both prior to and post HPT. Various evaluation metrics, including accuracy, precision, recall, F1 score, AUC-ROC, MCC, and K-fold cross-validation scores, were applied to assess performance. Additionally, computational resource requirements were analysed, and the interpretability of the effective algorithm was thoroughly examined to ensure practical applicability in a clinical setting. The detailed workflow is shown in the Figure 1 and 2.

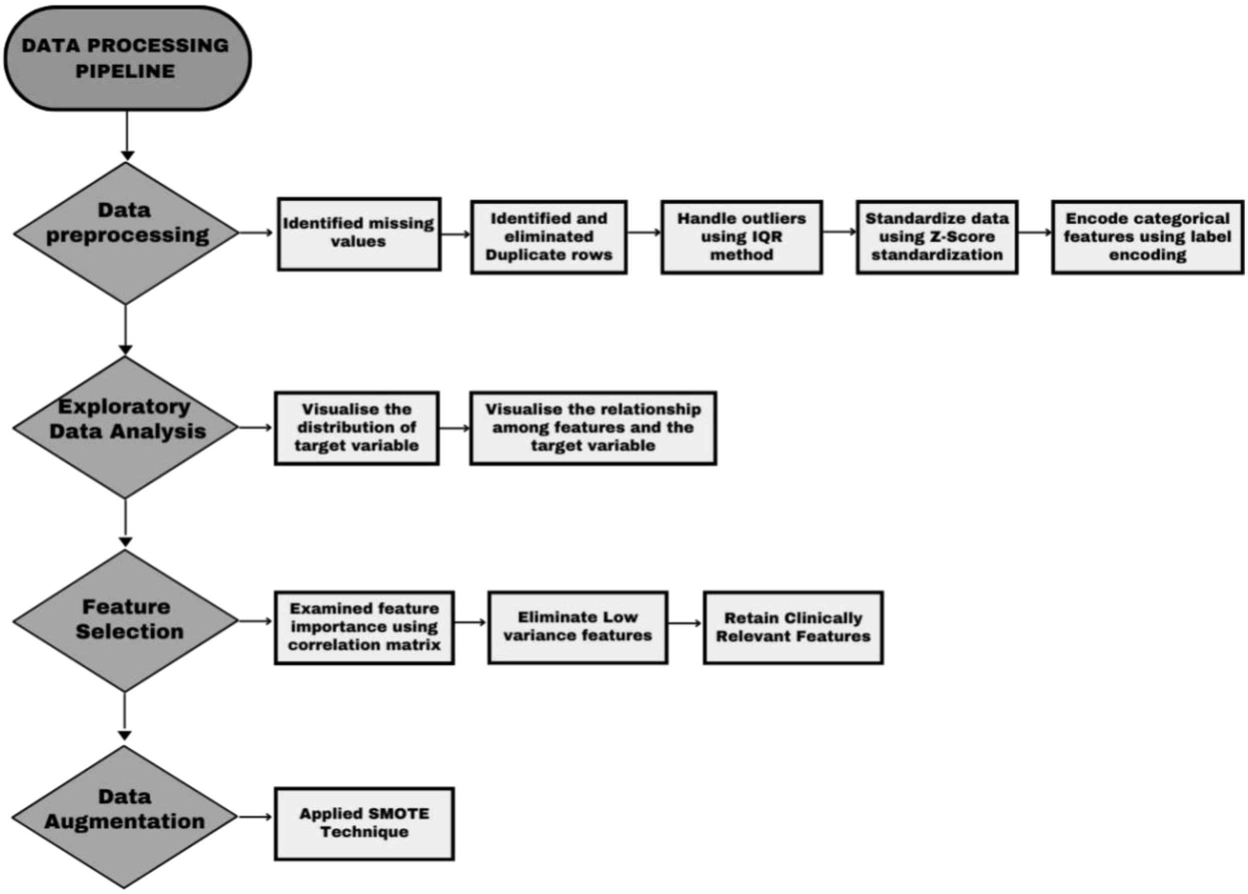

DATA PROCESSING PIPELINE
Data preprocessing
Identified missing values
Identified and eliminated Duplicate rows
Handle outliers using IQR method
Standardize data using Z-Score standardization
Encode categorical features using label encoding
Exploratory Data Analysis
Visualise the distribution of target variable
Visualise the relationship among features and the target variable
Feature Selection
Examined feature importance using correlation matrix
Eliminate Low variance features
Retain Clinically Relevant Features
Data Augmentation
Applied SMOTE Technique

Figure 1. **Data Processing Pipeline.**

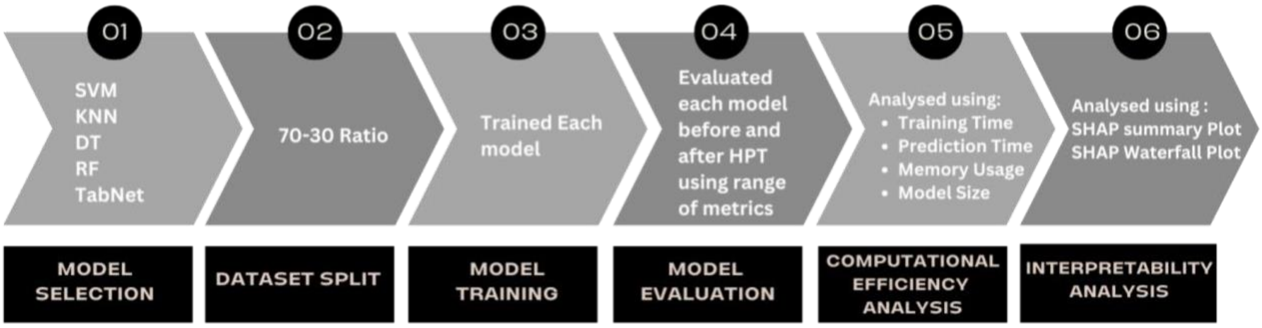


Figure 2. **Model Development and Evaluation Workflow.**

3.1 Dataset Collection

The dataset was acquired via Kaggle, an online community platform tailored for machine learning enthusiasts (www.kaggle.com, n.d.).

Moreover, the research conducted by (S. N. Mudassir and R. M, 2024) discovered that the dataset consists of medical records documenting patients diagnosed with various migraine-related disorders and it was carefully gathered by qualified medical experts at the Centro Materno Infantil de Soledad in the first quarter of 2013.

Furthermore, during the literature investigation, it was discovered that this migraine dataset has been extensively utilised in several studies on migraine analysis, including those conducted by (N. N. Aung and W. Srimaharaj, 2023), (Romould et al., 2024), and (Sanchez-Sanchez, García-González, and Ascar, 2020).

3.2 Dataset Description

The obtained data consists of 400 records with 24 features including both numerical and categorical variables. The attribute of the dataset is described in the Table 1. The "Type" column holds the different types of migraine diagnosed and it is the target variable in the study that need to be classified by the selected model.

SL. NO	ATTRIBUTES	DESCRIPTION	TYPE OF DATA	RANGE OF VALUE
1	Age	Age Of the Patient Reported	Continuous	15 To 77
2	Duration	Length Of Symptoms During the Most Recent Episode, Measured in Days.	Discrete	1 To 3
3	Frequency	Monthly episode frequency.	Discrete	1 To 8
4	Location	Either unilateral or bilateral pain location (none - 0, unilateral - 1, bilateral - 2)	Discrete	0 To 2
5	Character	Character depicts the throbbing or persistent sensation of pain. (none - 0, throbbing - 1, Persistant - 2)	Discrete	0 To 2
6	Intensity	The level of pain, categorised as mild, moderate, or severe (none - 0, mild - 1, moderate - 2, severe - 3)	Discrete	0 To 3
7	Nausea	Patient's sensation of nausea (not - 0, yes - 1)	Discrete	0 To 1
8	Vomit	Patient's sensation of vomiting (not - 0, yes - 1)	Discrete	0 To 1
9	Phonophobia	Sensitivity to noise (not - 0, yes - 1)	Discrete	0 To 1
10	Photophobia	Sensitivity to light (not - 0, yes - 1)	Discrete	0 To 1
11	Visual	The count of reversible visual symptoms	Discrete	0 To 4
12	Sensory	The count of reversible sensory symptoms	Discrete	0 To 2
13	Dysphasia	Impaired speech coordination (not - 0, yes - 1)	Discrete	0 To 1
14	Dysarthria	Disarticulated noises and words (not - 0, yes - 1)	Discrete	0 To 1
15	Vertigo	Symptom of dizziness (not - 0, yes - 1)	Discrete	0 To 1
16	Tinnitus	Patient's ability to hear things (not - 0, yes - 1)	Discrete	0 To 1
17	Hypoacusis	Deafness (not - 0, yes - 1)	Discrete	0 To 1
18	Diplopia	Dual vision (not - 0, yes- 1)	Discrete	0 To 1
19	Defect	Simultaneous frontal eye field and nasal field defect in both eyes (not - 0, yes - 1)	Discrete	0 To 1
20	Ataxia	Lack of muscular control (not - 0, yes - 1)	Discrete	0 To 1
21	Conscience	Compromised moral awareness (not - 0, Yes - 1)	Discrete	0 To 1
22	Paresthesia	Bilateral paraesthesia at the same time (not - 0, yes - 1)	Discrete	0 To 1
23	DPF	Family history (not - 0, yes - 1)	Discrete	0 To 1
24	Type	Type Of Migraine Diagnosed (Typical Aura With Migraine, Migraine Without Aura, Typical Aura Without Migraine, Familial Hemiplegic Migraine, Sporadic Hemiplegic Migraine, Basilar-Type Aura, Other)	Nominal	Nill

Table 1. **Dataset Description.**

3.3 Hardware and Software Specification

The analysis of the optimal algorithm for migraine classification was performed on Google Colaboratory or Colab, a cloud-based platform that offers access to robust computational resources (Carneiro *et al.,* 2018). The investigation was conducted using a MacBook Pro as the local interface to interact with the cloud environment.

3.3.1 Hardware Specification

- Local Device: MacBook Pro 13-inch (2022 model).
- Processor: Apple M2 chip that features an 8-core CPU and a 10-core GPU.
- Memory: 16 GB RAM facilitated efficient multitasking, enabling effortless transitions between several applications and browser tabs during project work.
- Operating System: macOS Sonoma, version 14.6.1 offers a reliable and protected platform for accessing cloud-based services.

3.3.2 Software Specification

- Programming Language: Python.
- Processor: Intel(R) Xeon(R) CPU @ 2.20GHz.
- Memory: 12.67 GB RAM.
- Disk: 107.72 GB total, 74.88 GB free.

3.4 Data Preprocessing

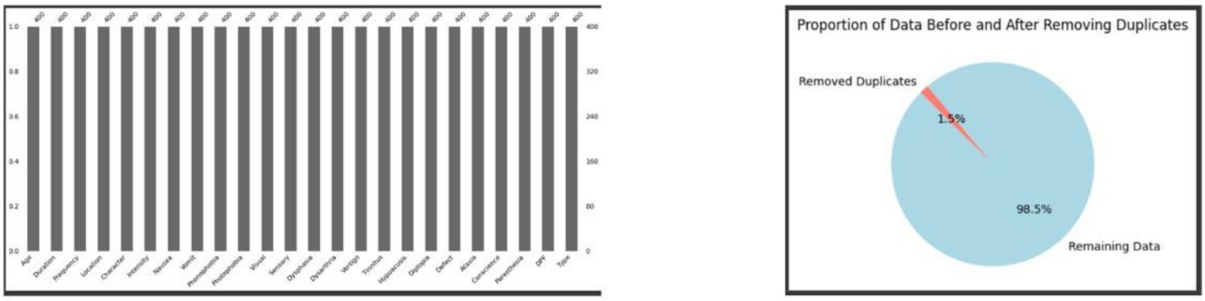
The data preprocessing involved multiple essential steps to ensure the dataset quality for analysis. Firstly, we checked for the missing values using the missingno library, and it was ensured that there were no missing values, with each column containing 400 non-missing entries. Secondly, duplicate rows were identified, revealing 6 duplicates, which were then removed, leaving 394 unique records.

Following this, outliers were identified using the Interquartile Range (IQR) method (Ur Rehman and Belhaouari, 2021), visualized through box plots. Instead of removing the outliers, they were capped to preserve the data integrity while controlling for extreme values. After addressing outliers, standardisation was applied using Z-score transformation via the StandardScaler library, ensuring each numerical feature had a mean of 0 and a standard deviation of 1 (Gao *et al.,* 2019), critical for algorithms sensitive to scale. This process confirmed that several features lacked variability, as indicated by a standard deviation of 0, suggesting minimal influence on classification outcomes.

Finally, categorical data in the "Type" column was encoded using label encoding method (Qiu and Liu, 2023), transforming the 7 unique values into numerical representations suitable for model training. The encoded value corresponding to the categorical value are represented below:

- Basilliar-Type aura: 0
- Familial hemiplegic migraine: 1
- Migraine without aura: 2
- Other: 3
- Sporadic hemiplegic migraine: 4
- Typical aura with migraine: 5
- Typical aura without migraine: 6

Together, these steps created a well-prepared dataset for reliable and interpretable analysis. Figure 3 represents the data preprocessing steps done.



Bar Plot representation of analysing missing values.

Pie chart representing proportion of data before and after removing duplicates.

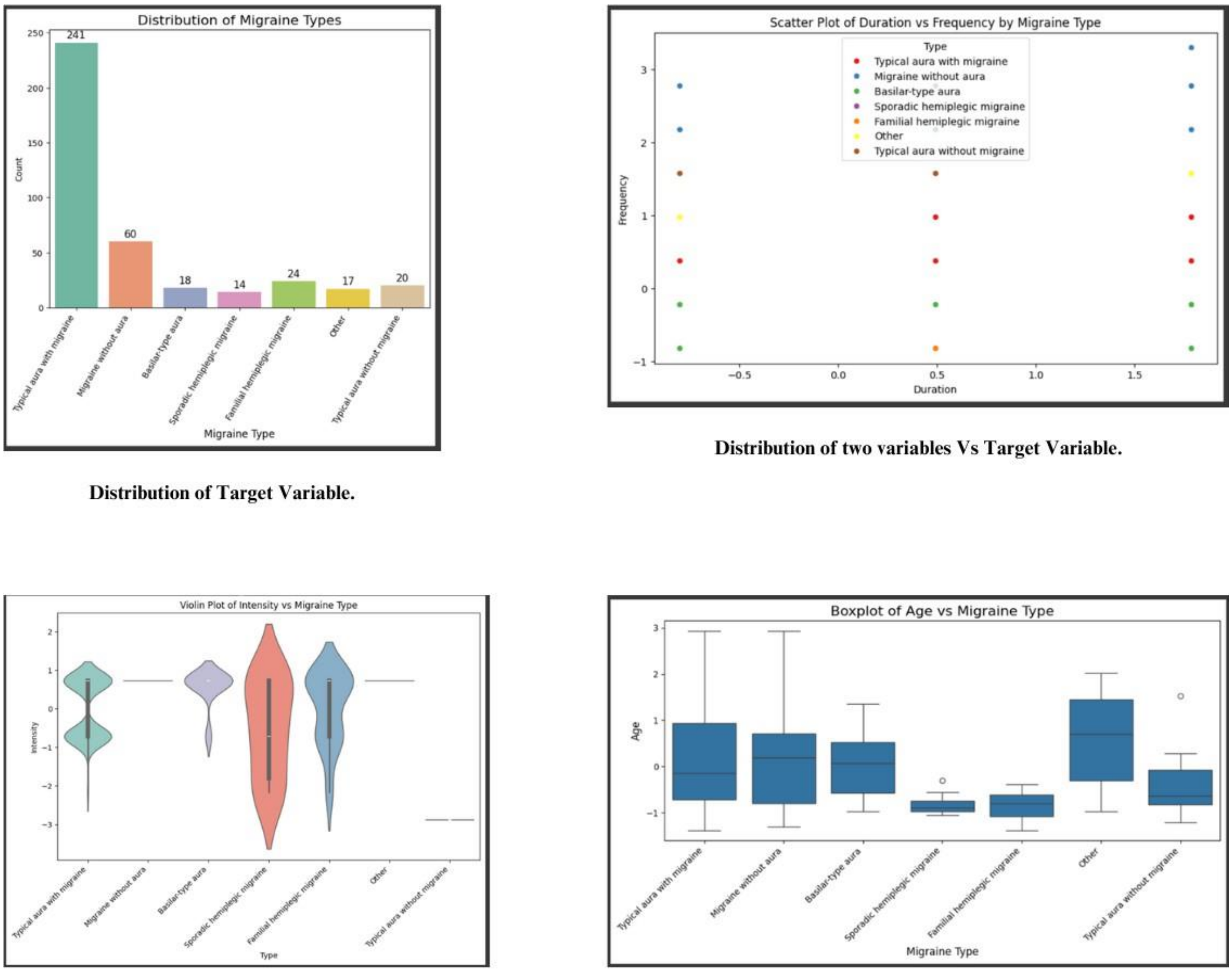
Figure 3: Data Preprocessing Steps.

3.5 Exploratory Data Analysis

The EDA aimed to uncover key insights and patterns (DSouza, 2020) within the migraine dataset through visual analysis. Initially, the distribution of the target variable was analysed using a bar plot, which revealed an imbalance, particularly favouring the "Typical aura with migraine" class. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied later to balance the dataset.

Secondly, a boxplot was used to examine the age distribution across migraine types, showing that certain types (e.g., "Basilar-type aura" and "Sporadic hemiplegic migraine") are associated with specific age ranges, suggesting age as a valuable classification feature. Following this, a violin plot highlighted the distribution of intensity scores among migraine types, indicating that intensity variations might enhance classification accuracy.

Finally, a scatter plot was used to explore the relationship between two continuous variables Duration and Frequency relative to the target variable. The plot indicated some clustering by migraine type, suggesting that these variables, when combined, could aid in differentiating between classes. These visual analyses provide foundational insights for selecting key features in the migraine classification process. Figure 4 depicts the EDA process.



Distribution of Target Variable.

Distribution of two variables Vs Target Variable.

Distribution of Intensity Vs Target variable.

Distribution of Age Vs Target Variable.

Figure 4. **EDA processes.**

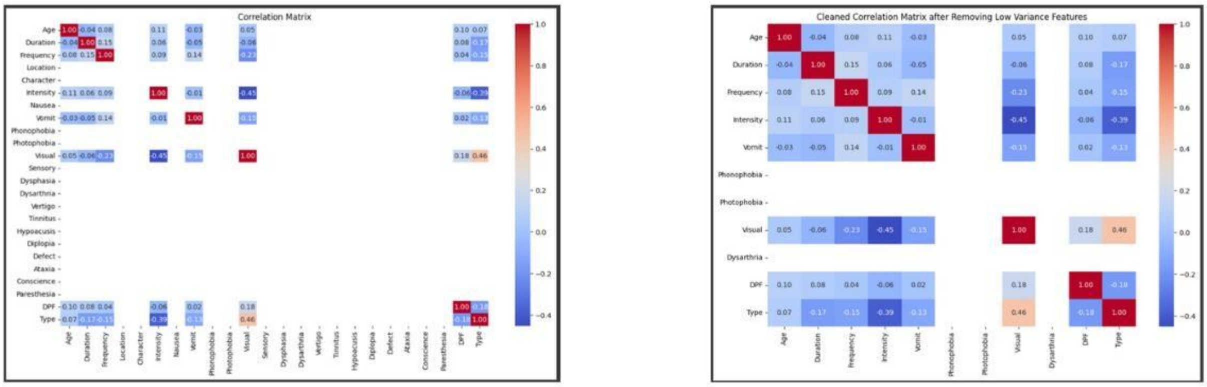
3.6 Feature Selection Process

To identify features most relevant to migraine classification, a correlation matrix analysis was conducted supplemented by insights derived from relevant domain literature. This approach enabled a nuanced selection process, balancing statistical relevance with clinical insights.

While evaluating the linear relationships among all numerical variables and their association with the target variable. Strong correlations, such as those between intensity, visual symptoms, and the target variable, were retained due to their potential predictive power. Conversely, features with minimal correlations, such as “location” and “character,” were flagged as less relevant for classification.

Features exhibiting low variance across the dataset, such as "Photophobia," "Phonophobia," and "Dysarthria," were reviewed for possible exclusion. Although these features displayed limited statistical contribution, they hold clinical significance in distinguishing migraine types (Pescador Ruschel & De Jesus, 2023; Demarquay et al., 2018). Given their medical relevance, these features were preserved to maintain a clinically comprehensive dataset, ensuring meaningful contributions to migraine classification.

This balanced approach allowed for a streamlined feature set that leverages statistically robust predictors while preserving essential clinical attributes, improving the model’s accuracy and applicability in migraine diagnosis. Figure 5 illustrates the correlation matrix prior to and after feature selection process.



Correlation matrix before feature selection.

Correlation matrix of selected feature.

Figure 5. **Correlation Matrix.**

3.7 Data Augmentation Process

To address the small sample size and imbalance in migraine types identified in preprocessing stage, SMOTE technique was applied. This data augmentation approach generates synthetic samples by interpolating between existing minority class instances and their nearest neighbours, enhancing class balance without duplication (Temraz & Keane, 2022). Following SMOTE, the dataset expanded from 394 to 1,687 records, achieving a balanced distribution across all seven migraine classes (see Figure 6). This balanced dataset better supports effective model training and classification across all migraine types.

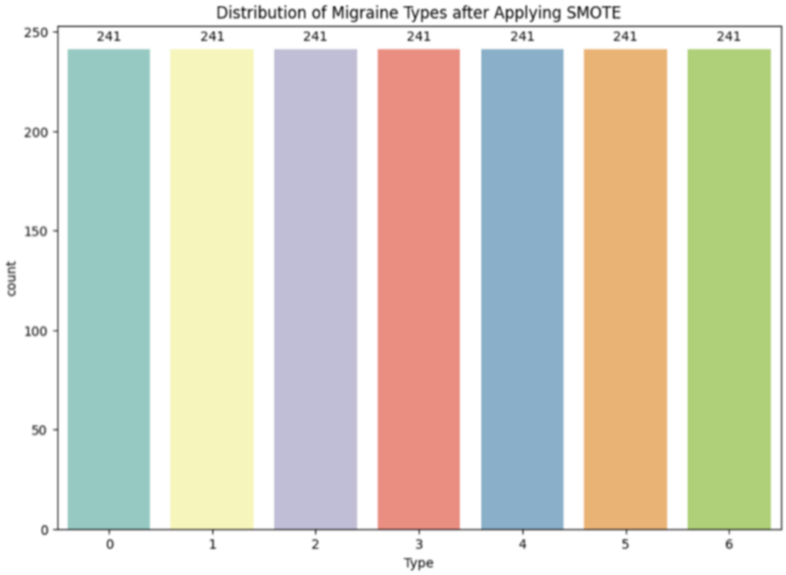


Figure 6. Distribution of Target variable following SMOTE.

3.8 Model Selection

From the literature analysis done, we chose some of the effective ML/DL models for migraine classification on tabular data. SVM, KNN, DT, RF, and TabNet. These models were chosen for their efficiency, interpretability, and proven performance on similar classification tasks.

3.9 Data Partitioning

The dataset was divided into training and testing subsets using 70:30 ratio, with 70% allocated for training and 30% reserved for testing to evaluate the model's predictive performance. This ratio was chosen to balance model learning and testing, providing a substantial amount of data for training while ensuring a sufficient test set to validate the model's generalisability to new, unseen data. This approach is particularly crucial in medical research, where data is limited, and models must undergo rigorous testing due to their potential impact on patient outcomes

(Gunawan Kurnia, 2024). After splitting, the training subset contained 1180 records, and the testing subset contained 507 records (see Figure 7), each with 10 features.

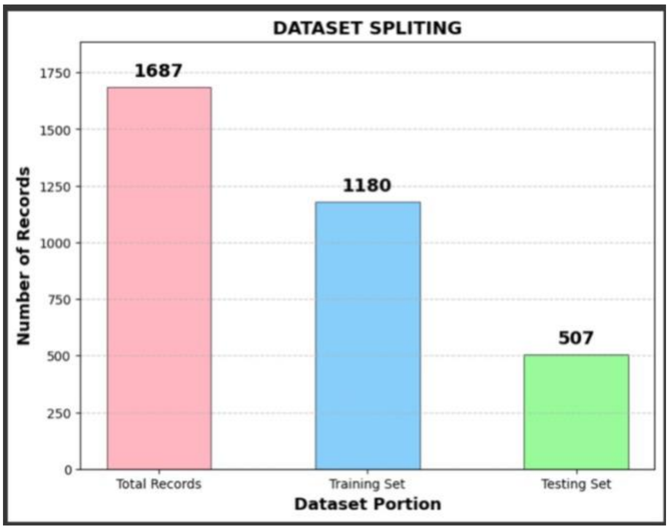


Figure 7. Number of records after data partitioning.

3.10 Model Evaluation

To classify migraine types effectively, five algorithms were evaluated prior to and after HPT. The performance was assessed using accuracy, precision, recall, F1 score, AUC-ROC, MCC and confusion matrix.

To prevent overfitting, we used stratified K-fold cross-validation with n_splits= 10, shuffle=True, and random_state=42. We employed GridSearchCV for hyperparameter optimisation, using n_jobs= -1 for parallelization and verbose=1 for monitoring the search progress. The Table 2 provides a comparison of the default and tuned parameters for each model, showing the parameter values pre and post HPT. The comparison illustrates how the tuning process influenced the model's performance.

Furthermore, the confusion matrix and AUC-ROC curve for each model, both before and after HPT, are illustrated in Figure 8 to 11. The performance comparison of the models are shown in Tables 3 and 4, respectively.

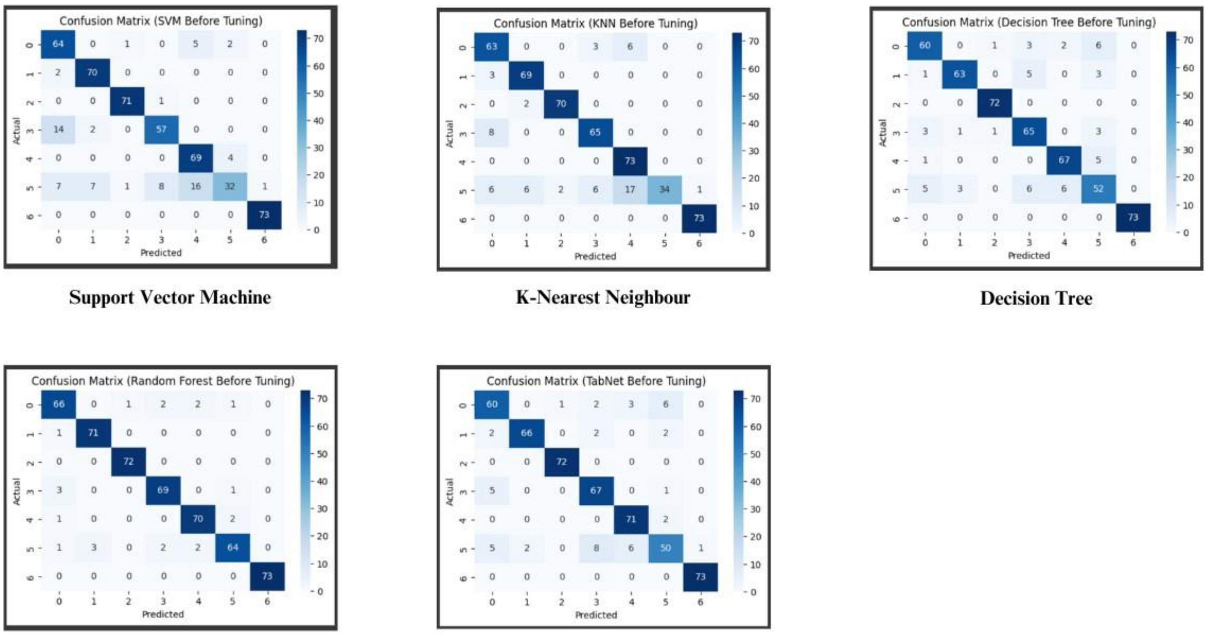
The effectiveness of the ML/DL models pre and post HPT reveals significant differences in their underlying mechanisms and their interactions with the data. Among all the classification methods considered, RF, the ML algorithm stands out as the most promising model, as it consistently outperforms others in terms of all the metrics with accuracy 95.8%. This is because of the RF’s ensemble learning approach, which consists of multiple DTs to reduce overfitting and improved generalisation. In contrast, DT showed underfitting due to its single-tree structure, limiting capacity to generalise well on unseen data which is reflected in its relatively lower CV score of 89.3% even with HPT.

On the other hand, SVM and KNN had significant improvement after HPT, with both algorithm’s accuracy reached up to 93% and 92.8% respectively. These models exhibited substantial improvements in F1-score and MCC, suggesting improved balance and reliability in their predictions. The improvement in SVM can be attributed to the fine-tuning of its regularization parameter C where others remain default. The enhancement in KNN is likely due to the optimisation of the number of neighbours and distance metrics, which results in more precise decision boundaries. Initially, KNN's performance was limited by its sensitivity to noise and local patterns. However, HPT mitigated these issues, enabling it to capture more global structures in the data.

TabNet, despite being a DL model, did not exhibit as such an improvement as SVM and KNN following HPT. This could be attributed to the inbuilt design of TabNet, which already includes mechanisms such as sequential attention and feature selection, thereby optimising its ability to learn from tabular data efficiently. This model was analysed with maximum epochs 100 with early stopping parameter 10 and the batch size was set to 64 since the dataset size is comparatively small. The optimiser algorithm used was adam. Like RF model, its architecture may already be well-suited to the task, as evidenced by its stable performance both pre and post HPT. Nevertheless, it was unable to surpass RF's performance, despite achieving comparable CV score of 91.6% post HPT. This may be due to RF's bagging approach that provides greater robustness against overfitting (IBM, 2023a), a well-known challenge in deep learning models such as TabNet.

Model	Parameter	Default Value	Tuned Value
SVM	C	1.0	10
	SVM Kernel	Radial Basis Function (RBF)	Radial Basis Function (RBF)
	SVM Gamma	Scale	Scale
KNN	n_neighbors	5	3
	KNN Weights	Uniform	Distance
	KNN p	2 (Euclidean distance)	1 (Manhattan distance)
DT	Criterion	Gini	Gini
	DT max_depth	None	None
	DT min_samples_leaf	1	1
	DT min_samples_split	2	2
	DT ccp_alpha	0.0	0.001
RF	n_estimators	100	100
	RF Criterion	Gini	Entropy
	RF max_depth	None	20
	RF min_samples_split	2	5
	RF min_samples_leaf	1	1
	RF max_features	sqrt	log2
TabNet	n_d	8	24
	TabNet n_a	8	24
	TabNet n_steps	3	5
	TabNet gamma	1.3	1.0
	TabNet lambda_sparse	0.04	0.001
	TabNet optimizer_fn	torch.optim.Adam	torch.optim.Adam
	TabNet optimizer_params	0.02	0.02
	TabNet mask_type	sparsemax	sparsemax
	TabNet max_epochs	100	100
	TabNet patience	10	10
	TabNet batch_size	64	64
	TabNet virtual_batch_size	128	128

Table 2. Comparison of parameters value Before and After Optimisation.



Support Vector Machine

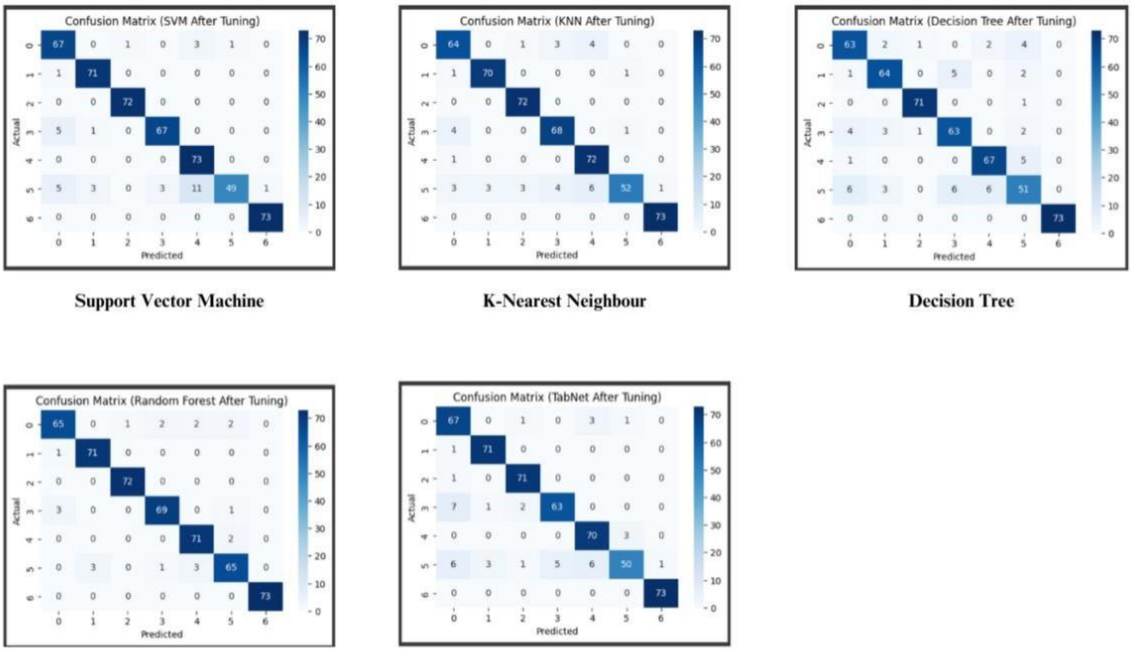
K-Nearest Neighbour

Decision Tree

Random Forest

TabNet

Figure 8. **Confusion matrix Prior HPT.**



Support Vector Machine

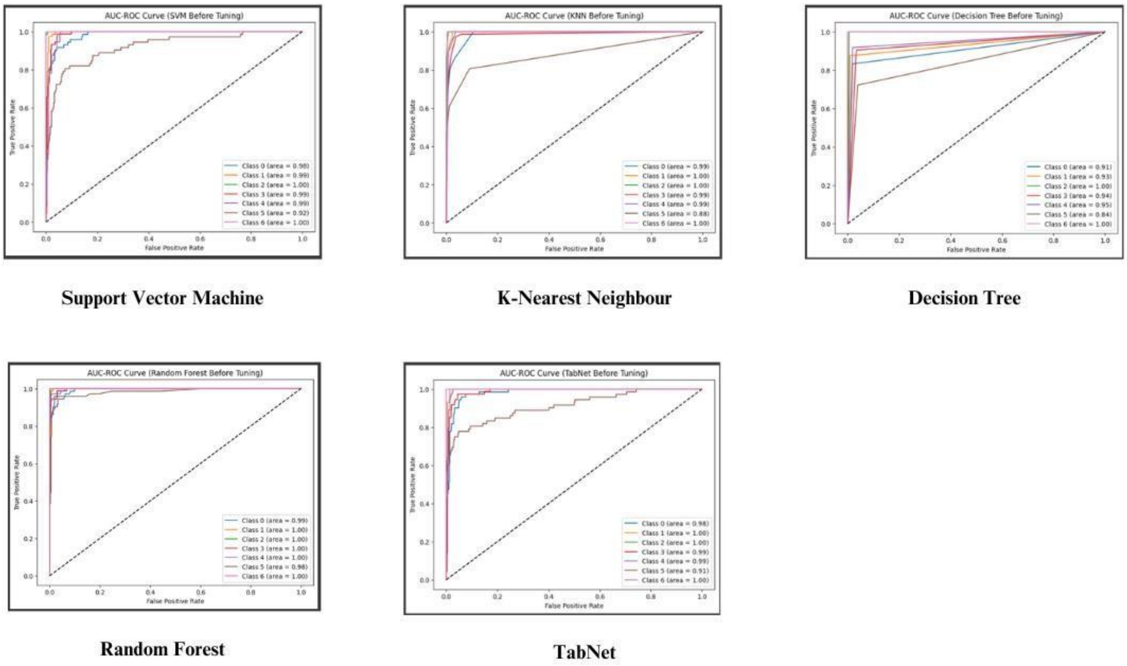
K-Nearest Neighbour

Decision Tree

Random Forest

TabNet

Figure 9. **Confusion matrix Post HPT**



Support Vector Machine

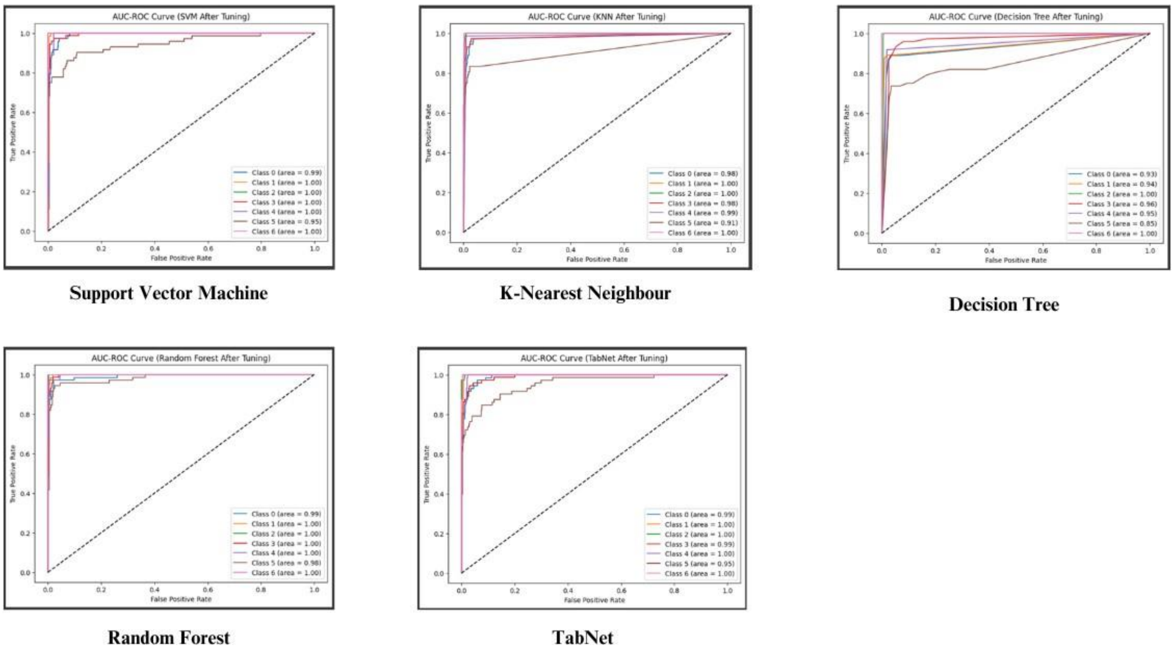
K-Nearest Neighbour

Decision Tree

Random Forest

TabNet

Figure 10. AUC-ROC Curve Prior HPT.



Support Vector Machine

K-Nearest Neighbour

Decision Tree

Random Forest

TabNet

Figure 11. AUC-ROC Curve Post HPT.

MODEL	PRECISION	RECALL	F1 SCORE	AUC-ROC	MCC	ACCURACY (%)	10-FOLD CV-ACCURACY
SVM	86.4	85.9	85.0	98.2	0.83	85.9	86.6
KNN	89.7	88.1	87.3	97.7	0.86	88.1	88.1
DT	89.1	89.1	89.1	93.7	0.87	89.1	88.0
RF	95.6	95.6	95.6	99.5	0.94	95.6	92.4
TABNET	90.4	90.5	90.3	98.2	0.88	90.5	88.1

Table 3. Classification report of all algorithms before HPT.

MODEL	PRECISION	RECALL	F1 SCORE	AUC-ROC	MCC	ACCURACY (%)	10-FOLD CV-ACCURACY
SVM	93.6	93.0	92.8	98.9	0.92	93.0	91.6
KNN	93.0	92.8	92.6	97.9	0.91	92.8	91.9
DT	89.0	89.1	89.0	94.7	0.87	89.1	89.3
RF	95.8	95.8	95.8	99.6	0.95	95.8	92.6
TABNET	91.8	91.1	90.7	98.6	0.89	91.1	91.6

Table 4. Classification report of all algorithms after HPT.

3.11 Computational Efficiency Analysis

We assessed the computational efficiency of each model based on the factors:

- Training time of the model
- Prediction time of the model.
- Memory usage during training.
- Memory usage during prediction.
- Model Size.

Where the time was calculated in Seconds and memory usage in Megabytes.

Model Size is the quantity of storage capacity needed to store the trained model (Saeed, 2023). This encompasses the overall count of parameters, in addition to any supplementary data structures, such as trees in decision trees or layers in neural networks, that are essential for the model's functioning. A lower model size is beneficial in situations when there are limited resources for deploying models. The computational efficiency analysed for each model is depicted in the Figure 12 and the comparison is shown in the table 5.

The computational efficiency of the models varied significantly across different metrics. KNN and DT exhibited the fastest training times at 0.1 and 0.11 seconds, respectively, followed closely by SVM at 0.26 seconds. RF took longer to train, requiring 0.9 seconds, while TabNet had the longest training time at 10.8 seconds. In terms of prediction time, TabNet was the fastest, predicting in just 0.04 seconds. The remaining models SVM, RF, and DT had similar prediction times, ranging between 0.10 and 0.13 seconds, while KNN was the slowest, with a prediction time of 0.19 seconds.

Memory usage also differed substantially across models. During training, TabNet consumed high memory at 1.13 MB, far exceeding the other models, which used between 0.02 MB (DT) and 0.14 MB (KNN and RF). Similarly, during prediction, TabNet continued to have the highest memory consumption (0.11 MB), while SVM and DT used the least 0.01 MB. Furthermore, the model sizes for all models were very small, with SVM, KNN, DT, and RF all at 0.046 MB, and TabNet being slightly smaller at 0.035 MB.

The computational efficiency analysis shows that KNN and DT are the fastest to train, making them suitable for environments requiring frequent updates. However, TabNet excels in prediction speed, despite its longer training time, which makes it ideal for real-time applications. KNN's slower prediction time limits its use for immediate decision-making, while SVM and DT's low memory usage makes them better suited for resource-constrained settings. Even though, TabNet’s small model size and fast predictions make it best suitable, it failed in the training time taken whereas, RF balances memory, speed, and size positioning it as an adaptive option.

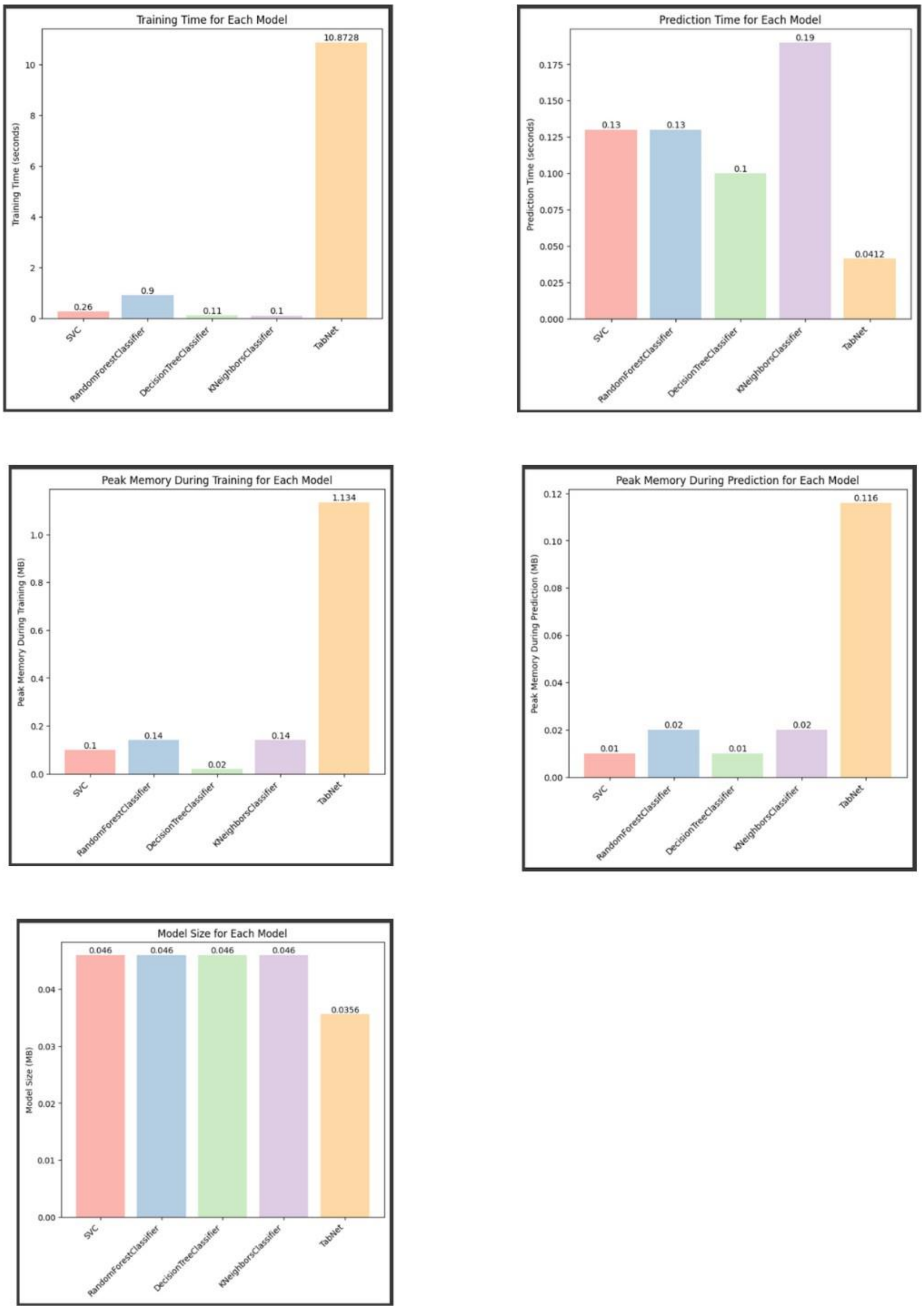


Figure 12. Computational Analysis of the model.

MODEL	Training Time (s)	Prediction Time (s)	Memory usage (Training) (MB)	Memory Usage (Prediction) (MB)	Model Size (MB)
SVM	0.26	0.13	0.1	0.01	0.046
KNN	0.1	0.19	0.14	0.02	0.046
DT	0.11	0.1	0.02	0.01	0.046
RF	0.9	0.13	0.14	0.02	0.046
TABNET	10.8	0.04	1.13	0.11	0.035

Table 5. Comparison of Computational Efficiency of all Algorithms.

3.12 Interpretability Analysis Using SHAP

To make model decisions clear and actionable for healthcare professionals interpretability analysis was performed for the effective ML and DL algorithm identified which are RF and TabNet. The technique used to identify the interpretability of the model was SHAP (Nguyen *et al.*, 2021). SHAP is used to better understand the model's prediction for a particular input, the contribution of each feature to this prediction is calculated.

The feature that holds the most importance is the one that exhibits the most extensive range of SHAP values acquired for that feature. The relative contribution is determined in relation to the base value. Each dot represents the contribution of a specific feature, where blue denotes lower values and red suggests higher values in a SHAP plot.

Random Forest Model

The RF SHAP summary plot (see Figure 13) provides an overview of the way in which various features interact and contribute to the model's predictions throughout the dataset. It was noted that features such as Age, Intensity, and Vomit have substantial interactions with other features, such as Frequency and Duration, which underscores their significant impact on the model's predictions. In contrast, features such as Phonophobia and Photophobia exhibit minimal interaction with other features as they exhibit narrower spreads of SHAP values. This suggests that these features have fewer complex relations with other features in the dataset, resulting in more direct and isolated contributions to the model's predictions.

At the same time, the RF SHAP waterfall plot (see Figure 14) provides a detailed explanation of how individual features contribute to the prediction for class 1 for the first instance in the test set. Starting from a baseline value of 0.143, reflective of the average probability for class 1 throughout the dataset, the Figure 14 illustrates how each feature either increases or decreases the predicted chance for this specific case. Phonophobia has the most significant positive impact, increasing the predicted probability by 0.11 units, pushing the prediction toward class 1. Conversely, Duration and Photophobia have the largest negative impacts, decreasing the prediction by 0.07 and 0.06, respectively. Other features such as Intensity and Vomit contribute smaller positive and negative effects, respectively. The final prediction of 0.143 reflects the cumulative influence of these features on the model's decision, showing that while some features increase the probability, others tend to reduce the probability of this instance being classified as class 1.

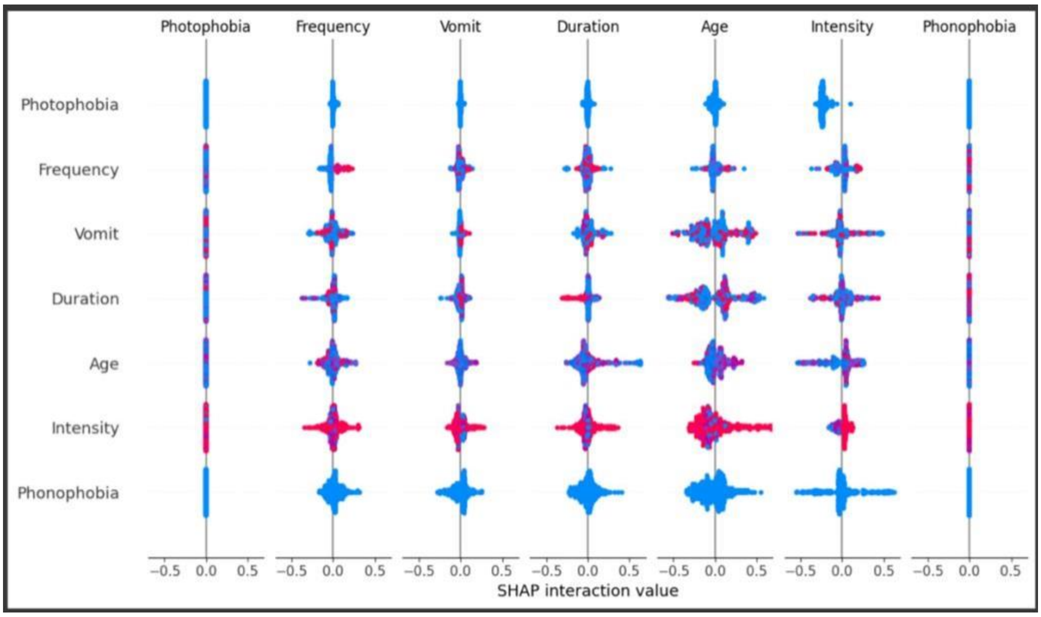


Figure 13. **SHAP summary plot for RF model Interpretability Analysis.**

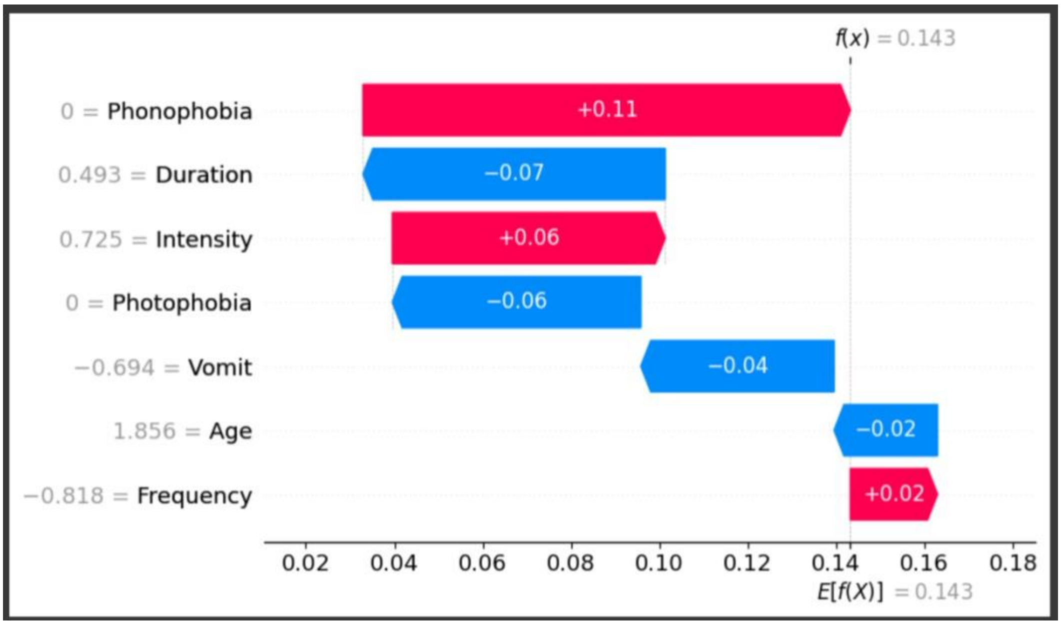


Figure 14. **Waterfall plot Analysis of RF model.**

TabNet Model

While observing the TabNet summary plot (see Figure 15) the features such as Age, Intensity, Frequency, and Duration exhibit substantial interactions with other features, indicating that their importance varies across different instances. In contrast, features such as Phonophobia and Photophobia exhibit narrower spreads of SHAP values, indicating that they contribute in a more direct and isolated way to the model's predictions. Additionally, the presence of a wider SHAP value distribution across multiple features strongly emphasis that TabNet personalises its decision-making by selecting different features for different instances, making it an adaptive model.

On the other hand, the waterfall plot (see Figure 16) provides a detailed explanation of how individual features contribute to the classification decision for a specific instance in the test set. Age has the highest positive impact (+1.25), pushing the prediction toward a higher classification probability, while Intensity (-0.63) and DPF (-0.58) contribute the most significant negative impacts, reducing the probability. Notably, Dysarthria appears in this instance, even though it was not identified as a key feature in the summary plot. This suggests that TabNet dynamically includes features based on each instance or cases, making it more flexible in its decision-making.

This instance-based feature selection strengthens the adaptability of TabNet allowing for more personalised predictions**,** but it also introduces challenges in model interpretability as the changing feature importance per case complicates clinical validation.

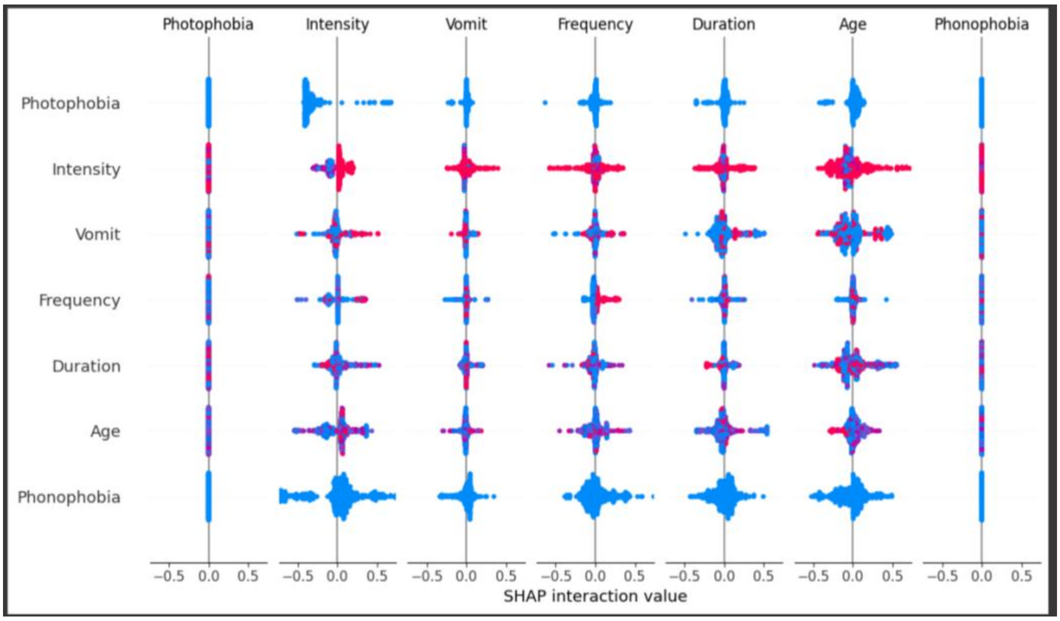


Figure 15. **SHAP summary plot for TabNet model Interpretability Analysis.**

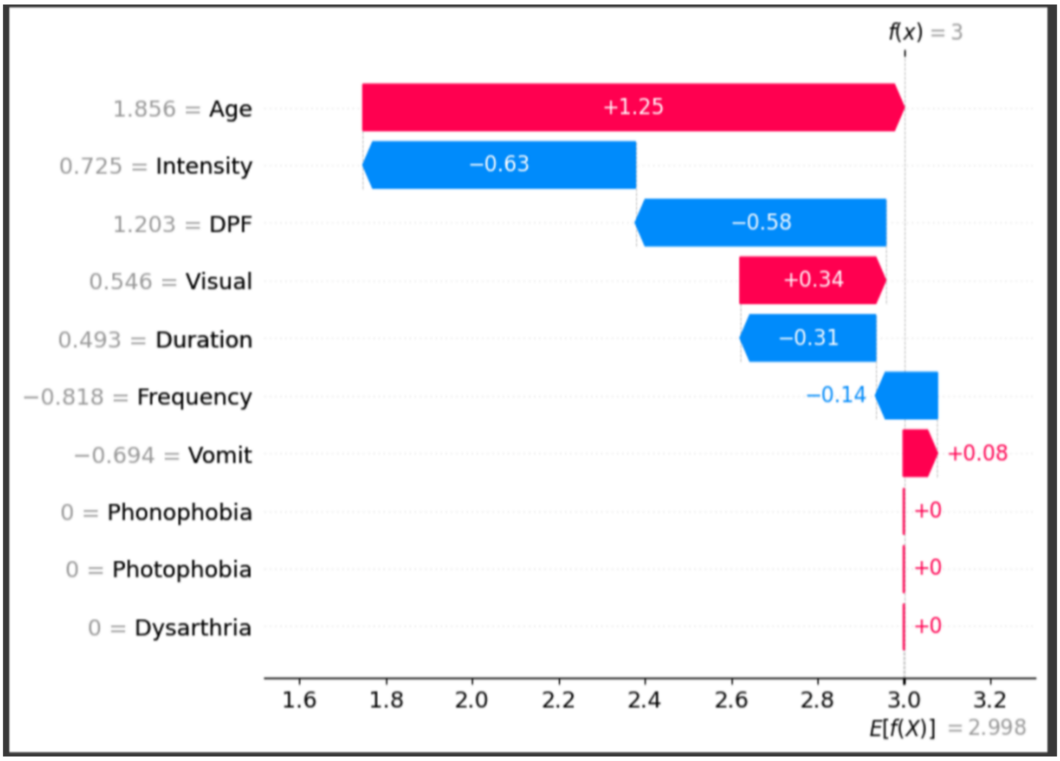


Figure 16. Waterfall plot Analysis of TabNet model.

4.0 Result and Analysis

In this study, we conducted a comprehensive evaluation of ML/DL models for classifying migraines in low-resource healthcare settings, assessing key performance metrics like accuracy, precision, recall, F1 score, AUC-ROC, and MCC, as well as computational efficiency and interpretability.

The ML model, RF, demonstrated high accuracy, with an accuracy score of 95.8% and a 10-fold cross-validation accuracy of 92.6%. RF also achieved superior metrics across the board, including an F1 score of 95.8, an MCC of 0.95, and an AUC-ROC of 99.6, indicating robust performance and high reliability compared to DL model, TabNet, which with an accuracy of 91.1%, MCC of 0.89, and AUC-ROC of 98.6.The RF model's high AUC-ROC score illustrates its strong discriminatory ability between classes, further affirming its suitability for migraine classification.

Comparatively, other ML models such as SVM, KNN, and DT achieved notable results, yet were less optimal than RF. For instance, SVM reached an accuracy of 93.0%, with an F1 score of 92.8, an MCC of 0.92, and an AUC-ROC of 98.9. While these metrics are strong, RF outperformed SVM, particularly in AUC-ROC and cross-validation accuracy, suggesting better generalisability. KNN, with an accuracy of 92.8% and an MCC of 0.91, and DT, on the other hand, yielded an accuracy of 89.1% and an MCC of 0.87, further reinforcing RF as the top choice in terms of balanced performance metrics.

Beyond these performance metrics, our study emphasised computational efficiency and interpretability to align with the needs of low-resource environments. RF’s computational requirements were efficiently met through cloud-based Google Colab, which provided a cost-effective and accessible environment. This setup allowed for HPT and model evaluation without demanding advanced local hardware, highlighting RF's suitability for economically challenged regions. TabNet, despite its rapid prediction capabilities, required intensive computational resources during training, presenting challenges in low-resource settings.

Interpretability was another critical focus, given the need for healthcare practitioners to understand the model’s decision-making process. SHAP interpretability analysis was used to interpret the RF and TabNet models where RF relied heavily on clinically significant features, such as photophobia and intensity, validating its decision-making in a clinically relevant manner and follows structured, consistent decision-making process, while TabNet feature selection process across each instance.

4.1 Tradeoff Analysis: RF vs. TabNet

RF and TabNet offer distinct advantages and trade-offs in terms of performance, computational efficiency, interpretability influencing their suitability for migraine classification in low-resource healthcare settings. RF demonstrates superior predictive performance, achieving higher accuracy, precision, recall, and F1-score, along with a more consistent feature importance across cases. This structured decision-making enhances trust and reliability, making it easier for healthcare professionals to validate and interpret predictions. In contrast, TabNet dynamically selects different features per instance, allowing for greater feature adaptability and potentially more personalised predictions. However, this adaptability introduces interpretability challenges, as the reasoning behind predictions varies across cases, making clinical validation more difficult. The dynamic nature of feature importance can make it difficult for clinicians to understand the model's reasoning, potentially hindering trust and adoption. Furthermore, RF is significantly more efficient, requiring only 0.9 seconds for training compared to TabNet’s 10.8 seconds and consuming far less memory. While RF has a slightly slower prediction speed, its overall computational efficiency makes it more suitable for deployment in resource-constrained environments.

Given these trade-offs, RF is the more practical choice for real-world deployment in healthcare due to its high performance, computational efficiency, and nature of the interpretability. TabNet, while offering greater feature adaptability, poses challenges in clinical adoption due to its dynamic decision-making and lower transparency. If personalised case-specific predictions are prioritised and computational resources are available, TabNet may provide value, but its lower clinical interpretability may limit its real-world impact in healthcare applications compared to RF.Detailed tradeoff analysis can be seen in Figure 17 and from the Figure 18.

In comparison to other studies in the literature (see table 6), which report accuracy levels of RF as 78–99% (Dhiyaussalam et al., 2020; S. S. Esfahan et al., 2023; David et al., 2023) but may overlook constraints faced in resource-limited environments, our results are intentionally contextualised. By evaluating models not only on accuracy but also on computational efficiency and interpretability, our work aligns closely with the practical requirements of low-resource healthcare settings. This approach ensures that the model not only performs effectively but also generalises well across various migraine subtypes, establishing it as a viable and reliable solution for healthcare providers in such contexts.

Sl. No	Study	Outcome by the model	Dataset type	Effective Algorithm identified	Accuracy of the effective model
1	(Romould et al., 2024)	Classification of 6 types of migraine	Migraine Tabular dataset	KNN	85%
2	(S. S. Esfahan et al., 2023)	Classification of migraine vs. tension-type headaches (TTH)	Psychological and demographic data	RF	97.92%
3	(David et al., 2023)	Classification of 7 migraine types	Clinical dataset with 24 features	RF with Scatter Search	98.26%
4	(S. N. Mudassir and R. M, 2024)	Classification of various migraine subtypes	Clinical and patient-reported data	TabNet	98% overall; 99% for specific migraine subtypes
5	(Fu et al., 2024)	Classification of migraine with aura vs. migraine without aura	Structural and functional MRI data	RF	78.1%

6	(Dhiyaussalam et al., 2020)	Classification of migraine, tension-type headache (TTH), and cluster headaches	Migbase dataset with 39 features	RF	99.56%
7	(Tahhan et al., 2024)	Risk assessment and classification of migraine among university students	Survey data on lifestyle, dietary habits, and behavioral factors from university students	Linear SVM	92.7%
8	Chen et al., 2024)	Classification of Migraine Without Aura vs. Healthy Controls	Resting-state fMRI data	SVM with MVPA	81.54% (First cohort); 76.47% (Second cohort)
9	(Subasi, Ahmed and Alickovic, 2018)	Classification of migraine vs. healthy controls	EEG data collected during flash stimulation (4 Hz)	RF	85.18%
10	(Fu et al., 2022)	Classification of migraine without aura (MwA) vs. healthy controls and prediction of tVNS treatment efficacy	fMRI data from 70 MwA patients and 70 healthy controls	SVM for classificat ion, SVR for prediction	79.3%
11	(Qawasmeh et al., 2020).	Classification of headache types: Migraine, Cluster Headache, Tension-Type Headache, and Secondary Headache	Data collected from patients	RF	99.1% (Migraine), 93% overall accuracy
12	(Hsiao et al., 2023)	Classification of chronic migraine vs. healthy controls	EEG data from	DT	87.5%
13	(Nie et al., 2023).	Classification of migraine patients vs. healthy controls	Resting-state fMRI data	SVM with combined features	96.81%

14	(Subasi et al., 2019)	Classification of migraine patients vs. healthy controls	EEG data with and without photic stimulation	RF	85.95%
15	(Kazemi and Katibeh, 2018)	Classification of pediatric migraine without aura patients vs. Healthy controls	EEG data	SVM	93%
16	Proposed Study	Classification of migraine subtypes	Tabular Data	RF - ML Model TabNet - DL model	95.8% 91.1%

Table 6. **Comparative Analysis of Effective Model Performance with Related Studies.**

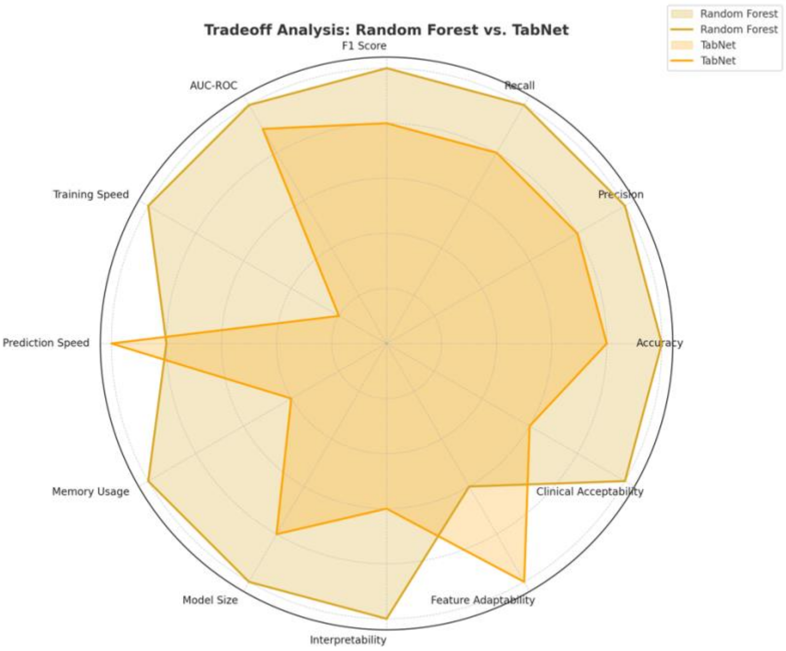
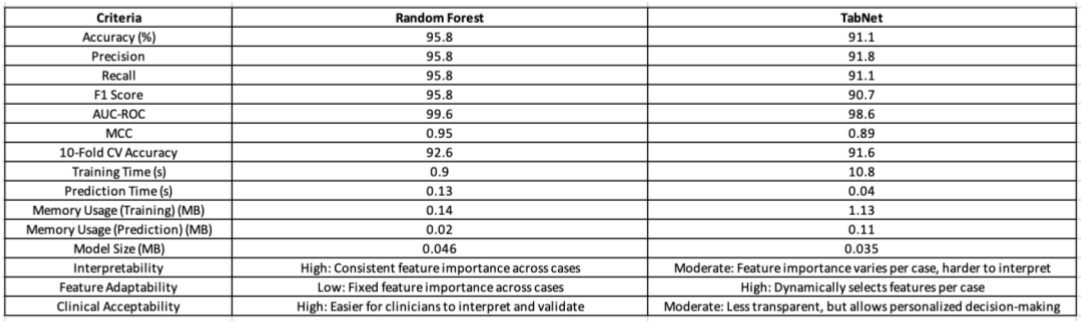


Figure 17. **Tradeoff Analysis Chart: Random Forest vs. TabNet.**



Criteria	Random Forest	TabNet
Accuracy (%)	95.8	91.1
Precision	95.8	91.8
Recall	95.8	91.1
F1 Score	95.8	90.7
AUC-ROC	99.6	98.6
MCC	0.95	0.89
10-Fold CV Accuracy	92.6	91.6
Training Time (s)	0.9	10.8
Prediction Time (s)	0.13	0.04
Memory Usage (Training) (MB)	0.14	1.13
Memory Usage (Prediction) (MB)	0.02	0.11
Model Size (MB)	0.046	0.035
Interpretability	High: Consistent feature importance across cases	Moderate: Feature importance varies per case, harder to interpret
Feature Adaptability	Low: Fixed feature importance across cases	High: Dynamically selects features per case
Clinical Acceptability	High: Easier for clinicians to interpret and validate	Moderate: Less transparent, but allows personalized decision-making

Figure 18. **Tradeoff Comparision: Random Forest vs. TabNet.**

5.0 Ethical Implications and Deployment Challenges

The deployment of our model in low-resource healthcare settings presents both ethical challenges and practical implementation constraints. Due to time limitations, we utilised a secondary dataset and applied the SMOTE technique to increase the sample size to 1,687 instances. However, this dataset remains relatively small, which may impact the generalisability of the model and lead to disparities in performance across different demographic groups, reducing fairness in clinical decision-making.

To mitigate these concerns, our future work will focus on validating the model on a larger primary dataset and comparing its predictions with professional diagnoses to assess real-world accuracy. Additionally, to address bias in training data, we plan to ensure more diverse and representative datasets and incorporate fairness-aware algorithms to enhance model equity and reliability in clinical applications.

6.0 Conclusion

In this study, we evaluated various ML/DL algorithms for classifying migraine subtypes, focusing on their applicability in low-resource healthcare settings. Our analysis found that RF, a ML model, was the most effective, achieving 95.8% accuracy and outperforming other models in terms of additional metrics. TabNet, the DL model, achieved slightly lower performance with 91.1% accuracy. RF also demonstrated better computational efficiency compared to TabNet, making it more suitable for resource-constrained environments. Furthermore, RF's interpretability was enhanced using SHAP summary and waterfall plots, ensuring transparency for clinical use. These findings suggest that AI-driven diagnostic tools, such as RF, could reduce reliance on costly medical imaging and specialised neurologists, offering a more accessible solution for migraine classification in low-resource healthcare settings.

References

Awad, M. and Khanna, R. (2015) 'Support vector machines for classification' *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* Springer, pp. 39–66.

Carneiro, T., Da Nóbrega, R.V.M., Nepomuceno, T., Bian, G., De Albuquerque, V.H.C. and Reboucas Filho, P.P. (2018) 'Performance analysis of Google Scholar colaboratory as a tool for accelerating deep learning applications', *Ieee Access,* 6, pp. 61677–61685.

Chen, Y., Xu, J., Wu, J., Chen, H., Kang, Y., Yang, Y., Gong, Z., Huang, Y., Wang, H. and Wang, B. (2024) 'Aberrant concordance among dynamics of spontaneous brain activity in patients with migraine without aura: A multivariate pattern analysis study', *Heliyon,* 10(9).

Chiang, C., Luo, M., Dumkrieger, G., Trivedi, S., Chen, Y., Chao, C., Schwedt, T.J., Sarker, A. and Banerjee, I. (2024) 'A large language model–based generative natural language processing framework fine-tuned on clinical notes accurately extracts headache frequency from electronic health records', *Headache: The Journal of Head and Face Pain,* 64(4), pp. 400–409.

Cowan, R.P., Rapoport, A.M., Blythe, J., Rothrock, J., Knievel, K., Peretz, A.M., Ekpo, E., Sanjanwala, B.M. and Woldeamanuel, Y.W. (2022) 'Diagnostic accuracy of an artificial intelligence online engine in migraine: A multi-center study', *Headache: The Journal of Head and Face Pain,* 62(7), pp. 870–882.

David, K.M.L., Sharmili, V.V.S., Babu, T. and Nair, R.R. (2023) *Migraine categorization using the scatter search and random forest classifier.* IEEE, pp. 1.

Dhiyaussalam, A. Wibowo, F. A. Nugroho, E. A. Sarwoko and I. M. A. Setiawan (2020) *Classification of Headache Disorder Using Random Forest Algorithm.* pp. 1.

DSouza, J. (2020) *Using exploratory data analysis for generating inferences on the correlation of COVID-19 cases.* IEEE, pp. 1.

Fu, C., Zhang, Y., Ye, Y., Hou, X., Wen, Z., Yan, Z., Luo, W., Feng, M. and Liu, B. (2022) 'Predicting response to tVNS in patients with migraine using functional MRI: A voxels-based machine learning analysis', *Frontiers in Neuroscience,* 16, pp. 937453.

Fu, T., Gao, Y., Huang, X., Zhang, D., Liu, L., Wang, P., Yin, X., Lin, H., Yuan, J. and Ai, S. (2023) 'Brain connectome-based imaging markers for identifiable signature of migraine with and without aura', *Quantitative Imaging in Medicine and Surgery*, 14(1), pp. 194.

Gao, Z., Ding, L., Xiong, Q., Gong, Z. and Xiong, C. (2019) 'Image compressive sensing reconstruction based on z-score standardized group sparse representation', *IEEE access*, 7, pp. 90640–90651.

Ge, R. and Chang, J. (2023) 'Disease burden of migraine and tension-type headache in non-high-income East and Southeast Asia from 1990 to 2019', *The journal of headache and pain*, 24(1), pp. 32.

Gulati, S., Guleria, K. and Goyal, N. (2022) *Classification of migraine disease using supervised machine learning.* IEEE, pp. 1.

GunKurnia (2024). *Choosing the Optimal Data Split for Machine Learning: 80/20 vs 70/30?* [online] Medium. Available at: https://medium.com/@gunkurnia/choosing-the-optimal-data-split-for-machine-learning-80-20-vs-70-30-0fd266710236 [Accessed 31 Sep. 2024].

Habehh, H. and Gohel, S. (2021) 'Machine learning in healthcare', *Current Genomics*, 22(4), pp. 291–300.

Hsiao, F., Chen, W., Wang, Y., Chen, S., Lai, K., Coppola, G. and Wang, S. (2023) 'Identification of patients with chronic migraine by using sensory-evoked oscillations from the electroencephalogram classifier', *Cephalalgia*, 43(5), pp. 03331024231176074.

IBM (2023a). *What Is Random Forest? | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/random-forest [Accessed 4 Sep. 2024].

Kazemi, S. and Katibeh, P. (2018) 'Comparison of parametric and non-parametric EEG feature extraction methods in detection of pediatric migraine without aura', *Journal of biomedical physics & engineering*, 8(3), pp. 305.

Khan, L., Shahreen, M., Qazi, A., Jamil Ahmed Shah, S., Hussain, S. and Chang, H. (2024) 'Migraine headache (MH) classification using machine learning methods with data augmentation', *Scientific Reports*, 14(1), pp. 5180.

Miotto, R., Wang, F., Wang, S., Jiang, X. and Dudley, J.T. (2018) 'Deep learning for healthcare: review, opportunities and challenges', *Briefings in bioinformatics,* 19(6), pp. 1236–1246.

Mitrović, K., Petrušić, I., Radojičić, A., Daković, M. and Savić, A. (2023) 'Migraine with aura detection and subtype classification using machine learning algorithms and morphometric magnetic resonance imaging data', *Frontiers in neurology,* 14, pp. 1106612.

Mortel, D., Kawatu, N., Steiner, T.J. and Saylor, D. (2022) 'Barriers to headache care in low-and middle-income countries', *Eneurologicalsci,* 29, pp. 100427.

N. N. Aung and W. Srimaharaj (2023) *Migraine Categorization based on the Integration of EMD and Naive Bayes Classification.* pp. 438.

Nguyen, H.T.T., Cao, H.Q., Nguyen, K.V.T. and Pham, N.D.K. (2021) *Evaluation of explainable artificial intelligence: Shap, lime, and cam.* pp. 1.

Nie, W., Zeng, W., Yang, J., Zhao, L. and Shi, Y. (2023) 'Classification of migraine using static functional connectivity strength and dynamic functional connectome patterns: A resting-state fmri study', *Brain Sciences,* 13(4), pp. 596.

ORHANBULUCU, F. and LATİFOĞLU, F. (2024) 'Development of a Machine Learning Based Clinical Decision Support System for Classification of Migraine Types: A Preliminary Study', .

Pradeep, R., Nemichandra, S.C., Harsha, S. and Radhika, K. (2020) 'Migraine disability, quality of life, and its predictors', *Annals of neurosciences,* 27(1), pp. 18.

Qawasmeh, A., Alhusan, N., Hanandeh, F. and Al-Atiyat, M. (2020) 'A high performance system for the diagnosis of headache via hybrid machine learning model', *International Journal of Advanced Computer Science and Applications,* 11(5).

Qiu, Q. and Liu, H. (2023) *Numerical Embedding of Categorical Features in Tabular Data: A Survey.* IEEE, pp. 446.

Rahman Siddiquee, M.M., Shah, J., Chong, C., Nikolova, S., Dumkrieger, G., Li, B., Wu, T. and Schwedt, T.J. (2023) 'Headache classification and automatic biomarker extraction from structural MRIs using deep learning', *Brain Communications,* 5(1), pp. fcac311.

Rathore, D.K. and Mannepalli, P.K. (2021) *A Review of Machine Learning Techniques and Applications for Health Care.* IEEE, pp. 4.

Romould, R.V., Singh, V., Gourisaria, M.K., Das, H. and Dash, B.B. (2024) *Deciphering Migraine Types: A Machine Learning Odyssey for Precision Prediction.* IEEE, pp. 1610.

Rundel, D., Kobialka, J., von Crailsheim, C., Feurer, M., Nagler, T. and Rügamer, D. (2024) *Interpretable machine learning for TabPFN.* Springer, pp. 465.

S. N. Mudassir and R. M (2024) *Enhancing Migraine Diagnosis and Classification with TabNet: A Data-Driven Approach.* pp. 679.

S. S. Esfahan, A. Haratian, A. Haratian, F. Shayegh and S. Kiani (2023) *Automatic classification of migraine and tension-type headaches using machine learning methods.* pp. 220.

Saeed, F. (2023). *Model Efficiency (The Unsung Hero of Deep Learning): Achieving Top-Notch Performance with Minimal Resource Utilization*. [online] www.linkedin.com. Available at: https://www.linkedin.com/pulse/model-efficiency-unsung-hero-deep-learning-achieving-top-notch-saeed-gkrke/ [Accessed 27 Aug. 2024].

Sanchez-Sanchez, P.A., García-González, J.R. and Ascar, J.M.R. (2020) 'Automatic migraine classification using artificial neural networks', *F1000Research,* 9.

Sasaki, S., Katsuki, M., Kawahara, J., Yamagishi, C., Koh, A., Kawamura, S., Kashiwagi, K., Ikeda, T., Goto, T. and Kaneko, K. (2023) 'Developing an artificial intelligence-based pediatric and adolescent migraine diagnostic model', *Cureus,* 15(8).

Subasi, A., Ahmed, A. and Alickovic, E. (2018) 'Effect of Flash Stimulation for Migraine Detection Using Decision Tree Classifiers', *Procedia Computer Science,* 140, pp. 223–229 Available at: 10.1016/j.procs.2018.10.332.

Subasi, A., Ahmed, A., Aličković, E. and Hassan, A.R. (2019) 'Effect of photic stimulation for migraine detection using random forest and discrete wavelet transform', *Biomedical signal processing and control,* 49, pp. 231–239.

Tahhan, Z., Hatem, G., Abouelmaty, A.M., Rafei, Z. and Awada, S. (2024) 'Design and validation of an artificial intelligence-powered instrument for the assessment of migraine risk in university students in Lebanon', *Computers in Human Behavior Reports,* 15, pp. 100453.

Taufique, Z., Zhu, B., Coppola, G., Shoaran, M. and Altaf, M.A.B. (2021) 'A low power multi-class migraine detection processor based on somatosensory evoked potentials', *IEEE Transactions on Circuits and Systems II: Express Briefs,* 68(5), pp. 1720–1724.

Temraz, M. and Keane, M.T. (2022) 'Solving the class imbalance problem using a counterfactual method for data augmentation', *Machine Learning with Applications,* 9, pp. 100375.

The Migraine Trust. (n.d.). *What is migraine?* [online] Available at: https://migrainetrust.org/understand-migraine/what-is-migraine/#page-section-6 [Accessed 2 Jun. 2024].

Torrente, A., Maccora, S., Prinzi, F., Alonge, P., Pilati, L., Lupica, A., Di Stefano, V., Camarda, C., Vitabile, S. and Brighina, F. (2024) 'The clinical relevance of artificial intelligence in migraine', *Brain Sciences,* 14(1), pp. 85.

Ur Rehman, A. and Belhaouari, S.B. (2021) 'Unsupervised outlier detection in multidimensional data', *Journal of Big Data,* 8(1), pp. 80.

www.kaggle.com. (n.d.). *Migraine Dataset*. [online] Available at: https://www.kaggle.com/datasets/ranzeet013/migraine-dataset [Accessed 15 Oct. 2024].

Conceptualizing Similarity Measurement in Data Marketplaces

Samrat Gupta[1,2,5]
samratg@iima.ac.in

Jana Peliova[2]
jana.peliova@euba.sk

Payel Sadhukhan[3]
p.sadhukhan.tmsl@ticollege.org

Pradeep Kumar[4]
pradeepkumar@iiml.ac.in

Polyxeni Vasilakopoulou[5]
polyxeni.vasilakopoulou@uia.no

Ilias Pappas[5,6]
ilpappas@ntnu.no

[1]*Indian Institute of Management Ahmedabad, Gujarat 380015, India*
[2]*University of Economics in Bratislava, Petržalka 85235, Slovakia*
[3]*Techno Main, Salt Lake, Kolkata, West Bengal 700091, India*
[4]*Indian Institute of Management Lucknow, U.P. 226013, India*
[5]*University of Agder, Kristiansand 4630, Norway*
[6]*Norwegian University of Science and Technology, Trondheim 7034, Norway*

Abstract

Data marketplaces have emerged to facilitate data trading through secure engagement and collaboration among actors, providing mechanisms similar to online marketplaces, such as connecting buyers with sellers and facilitating financial exchanges. The inherent complexity of data coupled with challenges related to marketplace integration hinder the realization of effective data trading mechanisms on data marketplaces. Moreover, the fragmentation of the data marketplace ecosystem necessitates data assets matching capabilities to enable the federation of different marketplaces. In this paper, we explore various similarity metrics for data assets and conceptualize how matching of data assets can be performed in a data marketplace. This can enhance tasks such as pricing advisory and revenue allocation on data marketplaces. This study is a step towards creating a reliable and equitable data trading environment.

Keywords: data marketplaces, data assets, similarity measurement, pricing data

1.0 Introduction

Data is inherently complex due to its non-perishable nature, ease of replication, broad user base, and possessing a combinatorial value that varies based on the buyer and the specific use case (Agarwal et al., 2019). This complex nature of data coupled with challenges such as pricing, data ownership protection, and integration of different marketplaces restricts our ability to comprehend the evolving landscape of data trading, the wide spectrum of data types, and the various business models and technologies involved (Mehta et al., 2021; Yuan et al., 2022).

These developments have led to data marketplaces in which actors engage and collaborate securely to search, consume, publish, or reuse data to stimulate innovation, create value and support new enterprises (Bergman et al., 2022; Yu and Zhang, 2017). Data marketplaces provide similar mechanisms as traditional electronic marketplaces such as matching supply and demand, infrastructure for creating sales contracts, and facilitating exchanges for payment and transportation of data assets (Azcoita and Laoutaris, 2022). These exchanges typically entail financial transactions of some kind, either through payments in fiat money (monthly subscription, fixed one-off payment, or pay-as-you-go) or in a cryptocurrency often managed by the platform. Data marketplaces can be classified as semi-private, wherein the platform must provide permission to any seller or buyer before allowing them to exchange data, or public, wherever any seller or buyer can exchange data (Azcoita and Laoutaris, 2022). Moreover, data marketplaces often handle metadata management, curation, and data categorization to assist purchasers in finding pertinent data products (Eichler et al., 2021).

One of the problems which data marketplaces face is how to allocate revenue fairly among sellers because data and digital goods can be replicated at zero marginal cost (Agarwal et al., 2019). The owner of data can sell multiple versions of the same data to different competitors to maximize revenue without adding any value to the precision of the prediction task (Cong et al., 2022). Secondly, if a retailer goes out to buy datasets to perform some prediction task, combination of some datasets may provide better prediction than other combination of datasets. For example, if traffic going to a mall needs to be predicted for inventory demand, different companies can be sellers of predictive data. Real-time traffic routing data of people driving through certain locations from Uber or Lyft or real-time foot-traffic data into stores such as KFC or Starbucks or sentiment data from social media indicating which type of brands are trending can be used for machine learning tasks.

In such situations, correlations between datasets offered by several companies may give rise to additional complications (Cong et al., 2022). Further, the fact that datasets can be easily replicated may become an underlying problem (Cong et al., 2022). As such, revenue allocation should be performed in a way such that companies are disincentivized from replicating their data. Such a notion of revenue allocation which is robust to replication is crucial in contemporary applications, such as battery cost attribution in smartphone apps and reward distribution among experts in a prediction market (Agarwal et al., 2019).

Data possesses a concept of pairwise similarity, which is not typically found in other goods (Agarwal et al., 2019). The problem of evaluating similarity within data assets has received little attention, despite the fact that data quality and retrieval have been the subject of much research. Current approaches lack a cohesive framework that considers many aspects of similarity across various data formats, instead concentrating mostly on semantic and syntactic comparisons. This gap presents challenges in fields such as data governance, data marketplaces, and data monetization.

The disincentivizing for replication of data assets on a data marketplace can be achieved by penalizing similar data i.e. assigning less revenue to sellers with similar data to others (Agarwal et al., 2019; Cong et al., 2022). To this end, similarity measurement can exploit the structure of data and be used to exponentially down-weight similar data or correlated features. Also, similarity measurement between data assets can be helpful in advising a price for a new data asset on the data marketplace. The actual trading may or may not occur at the advised price, however, it provides a relevant reference point for the seller to quote for their data asset.

This study proposes an approach to systematically measure similarity within data assets.

2.0 Data Marketplaces

Data marketplaces are intermediary platforms that connect data providers with potential buyers and facilitate the exchange of data (Azcoita and Laoutaris, 2022). The data providers source data from various internal and external sources converting it into data products (Eichler et al., 2021). These products are then available on the marketplace for consumers to browse, discover, negotiate, evaluate, and ultimately purchase and integrate into their business systems. Data marketplaces can operate globally, regionally, or within specific industries or technologies (Azcoita and Laoutaris, 2022). The products offered on these data marketplaces may include datasets, queries, software, machine learning models, reports, webinars, etc. Such exchanges on data marketplaces usually involve some kind of economic transaction. General-purpose data marketplaces such as AWS, DataRade or Advaneo trade all types of data, while niche ones focus on specific industries (like energy, automotive) or specific datatypes (such as spatio-temporal data) (Bergman et al., 2022; Azcoita and Laoutaris, 2022). Recently, there has been a notable trend towards real-time data streaming marketplaces leveraging the potential of IoT and those specializing in training machine learning models (Azcoita and Laoutaris, 2022).

The multitude of identified marketplaces, each with unique user interfaces, access protocols/APIs, and on-boarding procedures, makes it difficult for data providers to create a presence in each one of them and so reach the largest audience possible (Eichler et al., 2022). The present state of data marketplace ecosystem's fragmentation necessitates the creation of interoperability standards that will enable the seamless integration between different platforms resulting in federated data ecosystems (Nagel and Lycklama, 2022). One of the endeavours in this direction is the European Union's initiative to develop and implement Europe's first **F**ederated, decentralized, trusted d**A**ta **M**arketplace for **E**mbedded finance (FAME)[1] on which a consortium consisting of more than 30 partners from industry, academia, and technology sectors is actively collaborating.

The process of selling data assets on a data marketplace involves intrinsic and extrinsic information about data assets which can be used for assessing similarity among them. This similarity can aid in price recommendation of the data assets for buyers and fair revenue allocation among sellers.

3.0 Similarity Measurement within Data Assets

Several articles related to different types of similarity metrics exist in literature (Khojamli and Razmara, 2021; Gupta et al., 2016; Gupta and Deodhar, 2024). However, none of these articles focus on similarity metrics which can aid data marketplaces in tasks such as revenue allocation and pricing advisory as we have discussed above. To this end, based on our first-hand experience of engagement with the development of a FAME data marketplace and deep immersion with literature, we present the similarity metrics which can used to assess similarity among data assets in a data marketplace. Formally, a similarity metric is a function, $SM: R^T \times R^T \rightarrow$ [0, 1], that satisfies four properties (Agarwal et al., 2019; Gupta and Kumar, 2021) which are as follows:

- Limited Range: $0 \leq SM \leq 1$;
- Reflexive: $SM(X,Y) = 1 \; if \; and \; only \; if \; X = Y$;
- Symmetry: $SM(X,Y) = SM(Y,X)$;
- Triangle Inequality: $dSM(X,Y) + dSM(Y,Z) \geq dSM(X,Z) \; where \; dSM(X,Y) = 1 - SM(X,Y)$

[1] https://www.fame-horizon.eu/

3.1 Similarity Based on Logical Characteristics of Data Assets

Given the binary characteristics of data assets such as if the data asset is a sensor data or not, if there are copyrights associated with the data asset, and so on, generalization of Hamming distance can be used to compute similarity among data assets on these aspects. Suppose there are p logical characteristics of data assets available then similarity can be mathematically represented as follows:

$$SM_{logical} = \frac{\sum_{l=1}^{l=p} w_l\left(DA_i^l == DA_j^l\right)}{|p|} \tag{1}$$

where w_l indicates weight of a logical characteristic and DA^l represents a logical characteristic.

3.2 Similarity Based on Ordinal Characteristics of Data Assets

Some characteristics of data assets could be available as ordinal characteristics on a Likert scale such as how suitable is a data asset for a particular application, how credible is the source of data asset etc. Similarity computation based on these aspects can be performed using the generalization of cosine similarity. If there are q ordinal characteristics of data assets available, then following formulation can be used to compute similarity about these ordinal characteristics:

$$SM_{ordinal} = \frac{\sum_{o=1}^{o=q} w_o DA_i^o \times w_o DA_j^o)}{\sqrt{\sum_{o=1}^{o=q}(w_o DA_i^o)^2} \times \sqrt{\sum_{o=1}^{o=q}(w_o DA_j^o)^2}} \tag{2}$$

where w_o indicates weight of an ordinal characteristic and DA^o represents an ordinal characteristic.

3.3 Similarity Based on Numeric Characteristics of Data Assets

Moreover, the information about characteristics of data assets such as size of the data asset, resources required to assemble the data asset, and so on may be available in numeric form. If there are r such characteristics of data assets available, similarity computation among them can be performed by using the generalization of cosine similarity as follows:

$$SM_{numeric} = \frac{\sum_{n=1}^{n=r} w_n \log(DA_i^n) \times w_n \log(DA_j^n)}{\sqrt{\sum_{n=1}^{n=r}(w_n \log(DA_i^n))^2} \times \sqrt{\sum_{n=1}^{n=r}(w_n \log(DA_j^n))^2}} \tag{3}$$

where w_n indicates the weight of each numeric characteristic and DA^n represents an ordinal characteristic. In the above formulation, the numeric values are adjusted using logarithmic transformation before cosine similarity computation.

3.4 Similarity Based on Continuous Data Distribution within Data Assets

The data assets may be similar to some extent in terms of their underlying distribution. To consider the similarity based on data distribution the Inverse Hellinger distance (IHD) (Hellinger, 1909) or Total Variation Distance (TVD) (Verdú, 2014) can be used. Mathematically, Hellinger distance is represented as follows:

$$H(P,Q) = (\tfrac{1}{2}\sum_{x\in\chi}(\sqrt{P(dx)} - \sqrt{Q(dx)}\,)^2)^{1/2} \tag{4}$$

where P and Q denote two probability measures on a measure space χ and similarity between P and Q can be measured as follows:

$$SM_{IHD} = 1 - H(P,Q) \tag{5}$$

Sample based on Total Variation Distance: In line with above, the similarity using total variation distance between two distributions P and Q with support on measure space χ can be represented as follows:

$$SM_{TVD} = 1 - \tfrac{1}{2}\sum_{x\in\chi}|P(x) - Q(x)| \tag{6}$$

3.5 Similarity Based on Categorical Characteristics of Data Assets

The data associated with data assets may also be categorical. For example, if a customer service request is of high, medium, or low priority, gender of the users, continent/country to which users belong etc. The similarity between data assets with this kind of data can be measured through Normalized Mutual Information (NMI) (Sakumoto et al., 2024). Similarity measurement through NMI can be represented as follows:

$$SM_{NMI} = \frac{\sum_{x\in U}\sum_{x\in V} P(i,j) log\frac{P(i,j)}{P(i)P^{'}(j)}}{mean(\sum_{x\in U} P(i)\log\,(P(i),\sum_{x\in V} P'(j)\log\,(P'(j))} \tag{7}$$

where U and V are two categorical assignments of data between which similarity is being measured. $P(i) = {|U_i|}/{N}$ is the probability that an object picked at random from U falls into category U_i, $P'(j) = {|V_j|}/{N}$ and $P(i,j) = \frac{|U_i \cap V_j|}{N}$.

3.6 Similarity Based on Textual Data Associated with Data Assets

This is an era of generative AI which marks the abundance of human and AI generated text (Ooi et al., 2023). Due to this there has been an unprecedented surge in data assets containing text data (Jo, 2023). The text data could also be associated with non-textual data assets in the form of their description and/or title. Thus, similarity based on textual data can be computed between data assets using metrics based on lexical similarity or semantic similarity etc. (Furlan et al., 2013; Majumder et al., 2016; Islam and Inkpen, 2008). The lexical text similarity includes metrics such as cosine similarity, Jaccard similarity, Sørensen-Dice coefficient, and Levenshtein distance etc. (Majumder et al., 2016). Semantic text similarity includes approaches based on word or sentence embeddings, contextual language models, sentence transformers etc. (Islam and Inkpen, 2008). BERT (Bidirectional Encoder Representations from Transformers) enables text similarity computations by transforming text into dense vector representations that capture semantic meaning in context. By encoding two texts and calculating the cosine similarity between their embeddings, BERT effectively measures the similarity between sentences or documents, leveraging its contextual embeddings for precise comparison (Yang et al., 2023).

3.7 Similarity Based on Intrinsic Characteristics of Data Assets

A key aspect of a data asset is its usability for machine learning tasks (Agarwal et al., 2019). In a considerable number of scenarios, data is used to build a predictive model which will take decisions on unseen, upcoming instances (Gupta et al., 2016). To have good predictive power of the model, several intrinsic features of data need to be considered. For example, the distinctness of classes in the data and feature partitioning over space are favorable characteristics for machine learning tasks (Santos et al., 2023). The former aspect of class separability can be computed through a metric called *degOver* which quantifies the degree of overlap by finding non-overlapping and overlapping examples in a *k*-neighborhood (Santos et al., 2022).

$$degOver = \frac{n_{min_{over}} + n_{maj_{over}}}{n} \tag{8}$$

where $n_{min_{over}}$ and $n_{maj_{over}}$ are the number of overlapping examples of both classes and *n* is the total number of examples in the data space (Santos et al., 2022).

The later aspect of feature overlap can be assessed through measures such as maximum Fisher's discriminant ratio (F1) (López et al., 2013; Santos et al., 2022). For a feature f_i comprised in the dataset, the Fisher's discriminant ratio can be computed as follows:

$$F1 = \frac{1}{1+\max(d_{f_i})} \tag{9}$$

where

$$d_{f_i} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{10}$$

and μ_1, μ_2, σ_1^2, and σ_2^2 are the means and variances of class 1 and 2, respectively. The lower values of F1 indicate more overlap and vice versa.

Once these intrinsic characteristics of all the data assets are computed, cosine similarity metric or similarity using Euclidean distance can be used to calculate intrinsic similarity between the data assets (Khojamli and Razmara, 2021; Majumder et al., 2016; Islam and Inkpen, 2008).

4.0 Conceptualizing Similarity Measurement in Data Marketplaces

In this section, we present a conceptual framework for similarity measurement among data assets inspired from our involvement in a work package related to developing pricing and monetization schemes for data trading in the FAME data marketplace which is currently being developed under the European Union's Horizon research and innovation programme. The Pricing Advisory Tool (PAT) of FAME is based on the underlying logic of similarity measurement. FAME initially implements seven use cases from varying domains such as banking, transportation, and climate change. The data assets emanating from these use cases will be published and available for sale on the FAME platform. It is important to understand that historical trades for all data assets may not be available on the FAME platform. Hence, pricing advisory needs to consider the intrinsic and extrinsic characteristics of data assets based on which pricing recommendation could be performed.

The extrinsic characteristics of data assets (such as size of data asset, space needed for its storage, completeness of data asset, real vs. synthetic data asset etc.) could be known by administering a set of questions (such as *How suitable are data in the asset for a wide range of analyses and interpretations? -Response on Likert scale, Have these data been validated against independent sources or standards?- Response in Yes or No, How much resources were required to assemble and prepare the asset in question (in MD)?-Response in numeric*) through an API from the owner of data asset while publishing it. However, intrinsic characteristics such as ML-readiness of a data asset are challenging to know or infer. Class-overlap is one such characteristic of data from which its ML-readiness can be inferred (Vuttipittayamongkol et al., 2021; Santos et al., 2022; Santos et al., 2023). For instance, an efficient classification model can be built from a data asset consisting of fraudulent and authentic bank transactions with well-separated classes compared to the one with overlapping classes. Hence the price of the former should be similar to other data assets with less class overlap (while also considering the other intrinsic or extrinsic characteristics of data asset). However, deducing the class overlap in a data asset and subsequently using this information along with extrinsic characteristics of that data asset to compute its similarity with other data assets is a non-trivial task. Our approach is based on transformation of datapoints in a data asset to Euclidean Minimum Spanning Tree for measuring class overlap and subsequently synergistically using hamming, cosine and BERT model for matching the data assets (March et al., 2010; Eghbali and Tahvildari, 2018; Xia et al., 2015). This helps in price recommendation by considering similar data assets on a data marketplace. The below diagram (Figure 1) provides an overview of the pricing advising process, its actors including the similarity analysis tool, and their interactions.

As shown in Figure 1, the aforementioned notions can be implemented in a form of API that delivers price recommendations for assets and digital products based on user inputs. It aims to extract both subjective and objective factors that influence final pricing. Additionally, it offers price range estimates based on comparable assets with established historical prices, enabling end-users to see the likely price range within which an asset or digital product's price may vary. Such a tool employs a structured, two-step approach to precisely identify similar assets within a data marketplace's federated data assets catalogue. This approach supports focussed asset comparisons and improves the effectiveness of asset-related decision-making processes.

The initial phase uses Natural Language Processing (NLP) techniques to extract and preprocess textual data from asset titles and descriptions provided by users within the Asset Offering interface. This data is then analysed through clustering with BERT to systematically group assets into finely segmented subcategories. This classification

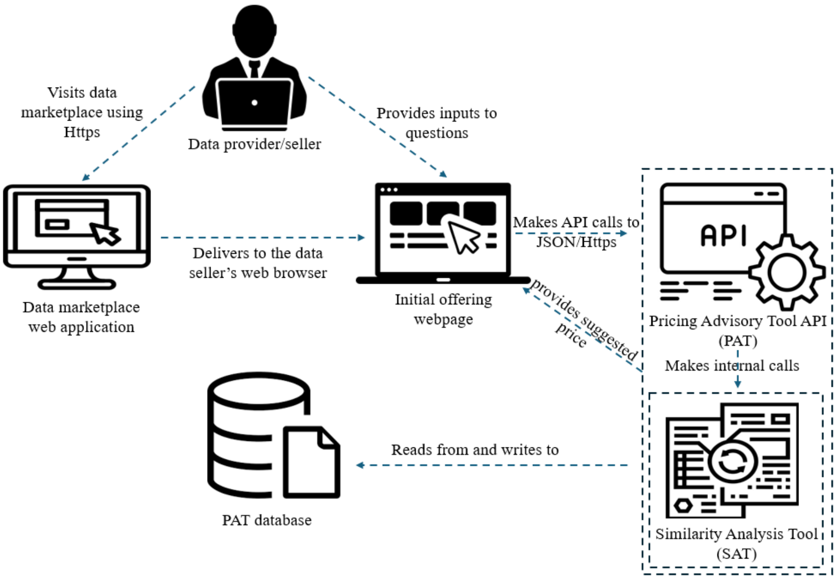

Visits data marketplace using Https
Data provider/seller
Provides inputs to questions
Data marketplace web application
Delivers to the data seller's web browser
Initial offering webpage
Makes API calls to JSON/Https
API
Pricing Advisory Tool API (PAT)
Provides suggested price
Makes internal calls
PAT database
Reads from and writes to
Similarity Analysis Tool (SAT)

Figure 1. **Component Level Architecture of pricing advisory in a data marketplace**

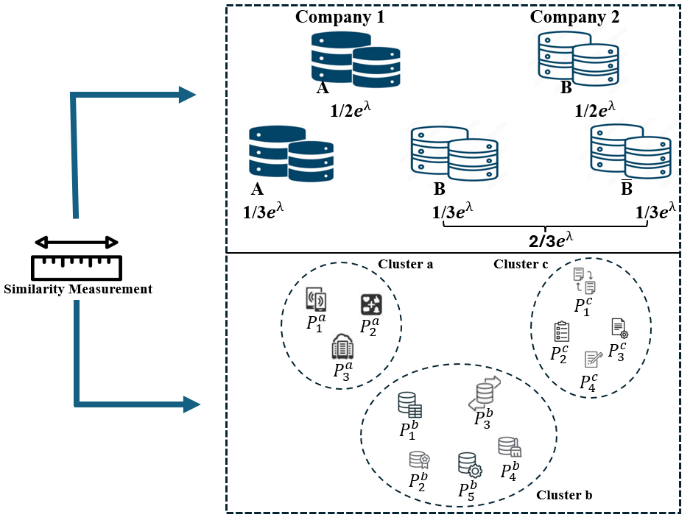
leverages semantic and contextual similarities in the asset metadata, enabling the formation of distinct asset clusters for further analysis.

Following the initial phase, the second step initiates a similarity analysis leveraging advanced algorithmic comparisons within each identified subgroup. This analysis employs vector space modelling and a composite similarity metric (based on the concepts such as cosine similarity, hamming distance, and minimum spanning tree) to quantitatively assess the closeness of each asset to the target asset under consideration. The objective is to pinpoint assets that exhibit the highest degrees of similarity to the target, based on the multidimensional feature space generated from the asset's descriptive metadata.

5.0 Matching of Data Assets Based on Similarity Measurement

Let's consider a data marketplace has two companies, one selling dataset A and other selling dataset B. Assuming A and B have same contribution to the accuracy of prediction task then half of the revenue will be allocated to the first company and half of the revenue will be allocated to the second company. Now, if the second company replicates its dataset B to $\overline{\mathbf{B}}$, thereby A, B and $\overline{\mathbf{B}}$ contributing to the prediction task, then each of these datasets get a revenue allocation of 1/3rd. Thus, even when there is no change in accuracy of the prediction task and aggregate payment remains the same, the first company receives a revenue of 1/3rd and the other receives a revenue of 2/3rd simply because of replicating its data (Agarwal et al., 2019). By penalizing comparable data assets, that is, giving sellers with identical data less money than others, the deterrence towards data duplication can be accomplished. To do this, similarity measurement can be used to exponentially down-weight associated characteristics or comparable data by making use of the structure of the data (as shown in Figure 2).

Additionally, determining how similar two data assets are to one another can assist determine how much to charge for a new data asset on the data marketplace. The purpose of similarity analysis in a data marketplace is to identify similar assets with a history of completed sales i.e., data asset must have past transactions. This analysis uses inputs such as responses to a questionnaire about each data asset and information from the asset listing, including the title, description, and business model chosen. An AI-based model (such as BERT) then analyses human-readable text from the title and asset



Revenue Allocation

Both datasets contribute equally to accuracy of prediction task (but are unique) and each gets equal revenue (λ=1)

Company 2 replicates data such that B and B' are same datasets. Similarity measurement downweights same data. Therefore, Company 2 makes less money (λ=2)

Pricing Advisory

Similarity measurement groups similar data assets which helps in advising the price of the focal data asset falling in a particular group

Figure 2. **An overview of the role of similarity measurement in revenue allocation and pricing advisory in data marketplaces**

description to create subsets of similar asset types. These subsets can subsequently be used to find relevant assets that match the focal data asset (as shown in Figure 2).

To measure intrinsic characteristics of data asset such as class overlap, first a dataset is transformed into a graph structure using the minimum spanning tree (MST) approach wherein the sum of edge weights is minimized (Sadhukhan and Gupta, 2025). MST is employed to capture the neighbourhood information and represent the intrinsic structure of heterogeneous datasets. We use Euclidean Minimum Spanning Tree (EMST) to form a connected network from a given set of datapoints (March et al., 2010). An EMST of a finite set of N points in a feature space, the points are connected in the most compact way by virtue of minimizing the total edge weights where connection weight (or the edge-weight) between any two points indicate their proximity. We define the EMST based class overlap in a dataset as follows:

Definition 1: Given a minimum spanning tree $G(V, E)$, in which vertex set V represents the data points and edge set E illustrates the proximity between data points, the class overlap in the dataset can be mathematically represented as:

$$EMST - O = \frac{w_{hom}}{w_{hom} + w_{het}} \times \frac{e_{het}}{e_{het} + e_{hom}} \tag{11}$$

where w_{hom} and w_{het} denote the average homogeneous edge weight and average heterogeneous edge weight respectively whereas w_{hom} and w_{het} denote the number of homogeneous edges and the number of heterogeneous edges respectively in the EMST of the given dataset. The value of $EMST - O$ ranges from 0 to 1.

This measure finds out the proportion of homogeneous edges and heterogeneous edges in the dataset, and the average weights of the homogeneous and the heterogeneous edges. The greater the number of heterogeneous edges, the more overlaps between the classes. Additionally, smaller homogeneous edge weights (shorter homogeneous edges) indicate compact class structures. A low-class overlap wherein homogeneous edges have lower weights than the heterogeneous edges is desirable as the classes therein are separated from each other and it is expected that the models trained on such a dataset would deliver better performance.

Once the information about class overlap in a data asset is computed using equation 11, the next step is to combine it with the extrinsic characteristics of the data asset. For this we use Definition 2.

Definition 2: Given the extrinsic characteristics of a data asset in the form of p logical values, q ordinal values and r numeric values, the intrinsic and extrinsic characteristics of a data asset can be represented as:

$$DA_i \; = \{EMST - O, l_1, l_2, \dots, l_p, o_1, o_2, \dots, o_q, o_1, o_2, \dots, o_q, n_1, n_2, \dots, n_r\} \qquad (12)$$

Therefore, depending on the number of data assets published on FAME platform, it will consider corresponding number of vectors as defined in equation 12 for each data asset to compute similarity among them.

Definition 3. For any two data assets DA_i, DA_j the similarity between them can be defined using a formulation that combines manhattan similarity, hamming similarity and cosine similarity. Manhattan similarity is used for deriving similarity between class overlap scores, Hamming distance is used for capturing similarity between binary characteristics of the data assets, cosine similarity is used for deriving similarity between ordinal characteristics of data assets, and numeric characteristics of data assets. However, due to scale of numeric characteristics, it is important to adjust the values using logarithmic transformation before cosine similarity computation. Mathematically the formulation can be expressed as follows:

$$Sim(DA_i, DA_j) = \frac{1}{4}\left(\left|DA_i^{EMST-O} - DA_j^{EMST-O}\right| + \frac{\sum_{l=1}^{l=p} w_l\left(DA_i^l == DA_j^l\right)}{|p|} \right. +$$

$$\left. \frac{\sum_{o=1}^{o=q} w_o DA_i^o \times w_o DA_j^o)}{\sqrt{\sum_{o=1}^{o=q}(w_o A_i^o)^2} \times \sqrt{\sum_{o=1}^{o=q}(w_o A_j^o)^2}} + \frac{\sum_{n=1}^{n=r} w_n \log\left(DA_i^n\right) \times w_n \log\left(DA_j^n\right)}{\sqrt{\sum_{n=1}^{n=q}(w_n \log\left(DA_i^n\right))^2} \times \sqrt{\sum_{n=1}^{n=q}(w_n \log\left(DA_j^n\right))^2}} \right) \qquad (13)$$

This mechanism can be implemented in real-time in a data marketplace and based on the searched keywords and queries asked to the data assets catalogue, the mean price of top 10 data assets can be recommended as price of the focal data asset. The pseudo code for matching of data assets based on aforementioned concepts and definitions is presented below.

Pseudo Code: Asset Matching in a Data Marketplace

```
Input:
  ${DA_1, DA_2, \ldots, DA_n} \in DA$ : A set of data assets hosted by the data marketplace
  $DA_f$: A focal data asset for which price is to be recommended
Output:
  ${DA_{f1}, DA_{f2}, \ldots, DA_{f10}}$: A set of 10 most similar data assets to the focal data asset $DA_f$
Begin:
  (1) For each data asset $DA_i$ , construct its EMST
  (2) For each data asset $DA_i$ , compute its class overlap EMST-O according to Definition 1
  (3) For each data asset $DA_i$ , merge its EMST-O value with the set of extrinsic characteristics according to Definition 2
  (4) Compute the similarity between each pair of data assets using Definition 3
  (5) Return and display the top 10 similar data assets in $DA$ in terms of their similarity with focal data asset $DA_f$
End
```

6.0 Discussion

This study investigates similarity measurement which can be useful for tasks such as revenue allocation and pricing advisory within data marketplaces. We underscore the potential of similarity metrics to address some of the inherent challenges faced by these marketplaces, such as data replication and equitable revenue distribution. For instance, when two companies contribute equally to a prediction task, the introduction of a duplicate dataset can distort revenue allocation (Agarwal et al., 2019). Similarity metrics can assist in pricing advisory, providing sellers with a reference point for setting prices based on the uniqueness and relevance of their data assets. Moreover, by utilizing a combination of similarity measurements discussed in section 3, we can ensure that replicated data receive a proportionately smaller share of the revenue, thus maintaining fairness and encouraging the originality of data contributions.

6.1 Theoretical Implications

This study offers a foundational conceptualization for evaluating relevance of data and valuation of data in a data marketplace setting by providing a methodology to bridge the gap between computational similarity measures and market-driven pricing strategies. The study has several theoretical implications for similarity measurement as it applies beyond recommendation systems and information retrieval. First, prior research on data similarity has primarily focused on enhancing recommendations, clustering, and retrieval accuracy (Khojamli & Razmara, 2021; Gupta & Deodhar, 2024), yet its role in economic decision-making within data marketplaces has remained underexplored. Second, this study is a step towards ascertaining the meaning and use of data by including hitherto unexplored attributes of data assets such as content relevance, data structure, format compatibility, and compliance requirements (Aaltonen et al., 2021). Third, this study contributes to the theories of data economics and information value by theorizing how the aforementioned similarity-based factors can be used in deriving value (Bonatti et al., 2024; Spiekermann and Korunovska, 2017). In doing so, this study also provides a foundation for further exploration of factors that influence market pricing and desirability of data assets. Fourth, this study implies how similarity measurement can serve as a mechanism for mitigating uncertainties by reducing information asymmetry between sellers and buyers. In summary, this study lays the foundation of how similarity measurement can advance both technical and commercial aspects of data trading, addressing limitations of prior models that either lack transparency or depend on arbitrary seller-defined values. We believe that future studies will empirically validate and refine the concepts of similarity measurement in data marketplaces presented in this study.

6.2 Practical Implications

The conceptualization of similarity measurement in data marketplaces enables both buyers and sellers to match demand and supply and enhance trust in data assets. By conceptualizing similarity measurement, this study provides actionable guidance for marketplace designers, data sellers, and buyers to effectively trade data assets based on their quality and relevance. First, similarity measurement can help in aligning search results with needs and characteristics of data buyers thereby improving discoverability of data assets within marketplaces. Second, similarity measurement enhances transparency in data quality thus boosting buyer’s confidence. Such transparency and trust are crucial for a thriving marketplace as it encourages repeat transactions (Wixom et al., 2023). Thirdly, similarity measurement helps data sellers in increasing the marketability of their data assets. Data sellers can position their data assets based on similarity to data assets in high demand, thereby enhancing visibility. Thus, similarity measurement can support dynamic pricing models and improve economic efficiency within a marketplace. Finally, similarity measurement helps organizational buyers to assess the interoperability of datasets while integrating external datasets with existing data systems across different formats. In summary, as data marketplaces grow in scale and complexity, the conceptualization of similarity measurement in this study will help in shaping an efficient data economy.

6.3 Limitations and Future Research Directions

However, our suggested approach for using similarity measures in data marketplaces is not without limitations. The complexity of calculating similarity metrics, especially for large and diverse datasets, poses a significant computational challenge. Additionally, the weight assignment in similarity measurements, which is crucial for their accuracy, can be subjective and may require domain-specific expertise. Another limitation is the potential variability in the types and quality of data assets across different marketplaces. Our proposed similarity metrics may need to be adapted or refined to accommodate these differences, which can vary widely depending on the specific characteristics of the data and the marketplace in question. Future research should focus on developing more efficient algorithms for calculating similarity metrics that can handle large-scale datasets without compromising accuracy. Additionally, studies should explore the dynamic adjustment of weight assignments in similarity measurements to better reflect the changing value and relevance of data assets over time. Further research could also investigate the integration of similarity metrics with advanced machine learning techniques to automate and refine the processes of revenue allocation and pricing advisory.

7.0 Conclusion

In conclusion, our study highlights the significant role that similarity measurement can play in enhancing the functionality of data marketplaces. By addressing issues such as data replication and providing pricing guidance, similarity metrics can contribute to a more equitable and efficient data trading environment. We believe that this study would pave the way for more integrated data marketplaces, ultimately supporting greater innovation and value creation in the data economy.

References

Aaltonen, A., Alaimo, C., & Kallinikos, J. (2021). The making of data commodities: Data analytics as an embedded process. Journal of Management Information Systems, 38(2), 401-429.

Agarwal, A., Dahleh, M. and Sarkar, T. (2019) A marketplace for data: An algorithmic solution. In Proceedings of the 2019 ACM Conference on Economics and Computation, pp. 701-726.

Azcoitia, S. A. and Laoutaris, N. (2022) A survey of data marketplaces and their business models. ACM SIGMOD Record, 51(3), 18-29.

Bergman, R., Abbas, A. E., Jung, S., Werker, C. and de Reuver, M. (2022) Business model archetypes for data marketplaces in the automotive industry: Contrasting business models of data marketplaces with varying ownership and orientation structures. Electronic Markets, 32(2), 747-765.

Bonatti, A., Dahleh, M., Horel, T., & Nouripour, A. (2024). Selling information in competitive environments. Journal of Economic Theory, 216, 105779.

Cong, Z., Luo, X., Pei, J., Zhu, F. and Zhang, Y. (2022) Data pricing in machine learning pipelines. Knowledge and Information Systems, 64(6), 1417-1455.

Eghbali, S., & Tahvildari, L. (2018). Fast cosine similarity search in binary space with angular multi-index hashing. IEEE Transactions on Knowledge and Data Engineering, 31(2), 329-342.

Eichler, R., Giebler, C., Gröger, C., Hoos, E., Schwarz, H. and Mitschang, B. (2021) Enterprise-wide metadata management: an industry case on the current state and challenges. In Business Information Systems (pp. 269-279)

Eichler, R., Gröger, C., Hoos, E., Schwarz, H. and Mitschang, B. (2022) From data asset to data product–the role of the data provider in the enterprise data marketplace. In Symposium and Summer School on Service-Oriented Computing, Cham: Springer International Publishing, pp. 119-138.

Furlan, B., Batanović, V., & Nikolić, B. (2013). Semantic similarity of short texts in languages with a deficient natural language processing support. Decision Support Systems, 55(3), 710-719.

Gupta, S. and Kumar, P. (2021) A constrained agglomerative clustering approach for unipartite and bipartite networks with application to credit networks. Information Sciences, 557, 332-354.

Gupta, S., & Deodhar, S. (2024). Understanding digitally enabled complex networks: a plural granulation-based hybrid community detection approach. Information Technology & People, 37(2), 919-943.

Gupta, S., Kumar, P. and Bhasker, B. (2016). A rough connectedness algorithm for mining communities in complex networks. In Big Data Analytics and Knowledge Discovery: 18th International Conference, DaWaK 2016, Porto, Portugal, September 6-8, 2016, Springer International Publishing. pp. 34-48

Gupta, S., Kumar, S. and Kumar, P. (2016): Evaluating the predictive power of an ensemble model for economic success of Indian movies. The Journal of Prediction Markets, 10(1), 30-52.

Hellinger, E. (1909): Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. Journal für die reine und angewandte Mathematik, 1909(136), 210-271.

Islam, A. and Inkpen, D. (2008): Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data (TKDD), 2(2), 1-25.

Jo, A. (2023): The promise and peril of generative AI. Nature, 614(1), 214-216.

Khojamli, H. and Razmara, J. (2021) Survey of similarity functions on neighborhood-based collaborative filtering. Expert Systems with Applications, 185, 115482.

Li, J., Sun, A., Han, J. and Li, C. (2020). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50-70.

López, V., Fernández, A., García, S., Palade, V. and Herrera, F. (2013): An insight into classifi-cation with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information sciences, 250, 113-141.

Majumder, G., Pakray, P., Gelbukh, A. and Pinto, D. (2016): Semantic textual similarity methods, tools, and applications: A survey. Computación y Sistemas, 20(4), 647-665.

March, W. B., Ram, P., & Gray, A. G. (2010). Fast Euclidean minimum spanning tree: algorithm, analysis, and applications. In Proceedings of the 16th ACM

SIGKDD international conference on Knowledge discovery and data mining, pp. 603-612.

Mehta, S., Dawande, M., Janakiraman, G. and Mookerjee, V. (2021) How to sell a data set? Pricing policies for data monetization. Information Systems Research, 32(4), 1281-1297.

Nagel, L. and Lycklama, D. (2022) How to build, run, and govern data spaces. In Designing data spaces: The ecosystem approach to competitive advantage. Cham: Springer International Publishing. pp. 17-28

Ooi, K. B., Tan, G. W. H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., ... & Wong, L. W. (2023). The potential of generative artificial intelligence across disciplines: Perspectives and future directions. Journal of Computer Information Systems, 1-32.

Sadhukhan, P., & Gupta, S. (2025). A graph theoretic approach to assess quality of data for classification task. Data & Knowledge Engineering, 102421.

Sakumoto, T., Hayashi, T., Sakaji, H. and Nonaka, H. (2024): Metadata-Based Clustering and Selection of Metadata Items for Similar Dataset Discovery and Data Combination Tasks. IEEE Access.

Santos, M. S., Abreu, P. H., Japkowicz, N., Fernández, A. and Santos, J. (2023): A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research. Information Fusion, 89, 228-253.

Santos, M. S., Abreu, P. H., Japkowicz, N., Fernández, A., Soares, C., Wilk, S. and Santos, J. (2022): On the joint-effect of class imbalance and overlap: a critical review. Artificial Intelligence Review, 55(8), 6207-6275.

Spiekermann, S., & Korunovska, J. (2017). Towards a value theory for personal data. Journal of Information Technology, 32(1), 62-84.

Verdú, S. (2014): Total variation distance and the distribution of relative information. In 2014 Information Theory and Applications Workshop (ITA), IEEE, pp. 1-3.

Vuttipittayamongkol, P., Elyan, E., & Petrovski, A. (2021). On the class overlap problem in imbalanced data classification. Knowledge-based systems, 212, 106631.

Wixom, B. H., Beath, C. M., & Owens, L. (2023). How to Have Better Strategy Conversations about Monetizing Data. MIT Sloan Management Review, 65(1), 1-5.

Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. Information sciences, 307, 39-52.

Yang, K., Lau, R. Y., & Abbasi, A. (2023). Getting personal: A deep learning artifact for text-based measurement of personality. Information Systems Research, 34(1), 194-222.

Yu, H., & Zhang, M. (2017). Data pricing strategy based on data quality. Computers & Industrial Engineering, 112, 1-10.

Yuan, N., Feng, H., Li, M., & Feng, N. (2022). Penetration or skimming? Pricing strategies for software platforms considering asymmetric cross-side network effects. Journal of the Association for Information Systems, 23(4), 966-998.

Acknowledgement
This research is supported by a funded project on FAME data marketplace under European Union's Horizon research and innovation programme (Grant Agreement No. : 101092639).

The Sharing Economy: How do the Affordances Influence the Continued Usage of Digital Platforms for Handyman Services in South African?

Phaswana M. Malatjie and Lisa F. Seymour
CITANDA, Department of Information Systems, University of Cape Town

Completed Research

Abstract

The sharing economy is gaining traction in South Africa, with platforms such as Uber, Bolt, HomePlus, Kandua and AirBnB leading the way. Some studies are even predicting that sharing economy services could significantly boost the global economy, contributing several billions of dollars. As a result, issues such as social exclusion in developing countries might be reduced due to the success of sharing economy services. This qualitative study follows an interpretive philosophy and inductive approach. The targeted audience was the general public who utilises sharing economy platforms that facilitates handyman services. Twenty-two interviews were analysed. The findings provide policymakers with insights on possible interventions that need to be done to align with the people's needs, concerns and preferences. Notably, the study found the affordance of increased inclusivity and equality and a contradicting barrier – increasing inequality. Additionally, the paper reveals a new gap (marketing) that is relevant, and actionable in South Africa.

Keywords: Affordances, Collaborative-consumerism, Sharing Economy

1.0 Introduction

The sharing economy (SE) is gaining traction in South Africa, with the adoption of platforms like Uber, Bolt, HomePlus, AirBnB and Kandua. It has become a catchphrase due to its global success (Malatjie & Seymour, 2023). As such, this paper defines it as a collaborative consumerism that promotes the temporary acquisition of goods and services mediated through digital platforms to create a pleasant living for everyone (Malatjie & Seymour, 2023; Belk, 2014). Researchers are turning their attention to the SE, as they believe this phenomenon has the potential to alleviate some of the social challenges such as the unemployment rate (Vallas & Schor, 2020).

South Africa, which has the highest unemployment rate in Africa (Statista, 2023), could potentially benefit from the growth of sharing economy services. By leveraging technology, these services can offer insights into the broader adoption of digital tools and internet applications in South Africa. But also provide new economic opportunities and help alleviate the country's pressing social issues. This can help in designing more effective and inclusive digital solutions which offer improved prospects on productivity and competitiveness. As such, the SE platforms are frequently considered for their contribution to sustainability (Aref, 2024; Frenken & Schor, 2017). Firms are now leveraging information systems (IS) to enable and ease the functioning of sharing platforms to contribute to society. They replicate and create new types of sharing marketplaces to connect people, organise transactions, and provide efficient services such as handyman services (Meng et al., 2022).

Having said that, the study defines handyman services as services such as plumbing, gardening, carpentry, locksmithing, and housekeeping. Consequently, there is a need to understand how users perceive the value, ease, and social impact of digital platforms for handyman services. These perceptions are essential for firms and policymakers because they impact behaviour and drive strategic decisions in the dynamic economic landscapes (Shao et al., 2023).

This paper provides insights on perceptions by answering the question: How do the affordances of the use of digital handyman platforms influence South Africans' perceptions? It summarises the related literature and provides findings based on the data that has been analysed thus far.

2.0 Literature Review

According to literature, researchers are debating the meaning of affordances, and it shows that there is no universally agreed-upon definition. This section will outline the definitions of the SE, and affordance in the context of the SE. It will show the types of the SE and show examples of the affordances of the SE. Relevant perceptions will also be briefed.

2.1 The Sharing Economy and Types of Sharing Economy

There are debates around the definition of the SE. Some researchers define the SE as a system of non-ownership transfer exchanges between three actors: consumers, platform workers and the platforms as the intermediator (Pelgander et al., 2022). Belk

(2014) defines collaborative consumption as people organising the purchase and distribution of a resource for a charge or other reward, emphasising that for the SE to be called "real sharing", no fees or compensations should be paid when sharing activities are performed. Thus, the SE is detailed in three categories:

The first is product-services system which is a form of sharing where users can share goods or services that belong to corporations or individuals (Malatjie & Seymour, 2023). This system enables the provision of collaborative products or services, including HomePlus, SweepSouth, and Kandua. Companies are permitted to provide products as services rather than selling them as physical items. Private goods can be shared or rented through these systems. These systems seek to offer the advantages of products without the requirement for ownership (Huang et al., 2024). The second collaborative lifestyle which allows individuals to share a common interest by leveraging intangible assets. This sharing primarily involves the exchange of money through crowdfunding platforms, time and skills (Huang et al., 2024; Malatjie & Seymour, 2023), or the sharing of one’s skills. The third is redistribution market which is the type of sharing that enables the exchange of goods, allowing the re-ownership of those goods. In other words, ownership is shared through gifting or selling of second-hand goods (Huang et al., 2024). Online platforms like Gumtree and BidorBuy serve as exemplary instances of redistribution markets within the South African context.

Following the summary of the SE categories, this study examined product-service systems offered globally and how the affordances influence the perception of South Africans towards utilising these systems. The product services chosen for this study are limited to handy services, such as household repair services, and house cleaning.

2.2 The Affordances of Sharing Economy Services

The behaviour of participants in the access-based consuming mode is currently understudied, both locally and globally. Researchers feel there are significant gaps that need to be filled (Govender, 2017; Lamberton & Rose, 2012). An in-depth analysis of participant behaviour in this context is required (Govender, 2017). Since the theme of this study is based on how affordances influence the continued usage of digital platforms for handyman services In South Africa, it is necessary to understand the phenomenon of affordances in the realm of IS. In the context of IS, literature defines affordance as the concept that establishes a relationship between users' behaviour and the functionalities of a system (Mesgari et al., 2023). Another definition is a concept that is presented as relations between the agencies of human actors and the material features of technology (Sutherland & Jarrahi; 2018). This study with the aid of literature defines the affordance in the context of technology as a concept that links the design or technology and how the technology is used by people (Faraj & Azad; 2013). The concept of affordance is relevant to this study and creates a link between the benefits obtained by platform workers and customers. The subsequent overview will delve into the affordances of the sharing economy services.

- **Affording a Safe Environment**. Sharing helps individuals to get to know the people in their neighbourhoods, which makes them safer. Friendships can be formed. Individuals who use the sharing economy service can help one another access resources quickly through recommendations. Consumers can save time by verifying the reputation of products online (Denisova, 2020).
- **Affording Convenience**. Globally, the SE sector is upending well-established businesses by providing consumers with easy access to resources at a reasonable cost, while also easing the financial and social responsibilities associated with ownership (Eckhardt & Bardhi, 2016).
- **Affording Cost Reductions**. Sharing allows consumers to divide the cost of owning high-quality, long-lasting items while also reducing the risk of loss, damage, or depreciation. Moreover, customers save money by accessing various goods without owning them; this lessens the price they would have spent if they had chosen ownership (Pulignano et al., 2024).
- **Affording Matchmaking**. Users (platform users—consumers and platform workers) are matched based on their needs or services they provide. The platform's matchmaking capabilities encourage users to generously donate items to those in need (Huang et al., 2024).
- **Affording Social Benefits**. The social dimension of perception considers the benefits to society, the community, and the environment. These benefits impact how users engage in SE activities (Shao et al., 2023).
- **Affording Trust**. Trust in the SE has some distinct characteristics. Trust needs to be established between consumers and platform workers, and between the platform and consumers. Trust is also facilitated through others' evaluation using reviews and ratings. It is created in the form of cyber trust (Pelgander et al., 2022).
- **Affording Sustainability**. Users are more likely to continue sharing after considering the sustainability viewpoint. By slowing down the rate at which natural resources run out, it preserves the environment by giving consumers easier access to pricey goods and services (Denisova, 2020). In contrast, sharing platforms can be regarded as fundamental pillars of sustainability (Al-Emran, M., & Griffy-Brown, 2023; Aref, 2024) as they promote resource efficiency by facilitating the shared utilisation of goods and services.
- **Affording Technological Adoption:** Understanding how these platforms leverage technology can offer insights into the broader adoption of digital tools and internet applications in South Africa. This can help in designing more effective and inclusive digital solutions. A technological adoption viewpoint, among the most prominent themes in IS research, is often considered when developing technologies like those described previously (Al-Emran, M & Griffy-Brown, 2023)**.**

2.2 Summary of Literature Review

Despite a significant quantity of academic research on SE, more studies are needed to understand how users perceive the impact of SE platforms which is essential for businesses, policymakers and researchers. Social exclusion is a pressing issue in South Africa and other African countries. Given the significant impact of social exclusion, IS researchers have opportunities to investigate potential solutions that are tailored to reducing unemployment and related issues in South Africa. There is also a need to identify actionable gaps to promote inclusivity, equality and economic growth. The literature assisted in creating the groundwork for this study. Table 1 presents the relevant literature that aids in answering the research question of this paper.

Themes	Findings	Literature source
Safe environment	- Sharing platforms provide users with insurance for safety. - Communities for fostering safe interactions on the platform.	Ahmadi, 2024; Malatjie & Seymour, 2023;
Convenience/Technology	- People are matched the technology and end up gaining access to services and good - Sharing platforms are disruptive and provide sustainability.	Ahmadi, 2024; Al-Emran, M &Griffy-Brown, 2023; Aref, 2024; Eckhardt & Bardhi, 2016; Huang et al., 2024
Cost reduction	- Individuals are able to save money by accessing services.	Pulignano et al., 2024; Valodia, 2023; University of Cambridge; n.d.
Social benefits	- Individuals can establish trustworthy connections on the platform.	Pelgander et al., 2022; Malatjie & Seymour, 2023
Definitions	- The definition of sharing economy - The definition of the affordances	Belk, 2014; Faraj & Azad; 2013; Mesgari et al., 2023; Sutherland & Jarrahi; 2018

Table 1. Thematic Matrix

3.0 Research Method

This study's objective is to understand how the affordances influence the continued usage of digital platforms for handyman services in South Africa. As such, this study adopted subjectivism as the ontology subjectivism as the ontology because participants' views and subsequent behaviours are what give rise to social phenomena (Saunders et al., 2019). The epistemology stance adopted to understand South

Africans' lived experience from their perspective is interpretivism. Furthermore, interpretivism asserts that each observer has a unique view and interpretation of reality (Saunders et al., 2019).

This qualitative paper is based on 22 interviews that were analysed and interpreted to answer the question: How do the affordances of the use of digital handyman platforms influence South Africans' perceptions? The experiences were used to better understand the affordances that influence South Africans' perceptions. The approach of the study was inductive because the goal was to explore new areas in the context of South Africa.

The targeted group is based on South Africans who are users of digital platforms for handyman services such as HomePlus and SweepSouth. Furthermore, the targeted group also includes people who are at trade schools and are interested in joining digital platforms for handyman services. To secure the participants, the study was guided by non-probability sampling techniques called Heterogeneous purposive sampling and snowballing sampling. Some of the users are platform workers who supply services on SE platforms, some are consumers who require services on SE platforms, and others are potential users.

The participants come from diverse regions of South Africa, representing a wide range of backgrounds. Their contributions offer valuable insights at various levels which could aid in developing digital solutions that can help in bridging the gap of inequalities in South Africa. Four participants are platform workers with tertiary qualifications residing in metropolitan cities. Thirteen participants are active platform consumers with tertiary qualifications and are employed in corporate companies, also residing in metropolitan cities. Two participants are final-year students at Technical and Vocational Education and Training (TVET) colleges pursuing artisan qualifications, and they are potential platform workers living in smaller cities. Three participants are facilitators at a TVET college with tertiary qualifications and potential platform workers residing in smaller cities. Table 2 provides a detailed demographic breakdown of the participants involved.

Participant IDs	Roles	Region	Stage
PAD001, PAD002, PAD003, PAD004, PAE005	Platform Consumer	Metropolitan city	Pilot list
UAD007, UAD008, UBD009, UAE015, UAD018, UAD020, UAE021, UAE022	Platform Consumer	Metropolitan city	Implementation list
WBD006, WBE016, WCE017, WCE019	Platform Worker	Metropolitan city	Implementation list
ZBE010, ZBE011, XF012, XF013, ZBE011	Potential Users	Small City	Implementation list

Table 2. Demographics of Participants

Data was collected using an interview instrument protocol and then analysed through inductive thematic analysis. The pilot group, indicated in Table 2, was involved in testing the research instrument**.** Microsoft Word Online was utilised for transcribing audio from in-person interviews, while Microsoft Teams transcribed online interviews. Inductive coding was used to analyse data using the Nvivo tool. Given the differing perspectives of platform workers, platform consumers, and potential platform workers, two sets of interview questions were developed. Platform workers were asked questions specific to their experiences and views of the handyman platforms. Likewise, consumers were asked questions based on their experiences and views. Potential workers were given questions tailored for platform workers but only needed to state what they thought would be beneficial to them.

To ensure ethical data collection, we informed participants that the study was approved by the Research Ethics Committee. Before accepting the interview, we asked participants to read and sign the consent form. Furthermore, we ensured that their identity would not be disclosed, and we used pseudonyms.

4.0 Findings and Discussion

This section presents the findings based on the data collected and analysed to answer the question: How do the affordances influence the continued usage of digital platforms for handyman services in South Africa? The affordances that influence the continued usage of digital platforms for handyman services are illustrated in Table 3 and are now discussed. The roles provided in Table 3 represents the views of a particular group of participants that were interviewed.

Master themes	Sub-themes	Roles
Affording inclusivity and equality	Affording inclusivity and equality	Consumers and Platform workers
Affording financial benefits	Earning extra income; Saving money; Ensuring temporary work	Consumers, platform workers and Potential workers
Affording support	Facilitating administrative functions; Facilitating continued improvement	Consumers and platform workers
Affording protection	Offering insurance coverage; Affording background checks	Consumers and platform workers
Affording trust	Affording trust	Consumers and platform workers

Table 3. The affordances that influence the continued usage of digital platforms for handyman services

Data based on the affordance of digital platforms for handyman services show that South Africans perceive digital platforms for handyman services as useful because the platforms afford them various opportunities that will now be discussed. The inductively derived model depicting an overview of the five affordances influencing people's views towards the continued usage of digital platforms for handyman services can be seen in Figure 1. The model shows the relationship between the affordances that are under investigation.

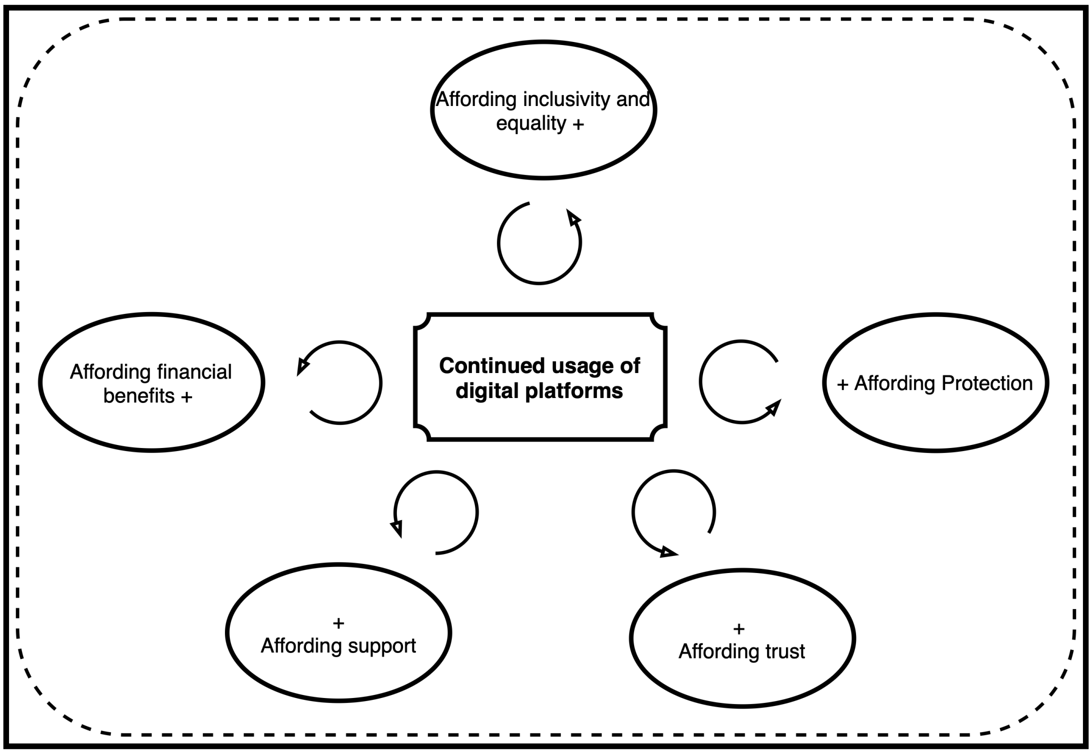

Affording inclusivity and equality +
Affording financial benefits +
Continued usage of digital platforms
+ Affording Protection
+ Affording support
+ Affording trust

Figure 1. **Explanatory conceptual model of affordances influencing the continued usage of digital platforms.**

4.1 Affording Inclusivity and Equality

Inclusivity typically talks about creating an environment where everyone feels welcomed, valued and able to contribute (University of Cambridge, n.d.), especially in countries like South Africa which are struggling with inequality (Valodia, 2023). Our study agrees with the literature as it found that sharing platforms have created a space that allows anyone to participate as a platform worker or consumer without being discriminated against. South Africans also perceive the platforms to reduce inequality and increase inclusivity. Women can participate in a male-dominated industry and the unemployed youth have access to temporary jobs. Platform workers believe that they can now offer their services in metropolitan cities without applying for jobs. As such, they believe by continuing to use the platform, they would be able to reduce inequality in South Africa. Furthermore, this study also found a contrary view to existing literature, suggesting that sharing platforms are not targeted to individuals relocating from less affluent areas. Here is the corroborating data

"Women in particular on the platform doing deliveries…which is not a typical thing because it's been a male-dominated industry." (PAD002)

"I don't think it's well marketed to lower LSM [living standard measure]" (UAE021)

“If you take the same concept and implement it in villages or small towns, you will create some economic upliftment in that village” (PAD003).

“Our government is struggling to put or is struggling to place people in consistent work. I think much like the way Uber opened up the transport [market], this would allow a lot of people to have options” (PAD002).

Despite the aforementioned challenges, South Africans believe that these affordances compel them to continue using the platforms. This is also evident in Figure 1, which suggests that SE platforms will remain in use as long as they continue to reduce inequalities and provide more opportunities for participation on the platform and in the economy.

4.2 Affording Financial Benefits

Financially, the benefits of using sharing platforms stem from consumers saving money due to more affordable services offered on the platforms, and platform workers gaining access to temporary work which either leads to them earning extra income or simply earning an income. Potential workers believe that sharing platforms could help them earn an income as the services they could provide on the platforms are fewer and have a lot of opportunities. The literature agrees with this finding, highlighting that platform workers can earn extra income by providing services on a platform (Pulignano et al., 2024). Here is the corroborating data:

“I will make money because the welders are few and there is plenty of jobs for welding” (XF013)

“The more I think I will use this app. The more people will want me to work for them, so I think I will make a lot of money” (ZBE011)

Furthermore, as the number of individuals earning an income or an additional income increases, they will likely continue utilising the platforms, thereby contributing to a reduction in income inequality.

4.3 Affording Support

This study defines affording support as the ability to afford both consumers and platform workers technical support such as feedback loops and administrative functions. The platforms offer support such as marketing, and invoicing. Furthermore, the platforms evaluate both consumers and platform workers to ensure that only trustworthy people use the services. Email accounts or forums are created for users to raise their technical issues. This finding suggests a favourable impact on the continued usage of digital platforms for handyman services. Existing literature contradicts this finding, emphasising that the platform merely facilitates connections between consumers and platform workers. Any services agreed upon by both parties are limited to the two parties, and the platform is not liable for any third-party issues (HomePlus, n.d.). Here is the corroborating data:

"It also assists us in marketing our product and services." (WCE017)

"The application's administrative function ensures that revenues collected by service providers are well documented in simplicity and easily accessible when one needs to file in for taxes with our government department SARS [South African Revenue Services]." (ZBE010)

Furthermore, the platform's user support can encourage more individuals to adopt the technology, leading to an increase in the continued usage of sharing platforms in South Africa.

4.4 Affording Protection

Sharing platforms encourage a community-based model with communal security and Botho/Ubuntu – to act with kindness and generosity towards others. Users undergo rigorous background checks before joining the platform. The idea is that vetting users helps create a trusted and safe community. Additionally, platforms offer users insurance coverage to illuminate the fear of users worrying about their items getting damaged. Here is the corroborating data:

"Before you join their platform… you go through a series of certification vetting, police clearance and they also need to check your background" (WCE017)

"The platform actually does offer insurance but then it comes at a cost" (UAD020)

When users perceive a sense of protection and safety, they are more motivated to continue utilising the platform. This, in turn, attracts a larger user base, leading to an increase in the platform's continued usage.

4.5 Affording Trust

Another affordance that emerged is perceived trust. People tend to trust services provided on popular or well-known platforms within their communities. This encourages them to continue using digital platforms for handyman services. Furthermore, consumers believe that by accessing handy services once on the platforms they already trust, their confidence and trust in the platform workers are boosted. This is because they can even use the platforms to access platform workers' ratings and reviews on them. The literature supports this finding, emphasising that users build trust within communities by utilising features on the platforms, such as rating services (Sutherland & Jarrahi, 2018). Here is corroborating data:

"Once you work with the person one time…I'm gonna call him directly" (PAD001)

"I'm not comfortable with strangers, but because they got to be reviewed via the handyman service platform, I'm a little more comfortable" (PAD004)

When users establish genuine connections within their communities, trust is fostered, which subsequently encourages users to persist in utilising the platform. Simultaneously, new individuals are drawn to the platform.

5.0 Conclusion and Recommendations

In summary, the growth and integration of sharing economy platforms in South Africa presents a unique opportunity to find potential solutions that could minimise economic disparities and provide more accessible services. However, the successful implementation and adoption of these platforms face significant challenges. This study explored the affordances of sharing economy platforms to understand how the affordances influence the continued usage of digital platforms for handyman services. This study identified the affordances of digital platforms for handyman services in South Africa. It discovered core affordances that are unique to emerging markets like South Africa, such as inclusivity and equality. Affordances such as access to opportunities reveal that South Africans perceive the platform to be useful and inclusive. This paper does concede, though, that the sample size may be too small to draw broad conclusions about all South Africans. The paper also acknowledges the need to comprehend how those who do not have access to sharing platforms might be exposed to or incorporated into them.

Lastly, studying the affordances of sharing platforms in IS studies is crucial because understanding these affordances aids in designing intuitive and user-friendly platforms. This often leads to a better user experience. South Africa presents unique challenges and opportunities present in the region such as regulatory environment, and cultural and social context. Researchers can investigate features that are tailored for South Africa which encourage user adoption and sustained engagement. This is significant for the growth of sharing platforms. Furthermore, IS researchers could compare the affordances and barriers of sharing platforms in South Africa to understand how they influence the assimilation of these platforms.

References

Ahmadi, M. (2024). *Enhancing shared living experiences through user-centered design: The design of a roommate matching platform* (Masters).

Al-Emran, M., and Griffy-Brown, C. (2023). The role of technology adoption in sustainable development: Overview, opportunities, challenges, and future research agendas. *Technology in Society*, *73*, 102240.

Aref, M. (2024). Sharing economy from the sustainable development goals perspective: a path to global prosperity, *Journal of Internet and Digital Economics, 4*(2), 116-138.

Belk, R. (2014). You are what you can access: Sharing and collaborative consumption online. *Journal of Business Research, 67*(8), 1595-1600.

Chua, E. L., Chiu, J. L., & Bool, N. C. (2019). Sharing economy: An analysis of Airbnb Business Model and the factors that influence consumer adoption. *Review of Integrative Business and Economics Research, 8*(2), 19-37.

Denisova, M. (2020). *Factors influencing success in sharing economy for consumer goods sector* Available from DSpace.

Eckhardt, G. M., & Bardhi, F. (2016). The relationship between access practices and economic systems. *Journal of the Association for Consumer Research, 1*(2), 210-225.

Faraj, S., & Azad, B. (2013). The materiality of technology: An affordance perspective. *Materiality and Organizing: Social Interaction in a Technological World, 237*, 258.

Frenken, K., & Schor, J. (2017). Putting the sharing economy into perspective. *A Research Agenda for Sustainable Consumption Governance, 23*, 3-10.

Govender, K. (2017). *The shift from ownership to access in South Africa: the shared economy* (Masters). Available from WIReDSpace. (2019-03-12T08:20:35Z).

HomePlus. (n.d.). *Home+ Terms and Conditions*. https://www.homeplus.africa/terms-and-conditions.

Huang, Y., Lin, C., & Wang, T. (2024). Benefits of Give Circle: Exploring the impact of collaborative redistribution platforms on user willingness to donate to charity and tendency towards consumer minimalism. *Computers in Human Behavior Reports, 14*, 100421.

Lamberton, C. P., & Rose, R. L. (2012). When is ours better than mine? A framework for understanding and altering participation in commercial sharing systems. *Journal of Marketing, 76*(4), 109-125.

Malatjie, P., & Seymour, L. F. (2023). *The Sharing Economy: Understanding the Affordances of and Barriers to the use of Digital Platforms*. Paper presented at the 9th African Conference on Information Systems and Technology, 22.

Meng, T., Ng, E., & Tan, B. (2022). Digital attrition: The negative implications of the sharing economy for the digital options of incumbent firms. *Information Systems Journal, 32*(5), 1005-1033.

Mesgari, M., Mohajeri, K., & Azad, B. (2023). Affordances and Information Systems Research: Taking Stock and Moving Forward. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, *54*(2), 29–52.

Pelgander, L., Öberg, C., & Barkenäs, L. (2022). Trust and the sharing economy. *Digital Business, 2*(2).

Pulignano, V., Muszyński, K., & Tapia, M. (2024). Variations of Freelancers' 'Effort-Bargain' Experiences in Platform Work. The Role of Skills. *ILR Review, 77*(6), 742-769.

Saunders, M. N. K., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students* (Eighth ed.). Pearson.

Shao, X., Jiménez, A., Lee, J. Y., & Taras, V. (2023). The impact of the perceived value of the sharing economy on consumer usage behavior: evidence from shared mobility in China. *Asian Business & Management, 22*(5), 1962-2003.

Statista. (2023). *Unemployment rate in Africa by country.* https://www.statista.com/statistics/1286939/unemployment-rate-in-africa-by-country

Sutherland, W., & Jarrahi, M. H. (2018). The sharing economy and digital platforms: A review and research agenda. *International Journal of Information Management, 43*, 328-341.

University of Cambridge. (n.d.). *What is equality, diversity, and inclusivity?* Cam.ac.uk. https://www.pdn.cam.ac.uk/intranet/equality-diversity-and-inclusion/what-equality-diversity-and-inclusivity

Vallas, S., & Schor, J. B. (2020). What do platforms do? Understanding the gig economy. *Annual Review of Sociology, 46*(1), 273-294.

Valodia, I. (2023). *South Africa can't crack the inequality curse. Why, and what can be done*. https://www.wits.ac.za/news/latest-news/opinion/2023/2023-09/south-africa-cant-crack-the-inequality-curse-why-and-what-can-be-done.html

Towards Sustainable AI Development: Challenges, Opportunities, and the Sustainable AI Development Card

Donghyeok Lee (CeADAR - Ireland's Centre for AI, University College Dublin), Yilin Li (CeADAR - Ireland's Centre for AI, University College Dublin), Junke Xu (CeADAR - Ireland's Centre for AI, University College Dublin), Christina Todorova (CeADAR - Ireland's Centre for AI, University College Dublin) and Alireza Dehghani (CeADAR - Ireland's Centre for AI, University College Dublin)

Completed Research

Abstract

With the expansion of AI systems, both in scale and complexity, it becomes ever more urgent to bridge the gap between advancement and environmental, social, and economic sustainability. In our desire to offer a practical tool for fostering sustainable AI development, we introduce an overall framework in guiding developers in adopting sustainable practices throughout the AI development lifecycle: the Sustainable AI Development Card. From data management to model training and deployment, this framework offers practical strategies in the direction of energy consumption minimisation and carbon emissions reduction, each with social equity promotion. Interposed upon recent advances in energy-efficient algorithms, hardware optimisation, and ethical AI deployment, the paper concentrates on detailing specific strategies that make alignment of AI performance with attainable sustainability objectives possible. The above shows how developers can minimise the sustainable risks of AI systems with no significant reduction in accuracy or scalability.

Keywords: Sustainable AI, Energy-efficient algorithms, AI development life cycle, Carbon footprint, Model optimisation, Ethical AI

1 Introduction

Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs (World Commission on Environment and Development, 1987). It includes environmental, social, and economic dimensions, all of which are essential to ensure the long-term viability of technological and industrial endeavour. In the field of artificial intelligence (AI), since the rising environmental, social, and economic impacts of large-scale AI systems, sustainability has taken on growing importance. As shown in Figure 1, the exponential growth in AI models, data, and infrastructure with increased computing power has led to significant energy consumption and increased carbon emissions, making sustainability an essential consideration for AI developers and researchers. To address these challenges, the concept of sustainable AI has emerged.

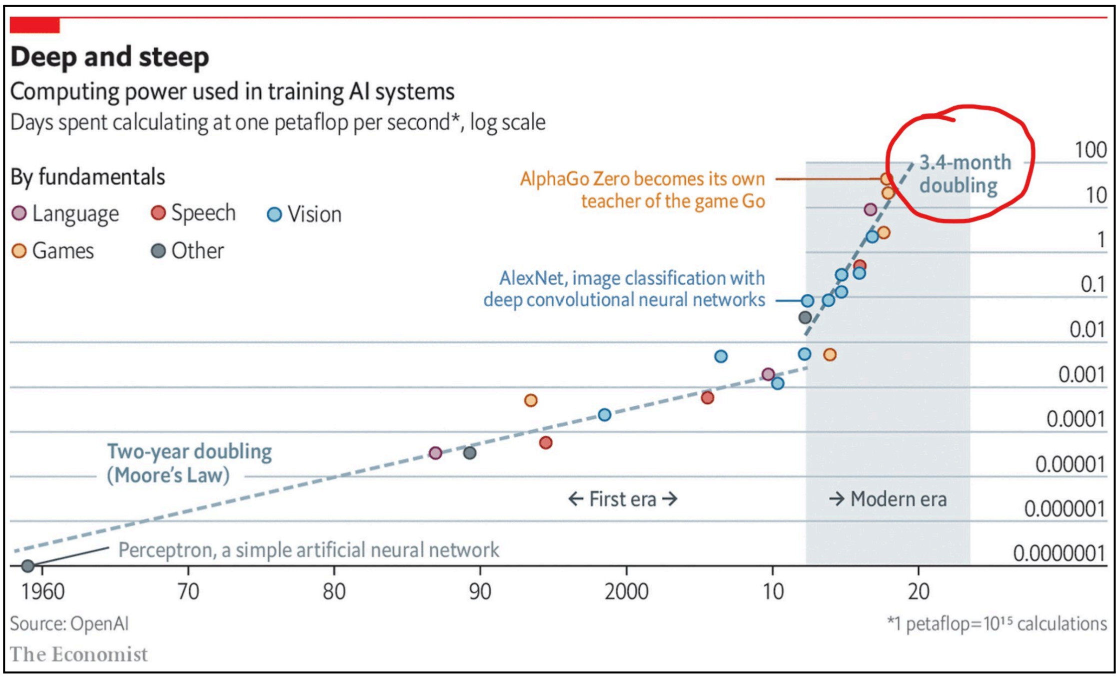


Figure 1. Computer power used in training AI systems has exponentially increased in the era of deep learning (*The Economist*, 2018)**.**

It involves designing AI systems that are economically feasible, socially responsible, and energy-efficient over the long term. On the environmental aspect, sustainable AI focuses on reducing the carbon footprint of AI systems, particularly during the training and inference stages, which are often the most resource-intensive. Energy savings of up to 115% have been reported in a study where AI models are optimised for efficiency, with savings of over 50% being common (Verdecchia, Sallou and Cruz, 2023). Such improvements are achieved through techniques like hyperparameter tuning, model pruning, and using energy-efficient hardware such as Tensor Processing Units (TPUs). By optimising these processes, developers can significantly reduce energy consumption, and this could minimise the impact on the environment. And also, social sustainability in AI refers to promoting fairness, transparency, and inclusivity by designing and deploying AI systems. As AI systems have made decisions in a wide range of important areas such as healthcare, finance, and criminal justice, it is crucial to ensure that these systems do not have to reinforce bias or worsen social inequality. Additionally, the economic dimension of sustainability in AI emphasises creating systems that are both cost-effective and scalable over the long term. High computational costs can create barriers to the widespread adoption of AI technologies, particularly for smaller developers and organisations with limited resources (Wu *et al.*, 2022). By improving energy efficiency and optimising resource use, sustainable AI practices help to break down these barriers, making advanced technologies more accessible to a broader range of users. Achieving sustainability in AI development is essential for coordinating innovation with environmental responsibility, social equity, and economic efficiency.

There is growing concern about the environmental impact of AI development and implementation. As AI models increase in scale and complexity, so do their computational demands, leading to significant increases in energy consumption and CO2 emissions. A number of strategies categorised under "Green-in AI" have been suggested to tackle these concerns, as indicated in their research. Important strategies encompass algorithm optimisation, hardware optimisation, data centre optimisation, and practical scaling factor reductions (Verdecchia, Sallou and Cruz, 2023). Algorithm optimisation is centred around creating models that demand reduced computational resources, achieved through sparse training, quantisation, and pruning techniques. Hardware optimisation includes the use of energy-efficient processors such as TPUs and the effective utilisation of parallelisation to reduce both execution time and energy consumption. Likewise, data centre optimisation seeks to decrease the environmental impact of AI by effectively managing server loads, enhancing cooling systems, and strategically placing data centres in areas with access to low-carbon energy sources. These technological advancements are significant in the effort to diminish the environmental impact of AI, emphasising the necessity for guidelines and resources that enable individual developers and small teams to engage in sustainable AI practices. Although algorithm optimisation makes the topic of sustainability more approachable for smaller organisations alike, the knowledge gap of how this works at a developer level is still a sustainability threat, requiring practical guidelines or checklists for technologists and developers. Our research suggests the development of a Sustainable AI Development Card, which serves as a checklist for developers to make environmentally, socially, and economically conscious decisions throughout the AI development process. We aim to provide the AI community with the necessary tools to collectively mitigate the sustainability risks of AI systems.

2 Related Literature

2.1 Overview of Sustainable AI

In order to make sure that AI systems are developed and implemented in accordance with more general sustainability objectives, sustainable AI is a multidisciplinary effort combining technology, ethics, and environmental science. Training a single large model like GPT-3 has been shown to produce emissions equivalent to driving a car for hundreds of thousands of miles (Patterson *et al.*, 2021). To mitigate these environmental impacts, the concept of Green AI (Schwartz *et al.*, 2020) has emerged, focusing on the development of AI systems that balance performance with energy efficiency. This includes techniques such as model pruning, quantisation, and distillation, which streamline model architectures and reduce their computational demands, effectively lowering energy consumption without sacrificing accuracy. Hardware advancements, designed to process machine learning workloads more efficiently, have led to substantial reductions in the operational carbon footprint of AI systems (Wu *et al.*, 2022). Additionally, recent research has also explored the optimisation of data management within AI pipelines as a key factor in sustainability. Techniques such as smart data sampling, data reduction, and dimensionality reduction are now being implemented to reduce the amount of data processed and stored, thus decreasing both the energy required for data handling and the overall computational footprint. In addition to addressing environmental sustainability, the social implications of AI deployment remain a key focus of sustainable AI efforts. As AI systems increasingly influence high-stakes decisions in areas such as healthcare, criminal justice, and employment, there is a growing concern about the fairness, transparency, and accountability of these systems. Recent advancements include the development of tools such as Model Cards (Mitchell *et al.*, 2019), which provide standardised documentation of AI models, detailing their intended use, potential biases, and performance across different demographic groups. On the economic front, energy-efficient AI infrastructures are critical in democratising access to AI technologies. The cost of developing and maintaining large-scale AI models continues to be a barrier for smaller organisations and academic researchers. Sustainable AI advocates for reducing these barriers through optimised energy usage, making AI systems more cost-effective and accessible to a wider range of developers (Verdecchia, Sallou and Cruz, 2023).

2.2 Use Cases Applying Sustainable AI Principles

Reducing the environmental impact of AI models has become a crucial concern. Research shows that AI systems can be made more sustainable without performance loss. One of the key strategies is fine-tuning pre-trained models, where large language models are adapted for specific tasks without retraining from scratch (Devlin *et al.*, 2019). This method saves computational resources and reduces the carbon footprint associated with training large models from scratch. In addition to fine-tuning, model compression techniques such as quantisation and structured pruning have proven highly effective in lowering AI's computational demands (Han, Mao and Dally, 2016) (Jacob *et al.*, 2017). By reducing the precision of neural network weights, methods like Zero Alignment ensure that compressed models still perform well while consuming less energy. For instance, quantising BERT models to INT8 precision maintained accuracy while significantly lowering the computational requirements (Zafrir *et al.*, 2019). Similarly, INT4 and INT8 quantisation was applied to models like BERT, Wav2Vec2.0, and ViT, demonstrating that these methods reduce training time and hardware utilisation, contributing to sustainable AI use (Wang *et al.*, 2022). These compression techniques are widely applied in NLP, speech recognition, and image classification tasks, providing sustainable alternatives for deploying AI models at scale (Shen *et al.*, 2019) (Touvron *et al.*, 2021). Another important approach is decentralised AI systems, such as Federated Learning. This method reduces energy consumption by distributing the computational load across edge devices, allowing data to be processed locally rather than in centralised data centres (McMahan *et al.*, 2023). Such systems significantly reduce the amount of data transmitted to the cloud, cutting down network and computational costs. Federated learning has been applied in scenarios like mobile devices and IoT, where decentralised computation is more resource-efficient than relying on energy-intensive centralised infrastructure (Kairouz *et al.*, 2021) (Li *et al.*, 2019). By enabling on-device learning, federated learning not only enhances privacy but also promotes the sustainable use of AI in resource-constrained environments (Yang *et al.*, 2019).

2.3 Previous Works and Research Gaps

The growing presence of AI technologies across different industries has prompted a heightened focus on sustainable AI due to concerns regarding their environmental, social, and economic implications. The challenges in evaluating the sustainability of AI was emphasised (Heilinger, Kempt and Nagel, 2024). The authors emphasise the importance of refraining from simplistic or excessively optimistic evaluations of AI's impact on sustainability, advocating for a meticulous consideration of environmental, social, and economic aspects. The authors introduce the distinction between "thin" and "thick" sustainability, positing that achieving genuine sustainability necessitates a thorough assessment of the entire life cycle of AI, encompassing considerations such as energy consumption, material resources, and social ramifications. Despite providing a strong theoretical framework, the research lacks specific guidance on implementing sustainability principles in the AI development process. Furthermore, the impact of AI on promoting sustainable urban development was elaborated on (Al-Raeei, 2024). The research emphasises the capacity of AI to facilitate intelligent urban planning through evidence-based decision-making while also drawing attention to ethical dilemmas, including bias in AI, transparency challenges, and privacy implications. One limitation of the research is the absence of explicit implementation guidelines or practical tools for developers to utilise. To address this gap, our research introduces the Sustainable AI Development Card, a practical tool designed to assist developers in making sustainably conscious decisions at every stage of the AI development pipeline. This card provides specific recommendations that go beyond corporate or organisational initiatives, enabling independent developers and small teams to contribute to the sustainability of AI systems.

3 Methodology

3.1 Traditional AI Pipeline

Traditional AI pipelines are organised into three main phases, which are data management, model management, and model deployment. The data management phase involves gathering raw data from different origin sources such as databases, APIs, sensors, or web scraping applications and is first stored in a centralised place or data lake. The purpose of this step is to collect all the relevant data and store it securely for analysis. Once data is collected, Exploratory Data Analysis (EDA) helps in learning the characteristics of the data and finding patterns and anomalies or outliers. EDA is about statistical analysis and visualisation for getting insights into the data distribution or relations between variables. Then, data preprocessing is done for cleaning and transforming the data. This includes handling missing values, correcting inconsistencies, normalising or standardising numerical features, and encoding categorical variables. Feature engineering then creates or selects the most relevant features to improve model performance. Techniques such as feature selection, extraction, and creation of interaction terms are used to enhance the predictive power of the dataset. The model management phase focuses on developing powerful machine-learning models. It begins with model selection, where suitable algorithms are chosen based on the problem type, classification, regression, clustering, etc., and the data attributes. Considerations include the algorithm's complexity, interpretability, and computational efficiency. To discover underlying patterns and relationships, the chosen models are trained on the prepared dataset. s. Model validation follows, employing techniques like cross-validation or hold-out validation to assess how well the model generalises to unseen data and to prevent overfitting. To improve model performance, hyperparameter tuning is a component of model optimisation. Methods such as grid search, random search, or Bayesian optimisation are used to find the optimal set of hyperparameters. Model evaluation is then conducted using appropriate metrics, like accuracy, precision, recall, F1-score for classification tasks, or RMSE for regression, to determine the best-performing model. This comprehensive evaluation ensures that the model meets the desired performance criteria before deployment. Finally, the model deployment phase involves integrating the validated model into a production environment to make real-time predictions or inform decision-making processes. Deployment can be executed through APIs, web services, or embedding the model into existing applications. Once deployed, continuous monitoring of the model's performance is essential to detect any degradation over time due to factors like data drift or changes in user behaviour. Monitoring involves tracking key performance indicators and setting up alerts for significant deviations. Model maintenance, including re-tuning or retraining the model, is performed based on the monitoring insights to maintain optimal performance. Continuous Integration and Continuous Deployment (CI/CD) practices are implemented to automate the deployment process. CI/CD pipelines facilitate rapid testing and deployment of model updates, ensuring that improvements are quickly and safely integrated into the production system.

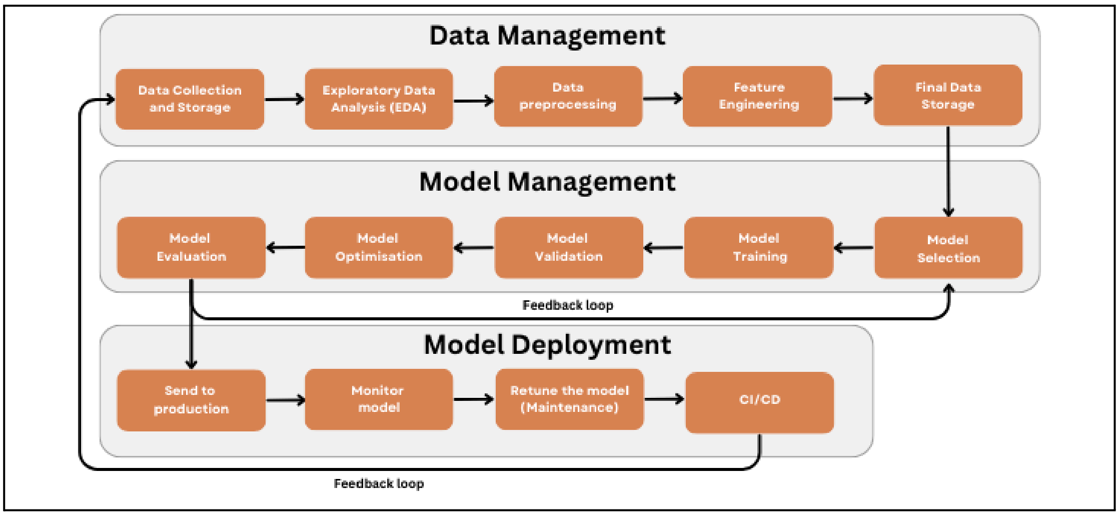


Figure 2. Traditional AI Pipeline

3.2 Problems in the traditional AI Pipeline

Due to their primary focus on performance optimisation, traditional AI pipelines frequently ignore sustainability issues, which results in significant energy consumption and resource inefficiencies during the data management, model management, and deployment stages.This section explores these key issues within the AI pipeline, underscoring the need for more eco-friendly practices in AI development. The data management stage of the AI pipeline presents significant sustainability challenges due to inefficiencies in data collection, storage, and processing. Many AI models rely on massive datasets, often collected without adequate filtering, which leads to unnecessary storage and increased energy consumption in data centres, contributing to a higher carbon footprint (Wu *et al.*, 2022). Although data scaling is commonly employed to enhance model performance, the environmental impact can be greatly increased by this method. Studies show that scaling data without optimising the storage and ingestion pipeline results in higher energy consumption while combining data scaling with model scaling can reduce energy usage but still comes with significant environmental costs (Sachdeva, Wu and McAuley, 2021). Additionally, EDA and pre-processing tasks are often computationally expensive and repetitive, especially when large, unfiltered datasets are involved, further driving up resource consumption. Moreover, data pre-processing and feature engineering, though essential for improving model performance, are typically resource-intensive and inefficient. Techniques such as data sampling or dimensionality reduction are underutilised, leading to excessive computational demands during model training (Verdecchia, Sallou and Cruz, 2023). Another key issue is data perishability, the fact that not all data retains its predictive value over time. Research has shown that some datasets, such as those in natural language processing (NLP), lose up to half of their predictive value within a few years, making long-term storage of large volumes of outdated data inefficient and environmentally costly (Valavi *et al.*, 2020). Finally, the long-term storage of processed datasets, particularly in cloud infrastructures, contributes significantly to the overall environmental footprint of AI systems, especially when energy-efficient storage solutions are not utilised (Patterson *et al.*, 2021). These inefficiencies in data management emphasise the need for more sustainable practices in AI development, including the adoption of intelligent data sampling, optimisation of data storage, and integration of energy-efficient infrastructures to reduce the carbon emissions associated with AI systems. The AI pipeline poses multiple sustainability challenges, especially in the field of model management. The essential issues have been identified within each phase of the model management process. In the process of model selection, a key challenge lies in the inclination towards employing highly complex models like deep learning despite the potential adequacy of simpler algorithms. This leads to avoidable energy usage and computational expenses. Furthermore, a significant number of developers emphasise optimising performance without taking into account the environmental implications, resulting in inefficient resource utilisation despite the feasibility of achieving comparable outcomes with simpler models. Training models is the stage that demands the most resources, particularly so with models of a large scale. The widespread utilisation of high-performance hardware such as GPUs contributes to elevated energy consumption. In addition, ineffective training methods, such as the absence of early stopping, can result in longer training durations and increased resource consumption, ultimately contributing to a greater environmental impact. Cross-validation methods, such as k-fold validation, necessitate retraining the model multiple times, leading to a significant increase in computational resources. The repetitive training of large datasets leads to a notable increase in energy consumption. Moreover, numerous validation procedures prioritise precision exclusively, neglecting to take into account the environmental repercussions of recurrent computational activities. For instance, conventional methods for hyperparameter tuning, such as grid search, explore a large number of combinations, resulting in a considerable computational burden. Despite the possibility of lowering time and energy demands through more effective techniques like Bayesian optimisation, extensive investigations take a substantial amount of resources. This lack of efficiency leads to the unnecessary consumption of energy for minimal enhancements in performance. During the evaluation of a model, emphasis is typically placed on enhancing accuracy or precision, potentially leading to the over-optimisation of models. These models offer only minimal improvement in performance at a significant computational expense. Moreover, the repetitive assessment of sizable models on extensive datasets amplifies energy consumption, contributing to the environmental footprint. The Model Deployment phase of the AI pipeline encounters several sustainability challenges. When sending the model to production, deploying complex AI models, especially deep learning models, requires substantial computational resources for real-time inference. This leads to high energy consumption and increased carbon emissions. During the monitoring of the model, continuous tracking of performance metrics and data drift necessitates ongoing computational operations. This persistent energy usage contributes to higher operational costs and a larger environmental footprint. How data centres, which house these AI models, consume a growing share of global electricity was discussed, projected to reach up to 8% by 2030 (Jones, 2018). The need to retune or maintain the model in response to concept drift or performance degradation presents additional sustainability issues. Frequent retraining of models is computationally intensive and increases energy consumption. According to a thorough analysis of learning under concept drift (Lu *et al.*, 2019), noting that traditional approaches to model adaptation can be resource-intensive and may not be sustainable in the long term. Implementing CI/CD pipelines introduces complexities that can accumulate as technical debt over time. The concept of "hidden technical debt" in machine learning systems was introduced, explaining that the maintenance and evolution of deployed models can lead to increased system complexity and inefficiencies (Sculley *et al.*, 2015). The sustainability of the AI deployment may be challenged by increased energy and resource consumption brought on by this technical debt.

3.3 Sustainable AI Development Card

In addressing the challenges outlined in the traditional AI pipeline, and informed by insights gained from our literature review, we identified specific sustainable techniques to counter each problem. Through this analysis, we developed a series of questions that form the foundation of our Sustainable AI Development Card, guiding developers in adopting sustainable practices at every stage. The Sustainable AI Development Card proposed in this research functions as a practical tool to assist individual developers in incorporating sustainability into the AI development process. The card provides strategies for reducing energy usage and effectively managing resources throughout different phases, such as data collection, model training, and deployment. For instance, it recommends utilising low-power hardware and employing model optimisation, quantisation, and pruning techniques in order to conserve computational resources and minimise superfluous consumption. This card aims to facilitate independent developers in incorporating sustainable AI practices. The Sustainable AI Development Card provides a series of questions and choices that assist developers in working towards sustainability objectives throughout the various phases of development, as seen in Table 1.

Sustainable AI Development Card				
Category	**Sub-category**	**Questions**	**O**	**X**
Hardware Selection		Have you selected low-power consumption hardware? (e.g., low-power processors, energy-efficient hardware)		
		Have you reviewed the energy efficiency of the hardware you are using? (e.g., energy usage monitoring tools, power consumption minimisation strategies)		
		Can you reduce unnecessary hardware usage by utilising cloud-based solutions? (e.g., cloud resource optimisation, serverless architecture)		
Data Management	Data Collection	Are you collecting only the necessary amount of data? (e.g., data filtering techniques, avoiding redundant data collection)		
		Have you considered using smaller, optimised datasets without sacrificing model performance? (e.g., data sampling techniques, smaller but representative datasets)		
		Have you minimised the environmental impact of data collection processes? (e.g., remote data collection minimisation, real-time filtering)		
		Have you ensured that the data collected does not unintentionally exclude underrepresented groups? (e.g., diversity checks, inclusive sampling methods)		
	EDA	Are you using representative data subsets for exploratory analysis? (e.g., data sampling for EDA, exploratory tasks on smaller data portions to reduce computational overhead)		
		Have you utilised automated tools to reduce the need for repetitive exploratory analysis? (e.g., Automated data profiling, Pre-built data summary reports to		

		minimise manual EDA efforts)		
	Data Pre-processing	Are you minimising unnecessary operations during pre-processing? (e.g., selective data cleaning, applying transformations only to required data)		
		Have you applied parallel or optimised methods for large dataset pre-processing? (e.g., Parallel data processing, Reducing the size of datasets before processing through techniques like filtering or dimensionality reduction)		
	Feature Engineering	Are you selecting only the most relevant features for your model? (e.g., feature selection, dimensionality reduction methods like PCA)		
		Have you considered techniques to reduce the computational cost of feature engineering? (e.g., sparse feature representations, parallel processing, reusing features across models)		
	Data Storage	Are you storing only essential datasets for training and validation? (e.g., data archiving or deletion for outdated datasets, retaining only critical datasets for ongoing models)		
		Have you implemented energy-efficient storage solutions? (e.g., energy-efficient cloud storage, cold storage for infrequently accessed data, avoiding duplication of data across storage systems)		
		Are you actively managing data life cycles to prevent long-term storage of unnecessary data? (e.g., automated data life cycle management, scheduled data purging or compression)		
Model Management	Model Selection	Have you considered the most energy-efficient algorithm for your task? (e.g., choosing simpler models when possible, assessing the trade-off between model complexity and performance)		
		Are you selecting models based on both performance and resource consumption? (e.g., using energy-efficient models, evaluating energy footprint alongside accuracy metrics)		
		Have you checked if the chosen model minimises bias and promotes fairness? (e.g., fairness auditing tools, bias mitigation strategies)		
		Is the selected model cost-effective, especially for small-scale applications? (e.g., simpler models with lower computational costs)		
	Model Training	Are you optimising your training process to minimise energy consumption? (e.g., early stopping, gradient accumulation, efficient learning rate scheduling)		
		Have you considered the use of distributed training or cloud resources to reduce the environmental impact? (e.g., distributed training across multiple machines, cloud-based resource optimisation)		
		Are you ensuring that the training process does not introduce bias that could affect social fairness? (e.g., balanced training data, algorithm fairness checks)		

	Model Validation	Are you using efficient validation techniques to reduce computational overhead? (e.g., k-fold cross-validation with fewer folds, proxy validation on representative data subsets)		
		Have you considered minimising the number of validation iterations without sacrificing model performance? (e.g., reducing the number of validation rounds, using holdout validation for large datasets)		
		Have you considered the societal implications of deploying this model? (e.g., ethical impact assessments, transparency in model deployment)		
		Are you ensuring that the deployment infrastructure is economically sustainable for long-term use? (e.g., affordable cloud options, resource-saving strategies)		
	Model Optimisation	Are you using energy-efficient hyperparameter tuning methods? (e.g., bayesian optimisation, random search instead of grid search)		
		Have you optimised the search space for hyperparameters to reduce unnecessary computations? (e.g., narrowing the hyperparameter range, using informed priors for search space)		
	Model Evaluation	Are you balancing performance improvement with resource costs during evaluation? (e.g., limiting evaluations to significant improvements, optimising evaluation cycles to reduce repetitions)		
		Have you reduced the size of test data to lower evaluation costs? (e.g., using representative subsets of test data and sampling techniques to reduce evaluation dataset size)		
Model Deployment	Send to Production	Have you optimised your model for efficient deployment? (e.g., model compression, pruning, quantisation, knowledge distillation)		
		Are you deploying only necessary components of your model? (e.g., feature selection, microservices architecture)		
		Have you considered the deployment environment's sustainability? (e.g., green data centres, edge computing)		
	Monitor the Model	Are you minimising resource usage in your monitoring processes? (e.g., adaptive monitoring, data sampling)		
		Have you optimised your monitoring tools for energy efficiency? (e.g., lightweight monitoring tools, efficient configurations, threshold setting)		
		Are you ensuring data privacy and security to prevent resource-draining breaches? (e.g., secure protocols, encryption, regular audits, vulnerability assessments)		
		Is the monitoring process equipped to detect and correct any social biases in real time? (e.g., bias detection tools, social fairness monitoring)		
		Are the monitoring tools affordable and efficient for ongoing maintenance?		

		(e.g., cost-efficient monitoring software, budget-friendly automation tools)		
	Retune the Model (Maintenance)	Are you using efficient methods for model retraining? (e.g., incremental learning, transfer learning)		
		Have you scheduled maintenance to optimise resource usage? (e.g., off-peak hours maintenance, batch updates)		
		Are you effectively detecting concept drift to avoid unnecessary retraining? (e.g., drift detection algorithms, performance monitoring, threshold-based triggers)		
	Continuous Integration/Continuous Deployment (CI/CD)	Have you optimised your CI/CD pipeline for efficiency? (e.g., caching mechanisms, parallel execution)		
		Are you minimising the environmental impact of your CI/CD processes? (e.g., selective testing, energy-efficient infrastructure, cloud optimisation)		
		Have you integrated security practices to prevent resource-intensive incidents? (e.g., automated security scans, access controls, authentication mechanisms)		

Table 1. Sustainable AI Development Card.

These inquiries and choices assist developers in making sustainable decisions throughout the life cycle of AI development. The options presented for each query are viable, sustainable approaches that empower individual developers to reduce sustainability risks when constructing AI systems.

4 Strategies and Benefits for Environmental, Social, and Economic Sustainability

Our proposed Sustainable AI Development Card functions as a tool created to aid developers in designing AI systems that align with sustainability principles encompassing environmental, social, and economic considerations. It functions as a valuable framework, providing specific strategies for AI developers to manage these aspects effectively and minimise adverse effects.

4.1 Environmental Sustainability: Carbon Emission Reduction

The concept of environmental sustainability aims to reduce the carbon emissions linked to the development of AI. Training a large language model may lead to carbon emissions equivalent to those generated by multiple vehicles throughout their entire life cycle. The Sustainable AI Development Card tackles this issue by advocating for energy-efficient methods across the entire life cycle of AI.

- **Optimising Models:** Developers are encouraged to apply techniques such as model pruning, quantisation, and knowledge distillation, which help streamline model architectures without compromising performance.

- **Energy-efficient Hardware:** Using specialised hardware like TPUs and energy-efficient GPUs reduces the energy demands during model training and inference. Furthermore, the deployment of distributed and cloud computing technologies can enable AI operations to be executed within environmentally friendly data centres, consequently reducing emissions.

- **Data Management:** Techniques like data reduction, smart data sampling, and dimensionality reduction are crucial for reducing the volume of data processed, thus saving energy during both storage and computation phases.

By implementing these strategies, developers of AI can effectively decrease the carbon footprint associated with energy-intensive activities such as model training and inference, thus supporting environmental sustainability.

4.2 Social Sustainability: Mitigating Bias and Ethical AI

Social sustainability focuses on the ethical implications of AI, especially the potential to perpetuate bias and inequality in decision-making. As AI is increasingly applied to areas such as healthcare, finance, and criminal justice, ensuring fairness and accountability is critical. The Sustainable AI Development Card offers recommendations for optimising equity and diversity, thereby guaranteeing that AI systems are aligned with principles of social equity. Essential practices consist of the following:

- **Bias Mitigation:** Developers are encouraged to implement algorithms that consider fairness and ensure that training datasets are diverse and representative. This aids in the mitigation of social biases being perpetuated by AI systems.

- **Transparency:** The Sustainable AI Development Card recommends the use of Model Cards, an approach introduced by Margaret Mitchell and colleagues, which documents the intended use, potential biases, and performance of AI models across different demographic groups. These cards enhance transparency and accountability, guaranteeing the ethical and responsible deployment of models.

- **Auditing and Monitoring:** Continuous monitoring and auditing of AI models are essential to detect and correct bias throughout the system's life cycle. This continuous evaluation guarantees that AI systems maintain fairness and inclusivity amidst evolving societal contexts.

The Sustainable AI Development Card ensures that AI contributes positively to society by incorporating ethical considerations into the development process, thereby minimising negative impacts such as bias and inequity.

4.3 Economic Sustainability: Reducing Energy Consumption and Costs

The economic sustainability of AI development is dependent on the cost-effectiveness and energy efficiency of systems, ensuring their long-term viability. The substantial expenses related to training large models computationally may pose a significant obstacle, particularly for smaller organisations and independent developers. The Sustainable AI Development Card proposes a range of strategies aimed at enhancing the cost-effectiveness of AI development.

- **Efficient Resource Allocation:** Utilising techniques such as early stopping, hyperparameter tuning, and mini-batch training can significantly lower computational expenses. These methods enable developers to reduce energy consumption during model training without compromising accuracy or efficiency.

- **Cloud-based and Edge Computing Solutions:** The Sustainable AI Development Card promotes the use of scalable cloud infrastructure that offers on-demand resources, allowing developers to only pay for what they need. Furthermore, the utilisation of AI models in proximity to the data source through edge computing diminishes the necessity for constant, energy-intensive communication with centralised data centres.

- Hardware Optimisation: Implementing energy-efficient hardware solutions, such as TPUs and low-power GPUs, ensures that AI systems run on optimised infrastructure that consumes less energy while maintaining high performance.

Through adherence to these practices, developers have the ability to decrease the comprehensive expenses associated with AI development, thereby broadening its accessibility and fostering economic sustainability. The Sustainable AI Development Card provides a thorough framework to assist developers in creating AI systems that prioritise efficiency as well as social and environmental responsibilities. The Sustainable AI Development Card assists in aligning AI development with sustainability goals by addressing carbon emissions, social bias, and resource optimisation. This tool enables developers, especially those facing resource constraints, to participate in sustainable AI principles, ensuring the constructive impact of AI in the future.

4.4 Balancing Performance and Sustainability

Confronted with the practical reality of needing powerful, accurate models, reducing the computational resources might seem like a performance downgrade and trade-offs in accuracy, speed and scalability, presenting a formidable and seemingly insurmountable challenge. Model simplification, for instance, is one of the ways to reduce energy consumption. Often achieved through reducing architectures in size, applying rule-based methods and model pruning, this simplification logically decreases the resources needed for training and inference-making. Nevertheless, simpler models, although their simplification could improve interpretability, struggle to capture complex relationships in large datasets, as opposed to more complex models, which might be more efficient. Ensemble techniques, such as combining simplified models, could compensate for accuracy loss without compromising much in the area of resource requirements. By the same token, techniques for model reduction, such as quantisation and compression, may introduce rounding errors, thus impacting model precision and consequently compromising the quality of results. Quantisation-aware training might support developers in enabling the model to learn and adapt to lower-precision constraints and help retain higher accuracy post-quantisation. Using mixed-precision quantisation could be especially beneficial for deep neural networks and allow for optimised resource deployment. In situations where training resources are limited, post-training quantisation can be applied to trained models. While less effective than quantisation-aware training, post-training quantisation can be considered a more resource-efficient method to achieve a reasonable trade-off between accuracy and efficiency. A well-recognised approach to enhance model performance is hyperparameter tuning, yet some traditional methods to achieve it can be resource-intensive. To create more sustainable AI models, developers can employ efficient alternatives that minimise computational costs while maintaining model quality. One alternative is Bayesian optimisation, as it optimises the focus of the model on promising areas of the hyperparameter space, reducing the need for exhaustive searches and conserving energy. Lastly, in some cases of real-time AI deployment, balancing latency with sustainability is essential, as latency directly impacts responsiveness. Methods like conditional execution and knowledge distillation allow models to perform calculations only when needed, saving time and energy by simplifying computations, especially when the confidence in predictions is high.

5 Discussion and Future Works

The Sustainable AI Development Card addresses sustainability concerns at every phase of AI development, from data management to model deployment. It offers strategies for reducing carbon footprint and promoting social and economic sustainability. The card’s recommendations help developers minimise the sustainability risks of AI without sacrificing performance much, making it a valuable tool for both small teams and large organisations aiming to achieve sustainability goals. The card’s emphasis on ethical deployment and reducing carbon footprints ensures that AI systems align with broader sustainability goals while maintaining operational efficiency.

The Sustainable AI Development Card is indeed exhaustive, but its delivery still faces numerous challenges. An obvious challenge is that developers often focus on performance metrics like accuracy, speed, and scalability over energy efficiency. The strong focus on these aspects can lead to the design of highly complex models and high resource consumption, creating challenges in following the sustainability guidelines suggested by this card. It can also demand a significant degree of change from the current process and mindset to incorporate sustainable practices. Developers used to focusing predominantly on performance may be reluctant to move towards sustainability, especially when it incurs extra costs or slows down time-to-market. While vitally important, this shift in culture and operations will not happen without organisational support and attendant incentives. A further challenge is scalability. While the Sustainable AI Development Card is highly applicable to smaller projects, its recommendations can be difficult to implement in larger, distributed systems. Organisations operating at scale must find a balance between meeting tight performance deadlines and ensuring that their AI infrastructure adheres to sustainable principles. Implementing the card's strategies consistently across large infrastructures presents logistical and technical challenges. Furthermore, it is imperative to contemplate incorporating the existing manual-based model card into a software application. This methodology would incorporate a user-friendly interface and automated features, thereby improving the practical usability of the card. These software solutions are designed to smoothly integrate with current workflows and facilitate scalability, thereby encouraging the broad implementation of sustainable AI development practices. While the card presents a promising approach to sustainable AI, future research is needed to empirically validate its effectiveness. Through quantitative analysis and case studies, we aim to substantiate the card's claims regarding reduced carbon footprint and energy efficiency, providing concrete data to support its recommendations. Such validation efforts will strengthen the card's utility and encourage its adoption across diverse AI development contexts.

As concerns about carbon emissions and energy usage continue to rise, developers and organisations are likely to prioritise sustainable AI development. Industry-wide standards for sustainable AI may emerge, encouraging the broader use of tools like the Sustainable AI Development Card to guide development decisions. Technological advancements also offer significant opportunities. Energy-efficient hardware, such as TPUs and low-power GPUs, is becoming more available, offering developers the ability to run AI models with less power consumption. As these technologies mature and become more affordable, they will likely drive greater adoption of sustainable practices. Moreover, automated tools that track energy usage throughout the AI development process present another opportunity. These tools could provide developers with real-time insights into their energy consumption, helping them make immediate adjustments and further optimise sustainability.

6 Conclusion

This research introduced the Sustainable AI Development Card as a key tool to guide AI developers in adopting sustainable practices throughout the AI development life cycle. The Sustainable AI Development Card addresses three critical dimensions of AI sustainability: environmental, social, and economic. Through the use of energy-efficient hardware, model optimisation techniques like pruning and quantisation, and responsible data handling, the card offers a comprehensive approach to reducing the carbon footprint of AI systems. Moreover, the framework emphasises the ethical use of AI, ensuring that social equity and fairness are prioritised alongside technical efficiency. Opportunities for the future of sustainable AI development are promising. Advances in hardware, automated sustainability monitoring tools, and increased awareness of AI's environmental impact will likely drive greater adoption of sustainable practices. Furthermore, as the demand for accountability and transparency in AI continues to grow, the role of the Sustainable AI Development Card in shaping industry-wide standards for sustainable AI development will become increasingly important.

In conclusion, while further work is needed to validate the card's effectiveness through empirical testing and quantitative analysis, the Sustainable AI Development Card represents a significant step toward more responsible, efficient, and eco-friendly AI technologies, this research contributes to a practical framework that aligns AI development with global sustainability objectives. In upcoming research, we aim to provide more quantitative results and case studies from the Card's integration, recognising that further investigation is essential to fully explore its impact on sustainable AI development.

References

Al-Raeei, M. (2024) 'The smart future for sustainable development: Artificial intelligence solutions for sustainable urbanization', Sustainable Development, n/a(n/a). Available at: https://doi.org/10.1002/sd.3131.

Devlin, J. et al. (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. arXiv. Available at: https://doi.org/10.48550/arXiv.1810.04805.

Han, S., Mao, H. and Dally, W.J. (2016) 'Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding'. arXiv. Available at: https://doi.org/10.48550/arXiv.1510.00149.

Heilinger, J.-C., Kempt, H. and Nagel, S. (2024) 'Beware of sustainable AI! Uses and abuses of a worthy goal', AI and Ethics, 4(2), pp. 201–212. Available at: https://doi.org/10.1007/s43681-023-00259-8.

Jacob, B. et al. (2017) 'Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference'. arXiv. Available at: https://doi.org/10.48550/arXiv.1712.05877.

Jones, N. (2018) 'How to stop data centres from gobbling up the world's electricity', Nature, 561(7722), pp. 163–166. Available at: https://doi.org/10.1038/d41586-018-06610-y.

Kairouz, P. et al. (2021) 'Advances and Open Problems in Federated Learning'. arXiv. Available at: https://doi.org/10.48550/arXiv.1912.04977.

Li, T. et al. (2019) 'Federated Learning: Challenges, Methods, and Future Directions'. arXiv. Available at: https://doi.org/10.48550/arXiv.1908.07873.

Lu, J. et al. (2019) 'Learning under Concept Drift: A Review', IEEE Transactions on Knowledge and Data Engineering, 31(12), pp. 2346–2363. Available at: https://doi.org/10.1109/TKDE.2018.2876857.

McMahan, H.B. et al. (2023) 'Communication-Efficient Learning of Deep Networks from Decentralized Data'. arXiv. Available at: https://doi.org/10.48550/arXiv.1602.05629.

Mitchell, M. et al. (2019) 'Model Cards for Model Reporting', in Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 220–229. Available at: https://doi.org/10.1145/3287560.3287596.

Patterson, D. et al. (2021) 'Carbon Emissions and Large Neural Network Training'. arXiv. Available at: https://doi.org/10.48550/arXiv.2104.10350.

Sachdeva, N., Wu, C.-J. and McAuley, J. (2021) 'SVP-CF: Selection via Proxy for Collaborative Filtering Data'. arXiv. Available at: https://doi.org/10.48550/arXiv.2107.04984.

Schwartz, R. et al. (2020) 'Green AI', Commun. ACM, 63(12), pp. 54–63. Available at: https://doi.org/10.1145/3381831.

Sculley, D. et al. (2015) ‘Hidden Technical Debt in Machine Learning Systems’, in Advances in Neural Information Processing Systems. Curran Associates, Inc. Available at: https://proceedings.neurips.cc/paper_files/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html (Accessed: 11 November 2024).

Shen, S. et al. (2019) ‘Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT’. arXiv. Available at: https://doi.org/10.48550/arXiv.1909.05840.

The Economist (2018) ‘The cost of training machines is becoming a problem’. Available at: https://www.economist.com/technology-quarterly/2020/06/11/the-cost-of-training-machines-is-becoming-a-problem (Accessed: 11 November 2024).

Touvron, H. et al. (2021) ‘Training data-efficient image transformers & distillation through attention’. arXiv. Available at: https://doi.org/10.48550/arXiv.2012.12877.

Valavi, E. et al. (2020) Time and the Value of Data. Harvard Business School.

Verdecchia, R., Sallou, J. and Cruz, L. (2023) ‘A systematic review of Green AI’, WIREs Data Mining and Knowledge Discovery, 13(4), p. e1507. Available at: https://doi.org/10.1002/widm.1507.

Wang, N., Liu, C. C. C., Venkataramani, S., Sen, S., Chen, C. Y., El Maghraoui, K., ... & Chang, L. (2022). Deep compression of pre-trained transformer models. *Advances in Neural Information Processing Systems*, *35*, 14140-14154.

World Commission on Environment and Development (1987) Our Common Future. Oxford: Oxford University Press. Available at: https://sustainabledevelopment.un.org/content/documents/5987our-common-future.pdf.

Wu, C.-J. et al. (2022) ‘Sustainable AI: Environmental Implications, Challenges and Opportunities’, Proceedings of Machine Learning and Systems, 4, pp. 795–813.

Yang, Q. et al. (2019) ‘Federated Machine Learning: Concept and Applications’. arXiv. Available at: https://doi.org/10.48550/arXiv.1902.04885.

Zafrir, O. et al. (2019) ‘Q8BERT: Quantized 8Bit BERT’. arXiv. Available at: https://doi.org/10.48550/arXiv.1910.06188.

Evaluating Organisational Readiness for Predictive Risk Intelligence Integration

Nicholas Quinn (Northumbria University - London), Usman Butt (Ajman University) and Mansour Alraja (Northumbria university)

Abstract

Predictive Risk Intelligence (PRI) is an AI use case currently deployed across various domains to forecast potential risks before they occur. At the heart of its predictive engine, is a confluence of machine learning (ML) and big data technologies referred to as smart information systems. Optimal utilization of this use case is expected to bring benefits to organisations and contribute to the drive towards sustainable development goals outlined by the United Nations. However, these advantages are hindered by system threats and vulnerabilities which are accompanied with ethical implications. There is a limited corpus of Empirical academic work in this particular field that is related to the governance of PRI and the wider applications of SIS. This paper aims to review the state of the art and contribute to empirical findings to evaluate how ready organizations are to integrate PRI as a driver towards SDGs whilst mitigating the ethical implications.

Keywords: Predictive Risk intelligence, Smart Information Systems, Artificial Intelligence, Machine Learning, Big Data. Big Data Analytics.

1. Introduction

Predictive risk intelligence (PRI) using Smart Information Systems (SIS) is an Artificial Intelligence (AI) use case adopted across several domains, that leverages advanced technologies and data analytics to identify, assess, and forecast potential risks before they occur.

SIS have emerged in recent years, as a broad technological concept to describe the combination of AI and advanced Big Data Analytics (BDA) to collect, process and analyse large and complex datasets, from various sources, in an intelligent and automated manner. Ryan et al (2020, p. 3) uses the term "SIS" to categorize Socio-Technical Systems (STS) that leverage a specific AI technique called Machine Learning (ML) often employing artificial neural networks, to extract insights from vast quantitiesof typically unstructured data. STS are a concept that denotes the design of a system that optimises the confluence of social arrangements and technical components (Bauer & Herder, 2009). The interplay between the components of SIS is evident in activities like sentiment analysis and trend prediction, where vast amounts of data generated by social media platforms are processed by ML algorithms, which augment and automate delineated tasks such as predictive modelling and sentiment classification (Ryan et al., 2020, p. 3). Akin to similar emerging technologies, the discourse surrounding the employment of SIS, is often directed towards the issue of managing and optimizing the consequences of innovation. Central to this discussion are the human rights affected by SIS as expressed United Nation's (UN) Universal Declaration of Human Rights and their implementation by the attainment of actionable and measurable Sustainable Development Goals (SDG's) (Ryan et al., 2020, p. 3). To ensure that AI systems work towards SDG's and not against them, requires a carefully co-ordinated global collaborative effort involving the use of dynamic and agile policy frameworks to safeguard and guide their implementation (Miailhe et al., 2020, pp. 214-216).

The ISO/IEC TR 24030 is a technical report is one step towards achieving driving Miailhe et al's (2020) collaborative effort. The report aims fosters collaboration between internal, external and potential stakeholders in AI applicability by gathering information to inform and guide AI standardization efforts, thus demonstrating how standardisation can be applied across various domains SDG's (ISO-IEC TR 24030, 2024). Another aim of this report is to identify new technical requirement to accelerate scientific and technological advancements, which akin to Ryan et al (2020) and Miailhe et al's (2020) is measured for success in terms of the attainment specific SDG'S. The document is organized inter alia 7 clause that define, a diverse range of application domains, where examples include Supply chain Management (SCM), Fintech and Healthcare as well as AI applications which include demand forecasting, loan screening, diagnosis support, before evidencing and referencing the actual use cases of these applications in active deployment (ISO-IEC TR 24030, 2024). Referenced as Use Case 164, PRI is in deployment across multiple domains and applications, including the aforementioned examples as a potential driver, specifically towards the contribution of SDG 9: Industry, Innovation and Infrastructure. The core technology at the heart of this use case, is a predictive engine that uses SIS to provide the foresight to clients, so that they can detect risks on a global scale days, sometimes weeks in advance of an unwanted occurrence (ISO-IEC TR 24030, p. 85). The rapid adoption of SIS and its enabling technologies like Internet of Things (IoT), ML, Cloud Computing and social media brings tremendous benefits, however, also introduce a number of system security threats and vulnerabilities that raise ethical considerations over Unauthorised Access and Misuse of Data, algorithmic integrity, fairness, accountability, transparency, bias, and accuracy (Jiya, 2019).

2. Systematic Literature Review

A systematic literature review followed an open and axial coding scheme as to identify the predominant technologies used for PRI and the system threats and vulnerabilities.

Data points	Axial Code	Open Code	Article
Quality Evaluation	Research type	Case studies	(Stahl et al., 2021, Tilimbe, 2019, Ryan et al., 2019, Stahl et al., 2022b, Ryan et al., 2021)
		Qualitative Interviews	(Trim & Lee, 2022)
		Policy Mitigation Strategies	(Stahl et al., 2021, Stahl et al., 2022b, (Ryan et al., 2021)
		Ethical Analysis	(Stahl et al., 2021, Andreou et al., 2019, Tilimbe, 2019, Ryan et al., 2019, Stahl et al., 2022b, Ryan, 2020, Ryan et al., 2021)
		Human Rights Analysis	(Stahl et al., 2022, (Andreou et al., 2019, Ryan et al., 2019, Ryan et al., 2021)
		Technical Analysis	(Stahl et al., 2021, Trim & Lee, 2022, Ryan et al., 2019, Stahl et al., 2022)
Technology	Technical Components of SIS	Machine Learning / Deep Learning	(Stahl et al., 2021, Tilimbe, 2019, Trim & Lee, 2022, Ryan et al., 2019, Stahl et al., 2022b)
		Big Data Analytics	(Stahl et al., 2021, Ryan et al., 2019, Stahl et al., 2022b)
		Predictive Analytics	(Tilimbe, 2019, Ryan et al., 2019, Stahl et al., 2022b)
		Artificial Neural Networks	(Stahl et al., 2021, Tilimbe, 2019, , Stahl et al., 2022b)
		Algorithms	(Tilimbe, 2019, Trim & Lee, 2022, Ryan et al., 2019, Stahl et al., 2022b)
Ethical Implications	System Threats & Vulnerabilities	Mis-implementation or Misuse or dual use of Machine Learning Systems	(Stahl et al., 2021, Andreou et al., 2019)
		Attacks on Machine Learning Models and Algorithms	(Stahl et al., 2021,Trim & Lee, 2022, Tilimbe, 2019)
		Malicious Uses of AI and Data Analysis	(Stahl et al., 2021, Tilimbe, 2019)
		Algorithmic Bias and Discrimination	(Stahl et al., 2021, Tilimbe, 2019, Ryan et al., 2021)
		Privacy and Data Protection	(Stahl et al., 2021, Tilimbe, 2019, Andreou et al., 2019, Stahl et al., 2022b, Ryan et al, 2021
		Security and Integrity of Systems	(Stahl et al., 2021, 2019, Stahl et al., 2022b, Andreou et al., 2019, Ryan et al., 2021)
		Transparency and accountability	Andreou et al., 2019 , Stahl et al., 2022b, Ryan et al, 2021
		Trust and Accuracy	(Ryan et al., 2019,Tilimbe, 2019, Andreou et al., 2019, Ryan et al., 2019, Ryan et al, 2021)
		Democracy, Freedom of Thought, Control, and Manipulation	(Ryan et al., 2019, Stahl et al., 2022b)
		Power Asymmetry	(Stahl et al., 2021, Tilimbe, 2019, Ryan et al., 2019, Stahl, Ryan et al, 2021)

Table 1: Axial and open coding scheme for PRI

3. Research Design and Methodology

3.1 Theoretical Foundation

The research for this study was focused on the intersection of a framework of processes and controls for managing the principles of organisational Information Governance (IG), Risk, and Compliance and the AI use case of PRI. In the context of this study the principle of IG aligns with Yusif & Hafeez-Baig's (2021) concept of cyber-security (CS) governance; as a set of policies, procedures and guidelines to achieve the organisations objectives and protect digital assets. Across domains such as CS and Information Security (IS), organisations may choose to adopt governance best practises that meet the requirements laid out by standards to successfully achieve and demonstrate policy compliance (Yusif & Hafeez-Baig, 2021, pp. 15-16). Best practices for Risk Management (RM) are fundamental to IG policy, therefore provide the foundation for standards such as the International Organization for Standardization (ISO) 27001, and National Institute of Standards and Technology (NIST) Cybersecurity Framework (CSF) (McIntosh et al., 2024). This is particularly relevant to organizations that adopt PRI as this reports previous sections have identified system threat and vulnerabilities that pose significant risks to organisations adopting this use case, potentially with implications for wider society. Policy should be aligned with a wider demonstration of compliance with the organisation's contractual, ethical, and regulatory obligations. This may take the form of various hard and soft regulatory instruments, (Birkstedt et al., 2023, p. 148) (Birkstedt et al., 2023, p. 14). Hard regulation includes the provisions from instruments such as the GDPR and the newly in force EU AI act, which although span the EU jurisdiction, have set an international regulatory precedent (McIntosh et al., 2024). Soft regulation refers to guidance such as the IEEE 700 series of standards of transparency of autonomous systems and the Principles for the Ethical Use of Artificial Intelligence in the United Nations System (Birkstedt et al., 2023). These ethical principle, represent a synthesis of commonly agreed upon shared values within wider international AI discourse (Jobin et al., 2019), often forming the basis for both hard and soft regulation. Whilst these, principles are widely adopted, it is argued that they are theoretical in nature, addressing the question of ‘what’, which means that the next logical question relates to ‘how’ these concepts are implemented, a question addressed through the exploration of tangible governance processes (Birkstedt et al., 2023), and ‘where’ the benchmark is for which the effectiveness of these are measured, which is widely discussed in the previous sections SLR as through the attainment of measurable SDG’s. Therefore, whilst UN SDG’s do not necessarily constitute a direct compliance requirement, the universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity are increasing becoming important factors in organisational social and sustainability responsibility (ISO-IEC TR 24030-2024). Thus, the objective of this study is to use a relevant standard for GRC as a benchmark to assess the maturity and preparedness of corporate approaches to GRC for organisations that use PRI and whether they were consciously aligning its usage with SDG 9.

3.2 Research Method

An inductive research approach was taken through qualitative research using semi-structured interviews. It was determined that this method was preferable to deductive research due to scarcity of predefined theories in the existing literature, therefore it was hoped that the inductive approach from which varying subjective observations are derived (Ameen et al., 2024), would enable a contribution of new theories. The rationale for semi structured interviews was largely due to suitability to qualitative research, enabling the SLR to inform a general guide, whilst at the same time, allowing flexibility for the researcher to adapt the questions to the flow of the interview accordingly (Saunders et al., 2015). This would allow for adjustments to be made for various domains and levels of expertise. Semi-structured interviews are less often used for exploration between two or more variables in both inductive and deductive research (Saunders et al., 2015). Therefore, it was hoped that this method would facilitating deeper and additional insights into the themes already addressed by the SLR in relation to the ethical implications of PRI and SIS, but not expected. Semi-structured interviews are often used for explaining and evaluating the relationship between two or more variables (Saunders et al., 2015), therefore it was expected that this method would explain how the organisation is addressing ethical concerns and using PRI to drive SDG 9. By using the ISO/IEC 42001 as a benchmark, the researcher could evaluate as how well organisation were implementing processes such as oversight, data management and risk management to handle these challenges.

.

References

Ameen, N., Hoelscher, V., & Panteli, N. (2024, November 1). Exploring how mumpreneurs use digital platforms' algorithms and mechanisms to generate different types of value. *Information Systems Journal*. https://doi.org/10.1111/isj.12518

Andreou, A., Laulhe-Shaelou, S., & Schroeder, D. (2019). *Current Human Rights Frameworks*.

Aşuroğlu, T., & Gemci, C. (n.d.). Role of ethics in information security. *International Conference of Advanced Technology & Sciences*, 141–144.

Bauer, J. M., & Herder, P. M. (2009). Designing Socio-Technical Systems. *Philosophy of Technology and Engineering Sciences*, 601–630. https://doi.org/10.1016/B978-0-444-51667-1.50026-4

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: themes, knowledge gaps and future agendas. In *Internet Research* (Vol. 33, Issue 7, pp. 133–167). Emerald Publishing. https://doi.org/10.1108/INTR-01-2022-0042

ISO/IEC. (2023). *ISO/IEC 42001:2023 - AI management systems*. https://www.iso.org/standard/81230.html

ISO-IEC TR 24030-2024.

Jiya, T. (2019). Ethical Implications of Predictive Risk Intelligence. *The ORBIT Journal*, *2*(2), 1–28. https://doi.org/10.29297/orbit.v2i2.112

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., Nowrozy, R., & Halgamuge, M. N. (2024). *From COBIT to ISO 42001: Evaluating Cybersecurity Frameworks for Opportunities, Risks, and Regulatory Compliance in Commercializing Large Language Models*. https://doi.org/10.1016/j.cose.2024.103964

Miailhe, N., Hodes, C., Jain, A., Iliadis, N., Alanoca, S., & Png, J. (2020). AI for Sustainable Development Goals. *Delphi - Interdisciplinary Review of Emerging Technologies*, *2*(4), 207–216. https://doi.org/10.21552/delphi/2019/4/10

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, *26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Ryan, M., Antoniou, J., Brooks, L., Jiya, T., Macnish, K., & Stahl, B. (2020). The ethical balance of using smart information systems for promoting the United Nations' sustainable development goals. *Sustainability (Switzerland)*, *12*(12). https://doi.org/10.3390/SU12124826

Ryan, M., Antoniou, J., Brooks, L., Jiya, T., Macnish, K., & Stahl, B. (2021). Research and Practice of AI Ethics: A Case Study Approach Juxtaposing Academic Discourse with Organisational Reality. *Science and Engineering Ethics*, *27*(2). https://doi.org/10.1007/s11948-021-00293-x

Ryan, M., Antoniou, J., Macnish, K., Brooks, L., Stahl, B., & Jiya, T. (2019). *Technofixing the Future: Ethical Side Effects of Using AI and Big Data to meet the SDGs*.

Saunders, M., Lewis, P., & Thornhill, A. (2015). *Research Methods for Business Students PDF EBook*. Pearson Education, Limited. http://ebookcentral.proquest.com/lib/northumbria/detail.action?docID=5175059

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Laulhé Shaelou, S., Patel, A., Ryan, M., & Wright, D. (2021). Artificial intelligence for human flourishing – Beyond principles for machine learning. *Journal of Business Research*, *124*, 374–388. https://doi.org/10.1016/j.jbusres.2020.11.030

Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K., & Jiya, T. (2022a). Organisational responses to the ethical issues of artificial intelligence. *AI and Society*, *37*(1), 23–37. https://doi.org/10.1007/s00146-021-01148-6

Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K., & Jiya, T. (2022b). Organisational responses to the ethical issues of artificial intelligence. *AI and Society*, *37*(1), 23–37. https://doi.org/10.1007/s00146-021-01148-6

Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K., & Jiya, T. (2022c). Organisational responses to the ethical issues of artificial intelligence. *AI Soc.*, *37*(1), 23–37. https://doi.org/10.1007/s00146-021-01148-6

Tilimbe, J. (2019). Ethical Implications of Predictive Risk Intelligence. *The ORBIT Journal*, *2*(2), 1–28. https://doi.org/https://doi.org/10.29297/orbit.v2i2.112

Trim, P. R. J., & Lee, Y. I. (2022). Combining Sociocultural Intelligence with Artificial Intelligence to Increase Organizational Cyber Security Provision through

Enhanced Resilience. *Big Data and Cognitive Computing*, *6*(4). https://doi.org/10.3390/bdcc6040110

Yusif, S., & Hafeez-Baig, A. (2021). A Conceptual Model for Cybersecurity Governance. *Journal of Applied Security Research*, *16*(4), 490–513. https://doi.org/10.1080/19361610.2021.1918995

Prospective Theorising in Blockchain Applications: Contrasting Academic Sentiment and Practical Adoption

Nicole Mäkineste
Durham University, NTNU

Spyros Angelopoulos
Durham University

Research In progress

Abstract

Blockchain applications are envisioned as transformative across sectors. Signals from practice, however, indicate obstacles in adopting such applications. We investigate the gap between the theoretical promise and practical adoption of blockchain applications in fields outside of finance. In doing so, we explore the cumulated body of knowledge on blockchain applications through the lens of prospective theorising and evaluate the speculative rigour of this field and its imagined futures. Through the meta-analysis of 126 review articles, we assess the state of blockchain adoption, academic sentiment and real-world pilot projects and initiatives. The anticipated findings of our study showcase that while the academic discourse maintains a sentiment of persistent optimism, real-world adoption remains limited, with discontinued projects exemplifying recurring obstacles. We emphasise the need for enhanced speculative rigour and inclusion of descriptive theorising in blockchain research to ground expectations in empirical evidence and delineate an agenda for future research on the topic.

Keywords: blockchain, technology adoption, prospective theorising

1.0 Introduction

The advancements of blockchain technology have enabled the development of smart contracts, decentralised applications, and decentralised autonomous organisations (Ellinger et al., 2023). This was seen as the Blockchain 2.0 era that would inevitably give way to Blockchain 3.0: blockchain applications outside of the field of finance (Wang et al., 2019). Academic literature on potential blockchain applications paints a beautiful picture of this era: consumers scanning QR codes to inspect the production journey of a wine bottle (e.g., Parry et al., 2023), traceability applications conquering corruption in the shipping industry (e.g., Sarker et al., 2021), and patients sharing their electronic health records with doctors with the click of a button (e.g., Hasselgren et al., 2021).

Although reality has not yet caught up with this desirable-future-view, industry analysts believe it to only be a matter of time. The latest Hype Cycle Report by Gartner, for instance, describes that while blockchain technology has not yet witnessed widespread adoption, its enterprise integration is at the brink of accelerated growth (Leow & Litan, 2024). However, in 2019, Gartner's corresponding report predicted a much faster adoption rate and stated that the blockchain market would witness an increase in adoption from 2021 onwards (Gartner, 2019). Such statements reflect a trend that can be witnessed across blockchain industry reports: the analysts are certain that widespread blockchain adoption is right around the corner, but year after year the goalpost is moved further (e.g., Fortune Business Insights, 2024). Concurrently, it is difficult to come by real-world blockchain applications outside of the field of finance and many of the widely used examples of enterprise blockchain adoption have silently disappeared into the shadows. As an example, TradeLens was a collaboration between IBM and Maersk to create a blockchain platform aimed at digitising the global shipping industry (Jensen et al., 2019). The research article on the project (*ibid*) is highly cited and the case study has been since used in several publications (e.g. Rani et al., 2024). However, the TradeLens platform has been discontinued (Maersk, 2022).

Despite the absence of real-world adoption, academic interest in blockchain applications remains high, envisioning a plethora of use cases, and conceptualised applications. Some of these applications have been built and validated in successful pilot projects. This shows that such applications can solve existing problems, and it is also technologically feasible to realise their potential. It is possible, however, to interpret signals from practice that indicate trouble in adopting such applications. There is a need, thus, for a clear-headed inspection of the state-of-the-art when it comes to blockchain adoption to level-set expectations and guide research to imagine desirable futures with increased speculative rigour. Thus, the first research question is:

What is the state-of-the-art of enterprise blockchain adoption in the literature?

The case of TradeLens exemplifies our second area of interest. As a pilot project for a blockchain application, TradeLens seemed to have it all; it was i) solving a proven real-world issue, ii) successfully built by industry experts, and iii) pioneered by a powerful industry player. Despite these traits, TradeLens was discontinued before it could reach wider adoption, which raises the question on whether this was an isolated failure or part of a pattern. Therefore, our second research question is:

What is the current status of real-world blockchain applications and pilot projects identified from academic literature?

2.0 Theoretical Background

To provide an answer to the research questions of our study, we incorporate a prospective theorising perspective. Prospective theorising is an approach that represents a dual shift from traditional theorising practices: from projection to imagination, and from values-neutral to values-led theorising (Gümüsay & Reinecke, 2024). It emphasises a future-oriented research approach and encourages the imagining of desirable futures with the intention of shaping the social reality rather than simply observing or predicting it (Gümüsay & Reinecke, 2022; Hanisch, 2024). The cumulated body of knowledge on blockchain applications demonstrates many characteristics of prospective theorising. The relevant research on the topic can be described as imagining a desirable future: it centres around theorising blockchain applications and use cases and imagining how their adoption will induct societal change. The research on the topic is also value-led: it generally posits that adoption of blockchain applications will result in desirable societal outcomes, such as, enhanced trust in processes and participants (Hawlitschek et al., 2018), privacy of personal information (Ahirao & Joshi, 2022), and decreased corruption (Sarker et al., 2021).

Future-oriented theorising requires speculative rigour. Gümüsay and Reinecke (2024) propose that speculative rigour composes of four main evaluative criteria: i) generative potency, ii) process transparency, iii) speculative plausibility, and iv) plausible desirability. Generative potency refers to the research implications' practical consequences and usefulness in solving problems and achieving the desirable future. Process transparency refers to the transparent documentation of the process that led to the future-oriented theory. Speculative plausibility means that the process of theorising is itself internally coherent and is based on scientific knowledge. Finally, plausible desirability refers to the research being grounded in shared values to ensure theorising of futures that are desirable by the intended audience. In addition to these elements, Hanisch (2024) posits the importance of theorising prescriptive instruments, which provide paths for how the imagined futures could be realised.

As a critique towards prospective theorising, Wickert (2024) emphasises the complementary nature of descriptive and prescriptive theorising and the importance of their synergy in addressing complex issues, warning of the dangers of de-contextualised, overly simplified prescriptive theories when applied to multifaceted problems. Achieving widespread adoption of a disruptive technology in the likes of blockchain can be considered such a complex issue. We, therefore, inspect the cumulated body of knowledge on blockchain applications through the lens of prospective theorising and evaluate the speculative rigour of this field and its imagined futures with the evaluative criteria proposed by Gümüsay and Reinecke (2024). In doing so, we also adopt Wickert's (2024) perspective of the potential pitfall of prospective theorising producing de-contextualised imaginaries as we examine the gap between academic sentiments and practical adoption of blockchain applications.

3.0 Methodology

We employ the paradigm of post-positivism, which allows us to take a structured, empirical approach while acknowledging that our findings are not a representation of an absolute truth but rather a reflection of the data gathered and analysed. We seek to answer our research questions by conducting a meta-analysis of academic review articles on blockchain applications across disciplines. By employing the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework for article selection (Figure 1) and a mixed-methods approach to data extraction and analysis, we aim to generate insights on the state-of-the-art of blockchain adoption.

Our dataset is constructed by systematically extracting data from the sample of 126 review articles. First, we extracted descriptive information about the publication, which included the title, authors, publication year, journal name and the discipline of the journal. Following this, we extracted qualitative data by extracting their abstracts and conclusions. Next, we extracted descriptive information from each article. This included keywords and phrases used to describe the adoption stage of the blockchain applications (e.g., '*potential*', '*theorised*', '*designed*', '*pilot project*', '*adopted*'), descriptions of all the mentioned blockchain use cases (e.g., supply chain traceability, electronic health records) and the main adoption challenges mentioned in each publication (e.g., '*regulation*'). Finally, we extracted the names of all mentioned real-world blockchain applications, research projects and pilot studies (e.g., TradeLens).

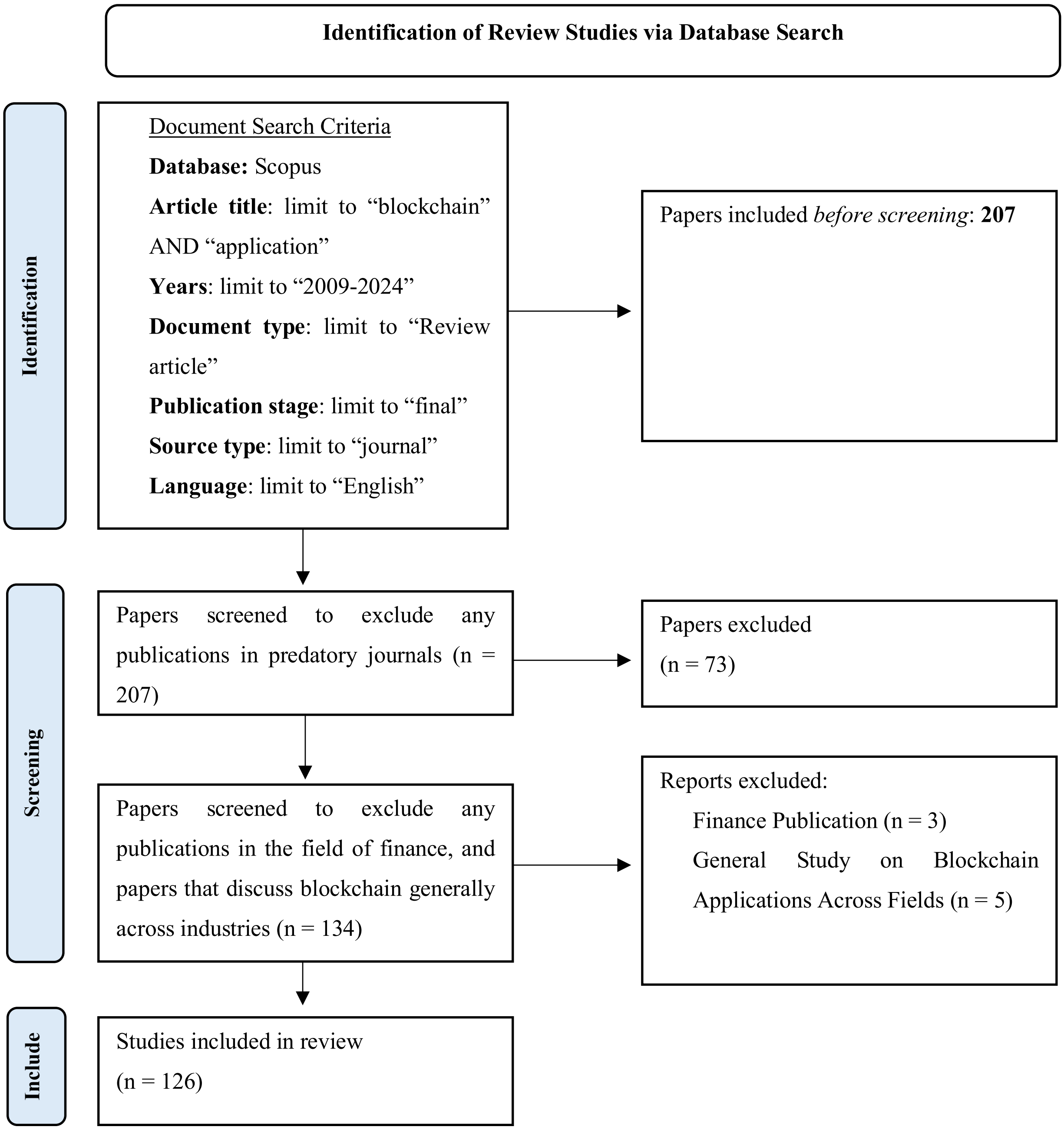

Identification of Review Studies via Database Search
Identification
Document Search Criteria
Database: Scopus
Article title: limit to “blockchain” AND “application”
Years: limit to “2009-2024”
Document type: limit to “Review article”
Publication stage: limit to “final”
Source type: limit to “journal”
Language: limit to “English”
Papers included before screening: 207
Screening
Papers screened to exclude any publications in predatory journals (n = 207)
Papers excluded (n = 73)
Papers screened to exclude any publications in the field of finance, and papers that discuss blockchain generally across industries (n = 134)
Reports excluded:
Finance Publication (n = 3)
General Study on Blockchain Applications Across Fields (n = 5)
Include
Studies included in review (n = 126)

Figure 1. **Data Collection**

Our data analysis will incorporate a mixed-methods approach. To answer our first research question, we will conduct a temporal analysis of blockchain adoption stages. The frequency of articles that describe blockchain applications in each of the pre-determined adoption stages will be tracked for the period 2009-2024. Time-series analysis will be used to identify trends and shifts in blockchain application maturity over time. To extend our findings, the adoption stages of blockchain applications will be compared across disciplines (e.g., healthcare, supply chain) to identify progress per sector. Thematic analysis will be conducted on the descriptions of blockchain use cases and the described adoption challenges to uncover patterns and recurring themes. This will enable us to contextualise how blockchain is used, which applications have gained the most traction, and why its adoption is slower within certain sectors.

To answer our second research question, we will analyse the real-world blockchain applications that are named in the sample documents. Each named application will be examined to determine its current operational status. This will involve web searches, database queries, and project updates to investigate if the application is i) operational, ii) discontinued, or iii) at a more advanced stage. A descriptive analysis will be used to present the findings. Finally, we will conduct sentiment analysis on the extracted abstracts and conclusions, which will provide us with insights into the overall academic sentiment on blockchain applications.

4.0 Anticipated Findings

The following anticipated findings have been derived from our initial analysis. First, studies on blockchain share a generally optimistic sentiment, yet, to date, descriptions of applications remain mostly theoretical. Research on blockchain applications largely focus on describing their potential and imagining prospective futures sparked by widespread adoption. However, certain disciplines, such as computer science and healthcare, describe comparatively advanced adoption stages, with more mature and tested applications emerging. Notably, many blockchain initiatives have faced setbacks, with TradeLens exemplifying a broader trend of discontinued projects. Furthermore, the challenges impeding blockchain adoption are consistent across industries, indicating common obstacles in realising blockchain's potential in practice.

Despite evidence of adoption challenges, the literature on the topic maintains a positive sentiment and a predominantly prescriptive focus. Such a contrast between academic sentiment and practical adoption trends could indicate that the prospective theorising of blockchain applications is de-contextualised, and the imagined futures are not grounded in real-world evidence. This indicates, that while prescriptive theorising can be useful in the initial stages of researching disruptive technologies and imagining desirable futures, research must eventually shift to include more descriptive elements. Incorporating descriptive approaches could also enhance the speculative plausibility of the imagined futures. Research on blockchain applications exemplifies how the speculative plausibility of a theorised future evolves with time as components of the imagined scenarios become reality. Further prescriptive research on blockchain applications could have diminishing speculative rigour as its plausibility is not informed by signals from the real-world or consider evidence from early adoption.

References

Ahirao, P., & Joshi, S. (2022). Social media users privacy protection from social surveillance using Blockchain Technology. *IEEE Bombay Section Signature Conference (IBSSC)*.

Ellinger, E. W., Gregory, R. W., Mini, T., Widjaja, T., & Henfridsson, O. (2023). Skin in the game: The transformational potential of decentralized autonomous organizations. *MIS Quarterly*, *48*(1), 245-272.

Fortune Business Insights. (2024, October 7). Blockchain technology market size, share & industry analysis by component, type, application, deployment, industry, and regional forecast 2024-2032, https://t.ly/yZLnC

Gartner. (2019, October 8). Gartner 2019 Hype Cycle shows most blockchain technologies are still five to 10 years away from transformational impact, https://t.ly/K6Qmg.

Gümüsay, A. A., & Reinecke, J. (2022). Researching for desirable futures: From real utopias to imagining alternatives. *Journal of Management Studies*, *59*(1), 236–242

Gümüsay, A. A., & Reinecke, J. (2024). Imagining desirable futures: A call for prospective theorizing with speculative rigour. *Organization Theory*, *5*(1), 1–23.

Hanisch, M. (2024). Prescriptive theorizing in management research: A new impetus for addressing grand challenges. *Journal of Management Studies*, *61*(4), 1693–1707.

Hasselgren, A., Rensaa, J. H., Kralevska, K., Gligoroski, D., & Faxvaag, A. (2021). Blockchain for Increased Trust in virtual Health Care: Proof-of-Concept study. *Journal of Medical Internet Research*, *23*(7), e28496.

Hawlitschek, F., Notheisen, B., & Teubner, T. (2018). The limits of trust-free systems: A literature review on blockchain technology and trust in the sharing economy. *Electronic Commerce Research and Applications*, *29*, 50–63.

Jensen, T., Hedman, J., & Henningsson, S. (2019, December). How TradeLens delivers business value with blockchain technology. *MIS Quarterly Executive*, 18(4), 221-242.

Leow, A., & Litan, A. (2024, July 29). *Hype Cycle for Web3 and Blockchain, 2024* (ID G00811967). Gartner.

Maersk (2022, November 29). A.P. Moller-Maersk and IBM to discontinue TradeLens, a blockchain-enabled global trade platform, https://t.ly/QYy4.

Parry, G., Revolidis, I., & Ellul, J. (2023). Bottling Up Trust: A review of blockchain adoption in wine Supply chain Traceability. *Social Science Research Network*.

Rani, P., Sharma, P., & Gupta, I. (2024). Toward a greener future: A survey on sustainable blockchain applications and impact. *Journal of Environmental Management, 354,* 120273.

Sarker, S., Henningsson, S., Jensen, T. W., & Hedman, J. (2021). The use of blockchain as a resource for combating corruption in global shipping: an

Interpretive case study. *Journal of Management Information Systems*, *38*(2), 338–373.

Wickert, C. (2024). Prescriptive theorizing to tackle societal grand challenges: Promises and perils. *Journal of Management Studies*, *61*(4), 1684–1691.

Research approaches to qualitative studies; the case of crisis management in Higher Education women academics

Maria Vardaki
Royal Holloway University of London, Matm015@live.rhul.ac.uk
Eleni Tzouramani
University of the West of Scotland, Eleni.Tzouramani@uws.ac.uk

Abstract

This paper analyses approaches to qualitative research, with the focus on women academics' narrative in the Higher Education in the UK. The authors have considered a number of different research strategies (case studies, ethnographies etc) and research designs (interviews, focus groups, surveys etc), in order to result in the appropriate approach to this particular research. This paper also considers the emerging role of AI tools for qualitative research analysis which can be efficient in thematic analysis and pattern recognition but they can pose challenges in capturing the nuanced, socially constructed experience of women academics. The paper also covers information on research participants and the data analysis process, focusing on thematic analysis for this study.

Keywords: research design, qualitative research, Higher Education, academics, women, research methods

1.0 General Introduction

Amongst the academic disciplines, social science covers a variety of disciplines, such as psychology, anthropology, political science, business studies (Kuper & Kuper, 2003). It is believed that any research analysis to any one discipline can be applied in other disciplines (Richardson & Fowers, 1998; Knoll et al, 2018). The debate that started in the 19th century reveals opposite opinions about what social science knowledge means, as researchers believe that it is about facts and not values whereas others believe in the authenticity of feelings and meaningful relationships (ibid, 1998). With so many technological advances and as our scientific knowledge progresses from year to year, we are all exposed to a vast number of tools, in order to conduct any type of research. Both empirical and academic research are engaging in the investigation of a subject, in gathering data, acquiring new knowledge and come to new conclusions, as a result of this research. There are a number of data gathering methods, and many argue that this is depended upon the research question itself (Saunders et al, 2012). Others (such as Guba and Lincoln, 1994) believe that "both qualitative and quantitative methods may be used appropriately with any research paradigm" (ibid, 105).

According to Kuhn (2012) a research paradigm is a set of common beliefs that are shared by researchers about how "problems should be understood and addressed". Saunders et al (2012) use the so-called "research onion" as a way to present the data collection methods right in its centre with the research philosophies and approaches, as well as methodological choices, research strategies and the time horizon of the research. The way researchers conduct social research is depended on the "plethora of choices in the design" (Lewis-Beck et al, 2003), which incorporates the research questions and problems, the strategies on how to investigate the problems and approaches to these strategies, the theoretical framework, the data and data collection methods, as well as the data analysis methods (Blaikie, 2000).

2.0 Epistemological and Ontological aspects of the research

The philosophical position of epistemology refers to the knowledge around us and how we acquire such knowledge. In the cases where the researcher focuses on the collection and analysis of facts, the method that is used is mainly through gathering quantitative data. The researcher can adopt a positivist position, as the focus is placed on "objects" or data and not human beings; this can be seen in the manufacturing, or the financial sector. The researcher focuses on an objective approach by collecting quantitative data, via survey questionnaires, published financial records, or from company annual production reports. This position is not applicable to this research, as the focus is not the object, the data collected by the researcher, but on the human beings, specifically women in academia.

On the other hand, the researcher who focuses on the human aspect of the data collection such as human resource processes, staff development, resilience and effects on human behaviour, performance management, career progression, staff motivation, via different methods (Lut, 2012), would concentrate on the importance of the narrative and the anthropological aspect of the research. The people's feelings cannot be measured or be quantified, so the researcher is often adopting an interpretivist position, by collecting qualitative data, through methods such interviews. In this case, the data are subjective and open for interpretation by the researcher. This method is the most applicable to this research, as it is focusing on the investigation of the perspective of women in academia during the transitional period from face-to-face to online learning and the mechanisms that they have used and are using, in order to endure the challenges of work and everyday life.

According to Crotty (1998, 10), “ontology is the study of being. It is concerned with 'what is', with the nature of existence, with the structure of reality as such. Were we to introduce it into our framework, it would sit alongside epistemology informing the theoretical perspective, for each theoretical perspective embodies a certain way of understanding what is (ontology) as well as a certain way of understanding what it means to know (epistemology)”. From an ontological point of view, the nature of our reality and how the world operates (Saunders at al, 2012), the researcher can assume either an objectivist or a subjectivist position, depending if the social actors are affecting our reality or not. For the purpose of the current research, we have to assume a subjectivist position, as the social actors operate within the real world and the researcher belongs to the same professional environment. We also approached the subject with open mind and let the data guide the themes that resulted from the research.

3.0 Research philosophies

3.1 Interpretivist

According to Saunders et al (2012) “Interpretivism is an epistemology that advocates that it is necessary for the researcher to understand the differences between humans in our role as social actors”. In order to assume an interpretive approach, the researcher considers the scientific knowledge too complex to be “reduced entirely to a series of law-like generalisations” (Gill & Johnson, 2010). Saunders (et al, 2012) make the distinction that the research is conducted amongst people and not objects, which is why humans are referred to as actors, as they are playing in the “stage of human life”. He believes that actors interpret the role they play and in the same way humans are playing a specific role in their daily social life (ibid, 2012).

For the purpose of this research, we assume an interpretivist role, in order to focus on the participants of the research, as this approach constantly evolves and does not remain static. The fact that the focus is on the interpretation of the data that are collected as we aim to explore women's degree and perceptions of resilience during the rapidly changing HE environment under the era of the pandemic. The interpretation of factors that enable and constrain resilience in women within HE is also focusing on the position of an interpretivist, as we draw a comparative analysis between the factors of resilience.

It is an integral part of the human nature to be able to receive information from others and interpret their behaviour according to their own disposition. In case of an event being witnessed by a number of people, the account they will provide is likely to be different, although the event is the same. This behaviour is explained by two intellectual traditions, phenomenology (the way we understand the world) and symbolic interactionism (how we interpret the world), (ibid, 2012).

One of the characteristics of the interpretivist approach is that the researcher will have to be empathetic and try to understand the research subjects from their point of view instead of just treating them as collection data (Saunders et al, 2012). This is particularly important in disciplines such as business and management, especially in human resources, staff development and organisational studies, as opposed to physical social science. The main difference in this comparison is the human factor that can determine people's behaviour and can provide differences in the data provision of a research theme. An aspect of an individual's working life can provide people with a vast number of differences in their professional career, such as differences in work patterns (Giannikis & Mihail, 2011). Ethnography is often important in this approach, in order to understand the culture of a group or organisation. On the other hand, ethnography cannot be used by an interpretivist approach in this study, as it is not relevant to the objectives of this type of research, which focused on individual narrative and individual perspectives of a situation.

3.2 Objectivism

According to Saunders et al, "objectivism is the philosophical position which holds that social entities exist in reality external to social actors" whereas the subjectivist view focuses on social phenomena which "are created from the perceptions and consequent actions of social actors". Bernstein (1983, p. 8) defines "objectivism" as the "basic conviction that there is or must be some permanent, a historical matrix or framework to which we can ultimately appeal in determining the nature of rationality, knowledge, truth, reality, goodness, or rightness."

Saunders (et al, 2012) refers to management studies as an example of a potential objective entity. In this case, each organisation has a set of rules and regulations, an organisational structure, a reporting procedure and all employees have their duties set by specific job descriptions. This could arguably form an objective entity. From this point of view, all similar organisations should operate in a comparable way. The fact that similar positions are held by different people, this provides a subjectivism that differentiates both the organisations and the specific posts. An objective research approach is normally found in physical science due to the nature of the research question. Objectivist approach can be viewed closer to a positivist position in social science research.

General characteristics of the objectivist approach entail the view of the world as external and objective, with the researcher being independent of personal values and beliefs, as these are not reflected in the research outcome. The researcher is using quantitative methods to gather facts. A deductive approach is used, in order to explain and prove certain theories. Heron (1996) argues, though, that values are the guiding force of all human beings and even the choice of the researcher’s philosophical approach demonstrates their personal values, beliefs and traditions. The importance of axiology as a philosophical disposition comes against the objectivist approach to the research subject.

Additionally, organisational culture is viewed differently depending on which approach the research is following. According to Smircich (1983) the objectivist approach sees the organisation as having a set culture but the subjectivist approach sees the organisation as a combination of what the organisation is through social interaction with its actors (humans). The differences between the objectivist and subjectivist approach is the number of factors, both social and physical, that affect the culture, which, subsequently, can affect a research conducted within that organisation. This research being conducted within Higher Education, under a crisis situation, focuses on a subjective approach, due to the varied conditions of work that women in academia face. Home working is a condition that can be seen in a completely different way by different academics, depending on their professional situation, family situation and their personal view point.

3.3 Realism

The question of “what exists in social reality?” is, according to Potter (2000) the focus of realism. Macquarrie (1973, 157) mentions that “'If there were no human beings, there might still be galaxies, trees, rocks, and so on-and doubtless there were, in those long stretches of time before the evolution of Homo sapiens or any other human species that may have existed on earth'. So reality exists without human beings, but it is those human beings and their social interactions that constitute the meaning of reality. According to Crotty (1998, 10), “realism (an ontological notion asserting that realities exist outside the mind) is often taken to imply objectivism (an epistemological notion asserting that meaning exists in objects independently of any consciousness)”. Social network analysis, though, provides evidence that show both positivist and critical realist views, depending on the methods used to analyse (Buch-Hansen, 2013). Extending critical realism to ethnography, Watson (2011) explains that in this case we look to answer “how the social world works”, taking into account that the people’s narrative is only the starting point of the research. To extend this type of research, events tend to be explained through the influence of human beings. This is a relevant way of explaining the current research undertaken, as the focus is on the human beings (women in academia) and their reality through a specific challenging event (pandemic). The perceptions of women and their narrative through the crisis is the main focus but the interpretations of resilience will draw the results of the research.

3.4 Pragmatism

In cases where the focus is the investigation of the reality in different ways, the researcher adopts a pragmatist approach, so different positions may be more relevant to different research questions, depending on the nature of the question. By adopting a pragmatist approach, the researcher can focus on which specific philosophy is suitable to the research, and this will allow the researcher to work with different philosophical approaches (Saunders et al, 2012; Tashakkori & Teddlie, 1998). The fact that different approaches can be used in one study provides the basis for different approaches to be combined. As a pragmatist believes that there are multiple realities, this could also provide grounds that a particular research question can be viewed by opposite approaches. In spite of this, though, it is important to use the methods that can provide credible and reliable data for the research output (Kelemen and Rumens 2008). The current research is conducted with the focus of exploring different philosophical approaches and applying the most relevant one. As the focus of the research is based on the human perspectives and women's feelings and approaches to the pandemic, a pragmatist approach will not be appropriate to use.

4.0 Research Design

4.1 Reflexivity

Identifying a way to understand how interrelated the above philosophical positions are, and how the interpretive and objectivist approach, for example, are affecting each other, the researcher has to consider the notion of reflexivity; how the researchers understand themselves "through thinking about their own thinking" (Johnson & Duberley, 2003). It is assumed that the researcher monitors their "behavioural impact upon the social settings under investigation" and, in this way, the methodology is validated and the research findings are not affected (ibid, 2003). The involvement of the researcher is removed, thus allowing objectivity in the research outcomes. On the other hand, the more the researcher is thinking about personal beliefs, the more this can reveal deeper and hidden dogmata. This will assist the researcher, though, to better understand themselves, so they will be able to address the research question fully. The two approaches, although different are interrelated through this research process.

Richardson & Fowers (1998) mention a different aspect of the study of social science, due to the fact that it combines theory and practice as opposed to other disciplines. They argue that the theory aspect of social science knowledge tends to be more "deterministic and reductionistic" whereas practise entails some kind of "freedom and creativity". This aspect is seen by Bernstein (1983) as combining a positivistic outlook, where a version of reality is given in a neutral way. The reflective manner of conducting the research is highly considered as the researcher will have to consider their own behaviour and how this is impacting on the participants, their own opinions and experiences through the pandemic, and their own resilience through certain experiences, in order to better understand the participants. The focus will have to subsequently remove personal opinions and interpret the data collected with objectivity and accept the version of reality for each participant. we understand and accept the fact that we are part of the research as we belong to the same sector as the interview participants and we have also faced similar challenges as the participants, which can also be seen as a limitation of this study. Having recognised this as a fact, we went into the research interview and analysis having accepted an element of subjectivity throughout the process.

4.2 Research Strategies

4.2.i Case studies

The usage of case studies provides flexibility for the researcher, as it can be conducted in any stage of the research, it captures reality/real life and can combine other methods, such as interviews or observations (Curtis et al, 2014). The usage of multiple case studies can test theoretical constraints and test a theory through repetition of study (Eisenhardt, 1989) but Flyvbjerg, (2006) believes that a single case study can provide significant and in-depth knowledge of a context or a situation or a group of people. Furthermore, some single case studies in the past have become paradigmatic cases, such as “Fordism” (Mullin, 1982). It is also argued that case studies may be popular as it is easy to predict a situation (Merriam, 2009).

On the other hand, case studies tend to be used to generalise contexts and situations (Smith, 1996 & Mjoset, 2006), which can lead into assumptions about a whole industry. Additionally, it is argued that they offer a convenient way of research but they provide only a “story” of the research subject (Sulaiman & Burke, 2009), which demands interpretation and explanation of phenomena, and critical review of the story and not just the narration (Curtis et al, 2014). A case study is depended upon the researcher’s bias, ethical stance and subjective approach due to the dependence in the choice of data to analyse, include and publish in the research output (Stake, 2005). We have considered the option of conducting a case study for this research but we have decided that this will not provide us with the in depth knowledge and perspective of women in academia, as the focus their narrative and not the case study. As the focus is on their perspectives of resilience and how they are coping during a crisis, a case study would not provide the relevant results for such research.

4.2.ii Ethnographic studies

Geertz (1988) has introduced the term "thick description" which could be the result of ethnography. This is not the means to achieve generalisations but the researcher looks into the subjects form their point of view with the observer assisting in this result by participating in the conversation. Geertz has also referred to the concept of spending a substantial amount of time observing the participants, in order to understand in depth what the narrative is as opposed to ask them what they mean. It is necessary to decode the participants' attitude and behaviour but not "go native" and become part of the group. The concept of triangulation can be used in this case, as multiple data sources can be used in this research.

Using interviews as a method to understand participants' experiences within ethnography will have to be conducted in a natural and informal way. The initiation of open or semi-structured questions as opposed to abstract ones will lead to a narrative of people's perspectives and experiences without restricting them and directing their responses. The participants will provide answers according to their perspectives and not response with what they think they should respond.

Due to the crisis and the restrictions in movement and travel, digital ethnography or netnography would be an appropriate method to follow. The term netnography has been related to cyber, virtual or online ethnography terms, although there are some distinctions made between them (Costello et al, 2017). In a wider concept, conducting online ethnography is related to using methods both online and offline and observations in virtual and offline spaces, as well as analysing behaviour and narratives of both online and offline settings (ibid, 2017). The main aspect of this method is that the researcher is identified and participates in observations of online communities without becoming part of those communities. This has been considered for the purposes of this study, but we found that such a method would be restrictive in terms of results, as the focus is on the individual women academics and not online academic communities.

4.3 Research Methods

4.3.i Survey Questionnaire

In surveys, large scale responses are more useful for any research question and the researcher can identify how a large part of the population responds to a particular opinion or belief. In essence, this is a low cost method to collect and analyse large scale data from a specific sample of the population without being impacted by the researchers' own beliefs and viewpoints, in a way that qualitative methods may be. This, on the other hand, is depended on how the questions are constructed, and if they are lacking from the researcher's bias. The sample does not necessarily represent the broader population, so the method selection of the sample will need careful consideration. In addition, the participants may respond with their biases coming thought the responses, or nuances may not be captured, as the narrative is missing, so the emotions are also not captured through the survey. From this point of view, this approach on its own would not be sufficient to look into our current research aims, as the focus is the participants' narrative, so it would need a qualitative element of research.

4.3.ii Focus groups

The purpose of using focus groups by researchers was not the primary purpose of their existence. Focus groups have moved from being used by advertising agencies to become an easy way (from the practical point of view) to source information from participants by using minimal effort (Morgan, 1993). This method cannot be used on its own but as a supplementary method combined with interviews or observations to provide a further narrative in participants' experiences. The researcher will need to participate in the focus group by always prompting and asking for clarification examples so the participants can share their experiences, but also feel comfortable and willing to share in order to avoid conformity and polarisation. Various techniques can be used in this occasion such as funnelling sequence questions (Morgan, 1993), which allow the researcher to move from a generic topic to a more specific one, so that the dynamic of the group is understood further.

The issue of homogeneity is also important as the focus group provide useful observations if variety exists. Ethical issues would have to be dealt with in advance, in order to avoid generalisation and superficial statements without necessarily reaching a successful result from the group. We have considered using this method, as a supplementary means of getting further results for the study, but the aim of the research is to focus on the individuals and not identify common problems for a group of women in academia, which has already been identified.

4.3.iii Observations

The researcher can also use observations either during interviews, focus groups or in the setting of the participants, such as teaching or training facilitation. The objectivity of the researcher is crucial for this method, in order to understand the subjects; opinions from the researcher will need to be avoided during this process but the researcher must focus on note taking as a reminder of the observations.

In observations, the researcher can draw a map of the location, the objects of importance and where things are, if this takes place in the personal setting of the participants or a classroom environment. The pandemic has created an additional hurdle for this method as teaching and training took place virtually, and practical or technical restrictions could jeopardise observations. It is worth noting that the researcher would need to reflect afterwards and be aware that the results of their observations may differ and the need to ensure validity of the data, integrity and objectivity in all stages is crucial, while accepting the personal opinions and interpretations of the written narrative from the observations conducted. In order to better understand the data and analysis of the fieldnotes, hermeneutics cycle can be used in the analysis stage (Palmer et al, 2010), so this will minimise any issues of subjectivity. As we have decided to concentrate on semi-structured interviews, we have applied this method during the interviews and reflect our observations as an additional support mechanisms during the data analysis.

4.3.iv Interviews

The use of interviews in trying to identify answers to specific research questions is useful when the researcher is looking to understand the narrative and the reasons behind a situation or people’s actions. Interviews produce rich, in-depth data and analysis can be achieved by identifying themes and trends. The benefit of online interviews focuses on time constraints, and transcription challenges. The drawback, in this case, is the loss of face-to-face contact with the interviewee, so the researcher is unable to identify and interpret body language or interact further in semi-structured interviews. The same applies for skype interviews (Deakin & Wakefield, 2014), where the researcher can overcome geographical restrictions and use it as a low cost method avoiding risks in being hosted at an unknown environment, but have the flexibility of the location choice while maintaining some kind of face-to-face interaction, although building rapport with the interviewee may be challenging.

In addition, interviews cannot reach a large population in the same way as surveys do. The data analysis can be impacted by the researcher's own beliefs and bias, and the interviewees are not representing the broader population. The participants' interaction with the researcher may affect the responses as their biases could come through in their responses, or the researcher's reaction may change their responses. Another aspect that has to be considered is the issue of language barriers, differences in contexts (political, scientific), cultural and social differences through diversity in experiences and values and natural adaptation of research methods, such as usage of ethnographies (Flick, 2014).

Furthermore, we attempted to avoid focusing on a specific type of participants only, as we believe that a more robust picture could be shown, if a number of participants could be included from different stages of their career as well. Occasionally, large set of variables may create issues in the research results, in the sense that they may show a relationship with each other but they do not have a correlation. In this case, the researcher will need to address the issue if this was coincidental or not. In order to ensure reliability and validity, the researcher has to ensure that the sample is not homogenous, so it covers a mixture of individuals from different backgrounds and, as previously mentioned, at different stages of their professional career.

We have interviewed 33 women academics, using the snowballing technique, and convenience samples. The sample included women across different academic disciplines (figure 2), and our aim was to have a representative (diverse) sample, based on different seniority levels, from early to middle and senior academics (Figure 1). We have included a mixture of ethnicity backgrounds, women academics on different type of contracts, such as permanent or fixed term contracts, full time and part time and in a variety of family situations (figure 3), as we wanted to gather data, both from academics with caring and non-caring responsibilities, so the sample would be more robust. We used semi-structured interviews, with the aim to last around one hour in duration, but aimed between 60 to 90 minutes of length. On two occasions, we had to reschedule a second part of the interviews with two academics, as we run out of time (in the first instance) and we had connectivity issues (in the second instance).

The interviews were conducted virtually, using MS Teams, due to the constantly developing situation at the time of the research (pandemic), as to not expose any interview participants at risk. We have ensured confidentiality which fully met during the online interviews and consent was sought in advance of the interview. We have constructed a preliminary list of interview questions, which was adapted during the interview process, in several cases, in order to reflect the application of semi-structured interviews according to individual circumstances.

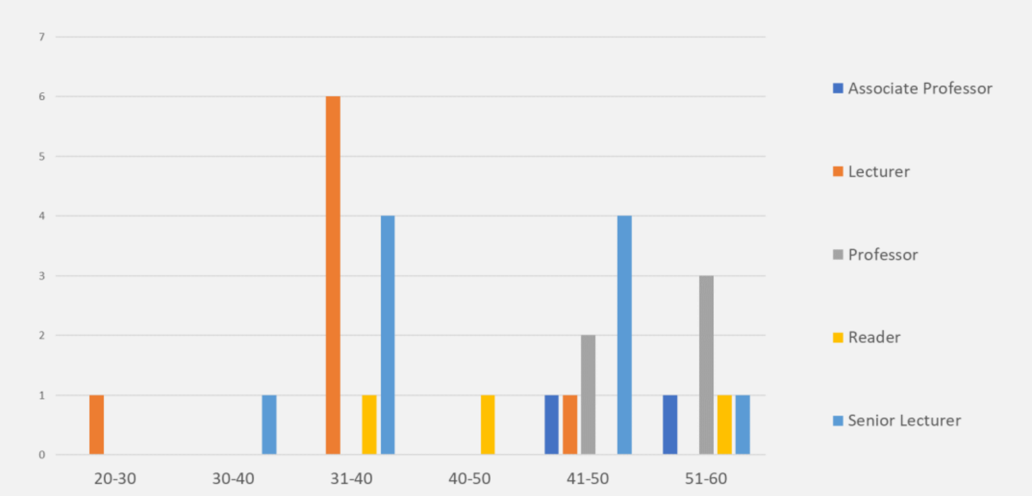


Figure 1. **Seniority Levels of women academics in this research**

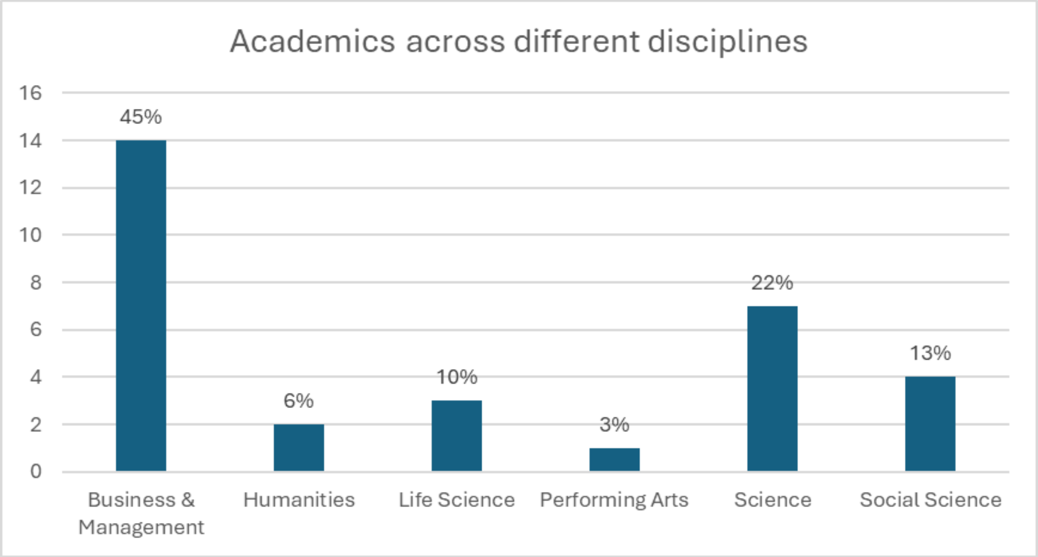


Figure 2. **Women academics across different disciplines in this research**

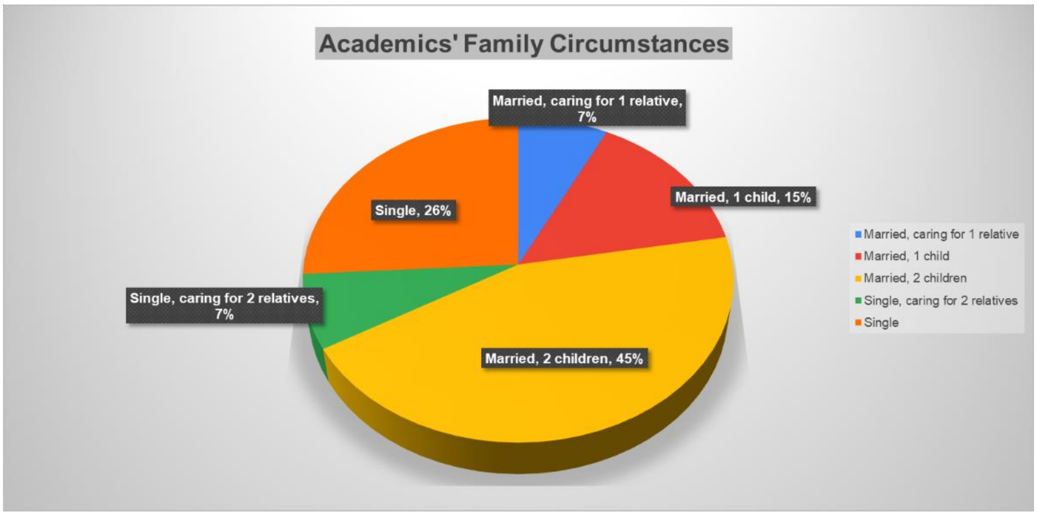


Figure 3. Women academics' family circumstances in this research

One of the main advantages of the semi-structured interviews is that it allows the flexibility to deviate further from the preliminary questions and, depending on the interviewee responses, the researcher can ask further questions to get in-depth knowledge of the subject. This is what is required in this study, as opposed to using structured or unstructured interviews (Saunders et al, 2012). Using structured interviews, we would not have the flexibility to explore further the interviewee narratives and identify potential themes. Using unstructured interviews, it would not have been possible to acquire knowledge of the main objectives and questions of this study, as this method is used as a narrative for interviewees to tell their stories, so this flexibility is not helpful for the type of research.

4.4 Mixed Methods approach

According to Creswell (2014), mixed methods research can combine or integrate qualitative and quantitative research techniques, in order to "neutralise the weaknesses and bias of each form of data" (p.43, 2014). Researchers recognise the limitations of both qualitative and quantitative research techniques. In case where the research can apply both techniques, and if the research questions allow for this application, the researcher can adopt the appropriate research philosophy, in order to be able to integrate the techniques of the data collection and the data analysis (Saunders et al, 2012). In this case, with the focus on the research question, the researcher can adopt a pragmatist approach. This would normally allow for research philosophies to be applied to different research subjects, depending on suitability (Saunders et al, 2012; Tashakkori &Teddlie, 1998). Tashakkori &Teddlie (1998) believe that the adopted philosophy is a continuum that “at some points the knower and the known must be interactive, while at others, one may more easily stand apart from what one is studying” (ibid, 1998, 26).

This type of flexibility allows for multiple realities which would assist in using opposite approaches to validate your data and support reliability. As Kelemen and Rumens (2008) believe, it is important to use techniques and approaches that can provide credible and reliable data for the research results. Charmaz (2006) believes the research study to be a “journey”, but it depends on our understanding and our conclusions through this journey, that we reach the research outcome. It is the researcher who decides the type of journey that will embark, in order to acquire the aspiring results.

Considering two research cases as an example, we could potentially look into investigating female academics’ research outputs during the pandemic and their challenges and perspectives during this period which could affect the research outcomes, we could attempt to adopt both approaches, as it depends on the research aims and objectives. In this case, both qualitative and quantitative methods can be combined, as the focus can be on both facts/data and human beings. The combination of the two different approaches can only be achieved when the research question needs it. In addition, any type of research that combines or includes the human element (feelings, beliefs, personal stance of the researcher) is likely to combine both approaches in the research journey. On this occasion, though, the research method focuses on qualitative as opposed to quantitative data collection, as the focus was the narrative of women academics, their perceptions and feelings at a given time.

4.5 Inductive VS deductive approach

Looking at the research from a deductive position, the researcher can begin "....with an abstract, logical relationship among concepts then move(s) towards concrete empirical evidence", (Neuman, 1997, 46). This means that a researcher can begin with a theory or an abstract concept and aim to relate these ideas to the research results and evidence that will be produced via data collection and analysis. On the other hand, researchers can collect data and analyse them fully through an inductive approach using as their basis the evidence they collect and categorise them, in order to then relate them to a theoretical idea, that will be the most relevant to each of the research questions. Neuman (ibid) also believes that researchers can be more flexible in their approach to data analysis and apply both a deductive and an inductive process, but as appropriate at different stages of each research.

One can consider the explanation that can be found in Kolb's diagram (figure 4) which is used by Ali (1998, 79) and refers to the process of induction moving from acquired facts to theories and the process of deduction moving from theories to explanations and predictions through facts acquired. This is a useful and practical diagram that was used in this research, as we opted for an inductive approach by gathering data/narratives from the participants and looking to identify common areas, explanations and theories that can be, subsequently, tested for their validity by considering the most appropriate data analysis approach.

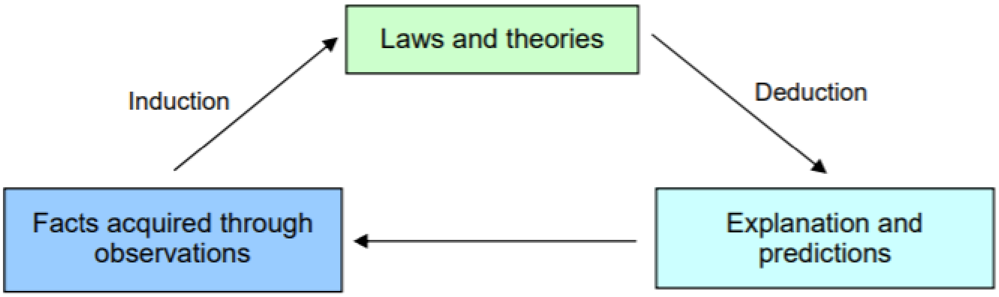

Laws and theories
Induction
Deduction
Facts acquired through observations
Explanation and predictions

Figure 4. Kolb's diagram – induction process (source: Ali, 1998, 79)

As Ali has referred to the deductive approach requiring "the development of a conceptual and theoretical structure prior to its testing through empirical observation, corresponding therefore to the left-hand side of Kolb's experiential learning cycle" (Ali, 1998:5), it would be most appropriate that the starting point of this research is the narrative of the participants (on the corner left hand side of Kolb's diagram). In this research, the focus was to apply an inductive approach to the study and look into the literature review with a deductive thinking, as the research will inevitably exhibit studies that can be comparable, even from a variety of sectors, to the current research study.

5.0 Data Analysis

During the data analysis stage, the researcher can adopt an interpretivist approach, considering a number of aspects regarding the validity of data and its uniqueness. As organisations differ amongst different sectors but also within the same sector (Munaf, 2009; Dirani, 2009), this has to be considered in the data analysis stage. As an example, the 2014 CIPD research report considers the focus that seven different organisations have with its social media usage, and it is evident from the research outcome that all seven organisations have different focus.

The research outcome is also depended on the timescale of a research. Evidently, the same quantitative or qualitative research would not produce the same results, if it is conducted longitudinally. A longitudinal survey (Ployhart & Ward, 2011) in the perceptions and resilience of women academics before, through and after the pandemic would provide a more in-depth knowledge for the way women operate within Higher Education and under a particular setting, as well as how the pandemic has changed, if at all, the way educational institutions support women according to their specific circumstances. Within the same organisation, timing of a research makes a difference in the production of the outcome, either because of the natural changes that occur within the organisation, or due to outside parameters (Knight, 2012). This demonstrates the complexity of social science research, which is accepted in an interpretivist approach (Sanders et al, 2012).

As we have used an inductive approach, a thematic analysis of qualitative data (Boyatzis, 1998; Braun & Clarke, 2006) is suitable for qualitative methods, although it can also be applied in quantitative and mixed methods approach. This can be used both by an objective and interpretive approach. Tashakkori &Teddlie (1998) view the adopted philosophy as a continuum and they believe that "at some points the knower and the known must be interactive, while at others, one may more easily stand apart from what one is studying" (ibid, 1998, 26).

For the analysis of the interview data, we have used the six phases of thematic analysis (Braun & Clarke, 2006), as this was an appropriate method for qualitative data analysis and most appropriate for interviews. For this reason, we familiarised ourselves with the data and, subsequently, generated an initial coding system. We have looked for different themes within the data and, then, reviewed the themes again for clarity and ambiguity avoidance. Our intention was to define the different themes and concluded on naming them, in order to produce the final results. We have also considered, created and developed a colour coding system, in order to achieve the objectives of the research, but this was not as helpful, as we had initially considered. The intention was to group similar themes according to different colours and search for commonalities, but it was not possible in the end to fully capture all themes with a colour coding system.

We have considered using a discourse analysing but the focus of the research is not on the language of the participants and how it is used in their narrative but the narratives themselves. As it is explained by Wood & Kroger (2000) discourse analysis is "a perspective on the nature of language", as well as the connection to social sciences matters. The authors also see discourse analysis as approaches to the narrative that include "theoretical assumptions" in addition to data collection and analysis practices. The main difference between discourse analysis and other qualitative data analysis methods is the interpretation of the reality that exists as opposed to the way that was produced, which discourse analysis focuses on (Phillips & Hardy, 2002).

Machine learning and natural language processing are being increasingly used in thematic analysis, discourse analysis and qualitative coding (Hitch, 2024; Christou, 2023; Gamboa & Diaz-Guerra, 2023) mostly facilitating large scale data processing and identify patterns (Fonseca, Chimenti & Susarez, 2023) as well as in coding large datasets and analysing qualitative data (Lennon et al., 2021). For this study, researching women's experiences of resilience in crisis contexts, the use of AI for analysis poses epistemological concerns. Interpretivist research engages in the analysis of meaning construction which cannot be interpreted though AI analysis of linguistic patterns instead of lived experience and human meaning making (Bano, Zowghi & Whittle, 2024). AI is based on pre-trained algorithms that often reproduce historical biases interpretations (Mehrabi et al., 2021) and hegemonic narratives (Zhang et al., 2023) instead of new insights. Studies on algorithmic bias indicate that AI might prioritise dominant discourses and underrepresent marginalised voices, especially those of women within institutional settings (Buolamwini & Gebru, 2018). Most importantly, AI cannot engage in reflexivity which is a key part of interpretivist studies where the researcher critically engages with their own subjectivity and biases (Shaffer & Lieder, 2022). AI treats qualitative data as quantitative patterns which does not allow for contextual interpretations of participants' narratives. In this study, it was important that a human researcher analysed the socially situated and context dependent narratives of women academics during the pandemic.

The affective and experiential aspects of resilience such as cannot be adequately captured through automated text analysis and AI does not have the required lived experience to interpret the socially constructed meaning and experiences of women academics (D'Ignazio & Klein, 2020). For this reason, the human interpretive analysis is more effective for the depth of theoretical sensitivity (Strauss & Corbin, 2014) and context based analysis (Braun and Clarke, 2006) needed. Although AI can facilitate qualitative research when aligned with the methodological approach, in this study it was not applied so that, women academics' experiences would not be devalued or depersonalised by AI.

6.0 Ethical and Legal Considerations

For the purpose of this research, all legal considerations have been respected as cases of hybrid working may be considered and specific requests may be made on the basis of personal circumstances. Within the spectrum of this research, we have attempted to be considerate to women's personal circumstances that might affect their working patterns and their resilience levels. In fact, a small number of women academics, during preliminary discussions (prior to the actual online interview), expressed worries about their personal circumstances and confidentiality issues; as a result, two academics chose not to be interviewed with a clear explanation for their reasons (not relevant to the interview process).

Job security (Wright et al, 2020) was extremely challenging during the pandemic, as a number of employees faced salary freezes, job losses and furloughing. In a CIPD report conducted since April 2020 (CIPD, 2020a), 25% of the workforce in the UK was initially furloughed in May and only 49% of the employees were working normal hours at that time. This is a sensitive topic on the professional career of women who may have found themselves in furloughing situations or other types of professional setbacks, so this is a topic that was carefully considered with outmost respect and confidentiality. From the interviewees of this research, two academics were particularly worried about this topic and one of them, in fact, had to move jobs in a different institution, due to the fact that her original position was no longer available, due to the pandemic.

Our access strategy was direct communication with the relevant women within the institution and, subsequently to this, communication with further academics through the initial selected interviewees. Ethical issues were considered at different stages of the research (Saunders et al., 2012, 236), but equally focusing on the participants and the interview process, in order to ensure that they are at ease and not affected negatively by the interview process. The intention was to follow the ethical issues by informing all participants of the confidentiality of their information and anonymity, by retaining our integrity and objectivity in all stages of the study and by providing all the information in advance to participants. Introductory communication was arranged directly with the participants, in order to explain the process and allow time to reflect and consider their full participation. According to Gardenier (2011) "ethics, whatever you perceive them to be, must apply to both friend and foe if they are to have any credibility as representing moral values".

In terms of the approach to the participants, the intention was to have preliminary discussions and agree their permission to conduct this research and ensure the participants are familiar with us as researchers. This was primarily conducted via email communication or via short online meetings with the participants. We ensured that confidentiality was applied at all stages of the research and we informed the participants accordingly, ensuring that their identity and personal information would not be compromised or exposed (Kaiser, 2009). We have provided all participants with a consent form in advance of the interview, so they could agree and sign before their participation. We had submitted an online ethical form through the University's website, in order to seek the relevant ethical permissions and comply with the ethical procedures in place for such research. In terms of any participant concerns, we have ensured that all participants were informed that they could withdraw from the process at any time and request their information to be deleted and not to be used as part of the research (Research Ethics Guidebook 2015). We ensured that our questions would not bring any anxiety to the participants, by explaining the topic and subtopics of the questions in advance and we aimed to pause the interview at any stage we felt this was needed, in order to bring the participants back to their initial emotional state. We have attempted, when time permitted, to have an informal conversation with the participants, at the end of each interview, in order to ensure that participants were happy with the interview and ask for their opinion of the interview.

7.0 Limitations

During any research process, limitations can be identified due to a number of reasons, which vary from the researcher's approach to the subject (taking into account the researcher's biases and personal beliefs, as well as the point of view to the specific subject); willingness of the participants (notwithstanding the participants' responses and how truthful they are, in both survey questionnaires and interviews); access to information (as in a case study approach or reviewing of documentation of an organisation, the researcher may not be provided with the necessary and relevant documentation), validity of data and the researcher's own ideas and interpretation of the data gathered, to name a few. The results of the research can also be affected by the level of the author's familiarity with the subject (Boyatzis, 1998) and their personal theoretical positions and values (Heron, 1996; Braun & Clarke, 2006).

The authors will have to achieve a balance in the data manipulation (Boyatzis, 1998), but to also ensure reliability by avoiding assumptions in advance of the data analysis (Saunders et al., 2012). This applies in both qualitative and quantitative data, in different ways. On many occasions, researchers have to rely on any studies or reports publicly available (CIPD, 2020a; CIPD, 2020b; Robinson et al, 2004). This may bring challenges, such as outdated information to rely upon or variations to the conditions of the research, which cannot be known to the researchers. Charmaz (2006) mentions that the research is a "journey", but it depends on what the researchers understand and conclude through this journey, that it becomes a research outcome. This extends to different research outcomes by different researchers, even if they are examining the same subjects.

References

Ali, S. (1998), Research Methodology: Back to Basics, Abac Journal, 75 – 98.

Bano, M., Zowghi, D. & Whittle, J. (2024) AI and human reasoning: Qualitative research in the age of Large Language Models. The AI Ethics Journal, 3(1).

Bernstein, R (1983), Beyond objectivism and relativism, Philadelphia: University of Pennsylvania Press.

Black, I. (2006). The presentation of interpretivist research. Qualitative Market Research: An International Journal, 9(4), 319–324

Blaikie, N.(2000), Designing social research: The logic of anticipation, Cambridge, UK: Polity.

Boyatzis, R. E (1998), Transforming qualitative information: thematic analysis and code development, Sage, USA.

Braun, V and Clarke, V (2006), Using thematic analysis in psychology, Qualitative Research in Psychology, 3, 77-101.

Buch-Hansen, H., (2013), Social Network Analysis and Critical Realism, Journal for the Theory of Social Behaviour, 44, 3.

Buchanan, D, Boddy, D and McAlman, J, (1988), Getting in, getting on, getting out and getting back, in Bryman, A. (ed.), Doing Research in Organisations, London, Routledge, 53–67.

Buolamwini, J. & Gebru, T. (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81:1-15

Carson, D., Gilmore, A., Perry, C., and Gronhaug, K. (2001). Qualitative Marketing Research. London: Sage.

Charmaz, K, (2006), Constructing Grounded Theory: A Practical Guide through Qualitative Analysis, SAGE, London.

Christou, P. A. (2023). The use of Artificial Intelligence (AI) in qualitative research for theory development. The Qualitative Report, 28(9): 2739-2755.

CIPD. (2012), A collection of Thought Pieces: harnessing social media for organisational effectiveness, London: Chartered Institute of Personnel and Development. Available at: http://aspirehrbp.org.uk/wp-content/uploads/sites/51/2016/11/Harnessing-social-media.pdf [Accessed 05/05/2019].

CIPD (2013a). Deterioration in employee voice and employee engagement at record low. Research Report, London: Chartered Institute of Personnel and Development.

CIPD (2013b). Social Technology, Social Business?, Survey Report. Chartered Institute of Personnel and Development. Available from: https://www.cipd.co.uk/Images/social-technology-social-business_2013_tcm18-10323.pdf [Accessed on: 05/05/2019]

CIPD (2018) Pre-employment checks; Guidance for Organisations, Chartered Institute of Personnel and Development, Available at: https://www.cipd.co.uk/Images/preemployment-checks-guide-2018_tcm18-51572.pdf [Accessed 04/05/2019]

CIPD (2014). Putting Social Media to Work: Lessons from Employers, Chartered Institute of Personnel and Development, Available from: https://www.cipd.co.uk/Images/putting-social-media-to-work-lessons-from-employers_tcm18-10319.pdf [Accessed on 05/05/2019]

Costello L, McDermott M-L, Wallace R. Netnography: Range of Practices, Misperceptions, and Missed Opportunities. International Journal of Qualitative Methods. December 2017

Cresswell, J W (2014), Research Design: Qualitative, Quantitative and Mixed Methods Approaches, SAGE, London.

Crotty, M. J. (1998), The Foundations of Social Research: Meaning and Perspective in the Research Process. London: Sage.

Crouse, P, Doyle, W and Young J.D (2011), Workplace learning strategies, barriers, facilitators and outcomes: a qualitative study among human resource management practitioners, Human Resource Development International, 14:1, February, 39-55.

Curtis, W., Murphy, M., Shields, S. (2014). Research and Education, London: Routledge.

Deakin, H., & Wakefield, K. (2014). Skype interviewing: reflections of two PhD researchers. Qualitative Research, 14, 5, 603–616.

D'ignazio, C. & Klein, L. F. (2023). Data feminism. MIT press.

Dirani, K M (2009), Measuring the learning organization culture, organizational commitment and job satisfaction in the Lebanese banking sector, Human Resource Development International, April, 12: 2, 189–208.

Edwards, T and Rees, C (2011), International Human Resource Management: Globalization, National Systems and Multinationals Companies, Pearson Education Ltd, Harlow.

Eisenhardt, K. M. (1989), Building theories from case study research. Academy of Management Review, 14, 532-550.

Flick, U. (2014). Challenges for Qualitative Inquiry as a Global Endeavor: Introduction to the Special Issue. Qualitative Inquiry, 20 (9), 1059–1063.

Flyvbjerg, B. (2006), Five Misunderstandings About Case-Study Research, Qualitative Inquiry, 12, 2, April 2006, 219-245.

Fonseca, A. L. A. D., Chimenti, P. C. P. D. S. & Suarez, M. C. (2023). Using deep learning language models as scaffolding tools in interpretive research. Revista de Administração Contemporânea, 27, e230021.

Gamboa, A. J. P. & Díaz-Guerra, D. D. (2023). Artificial Intelligence for the development of qualitative studies. LatIA, 1: 4.

Gao, J., Choo, K. T. W., Cao, J., Lee, R. K. W. & Perrault, S. (2023) CoAIcoder: Examining the effectiveness of AI-assisted human-to-human collaboration in qualitative analysis. ACM Transactions on Computer-Human Interaction, 31(1): 1-38.

Gardenier, J S (2011), Ethics in Quantitative Professional Practise, in Panter A T & Sterba, S K (eds), Handbook of Ethics in Quantitative Methodology, Taylor & Francis, New York.

Geertz, C., 1988, Thick description: toward an interpretive theory of culture, High Points in Anthropology, 531-552.

Giannikis, S K and Mihail, D M (2011), Modelling job satisfaction in low-level jobs: Differences between full-time and part-time employees in the Greek retail sector, European Management Journal, 29, 129– 143.

Gill, J and Johnson, P, (2010), Research Methods for Managers, 4th ed., London, SAGE

Guba, E G, & Lincoln, Y S (1994),Competing Paradigms in qualitative research, In N.K. Denzin & Y.S. Lincoln (Eds.), Handbook of qualitative research, 105-117,Thousand Oaks CA: Sage.

Hakim, C, (2000), Research Design: Successful designs for social and economic research, (2nd ed) London, Routledge.

Heron, J. (1996), Co-operative inquiry: research into the human condition, London, Sage.

Hitch, D. (2024) Artificial intelligence augmented qualitative analysis: the way of the future? Qualitative Health Research, 34(7): 595-606.

Johnson, P and Duberley, J (2003), Reflexivity in Management Research, Journal of Management Studies, 40:5.

Kaiser K., (2009), Protecting Respondent Confidentiality in Qualitative Research. Qualitative Health Research, 19(11), 1632-1641.

Kelemen, M, and Rumens, N, (2008), An introduction to critical management research, London, SAGE.

Knight, M (2012), Changing Times in UK Universities: what difference can HR make?, The Guardian, 12 May, available from www.theguardian.com [Accessed 26 February 2021]

Knoll, J., Matthes, J. and Heiss, R. (2018) ‘The social media political participation model: A goal systems theory perspective’, Convergence. doi: 10.1177/1354856517750366.[Accessed 02/05/2019]

Kuhn, T S, (2012), The Structure of Scientific Revolutions, 4th edition, The University of Chicago Press Ltd, London.

Kuper, A. and Kuper, J., (2003), The Social Science Encyclopedia, Taylor & Francis, London.

Lennon, R. P., Fraleigh, R., Van Scoy, L. J., Keshaviah, A., Hu, X. C., Snyder, B. L. & Griffin, C. (2021) Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. Family medicine and community health, 9(Suppl 1), e001287.

Lewis-Beck, M., Bryman, A., & Liao, T. (2003). The SAGE encyclopedia of social science research methods. Thousand Oaks, [Calif.] ; London: SAGE.

Lut, D M (2012),Connection between Job Motivation, Job Satisfaction and Work Performance in Romanian Trade Enterprises, Annals of “Dunarea de Jos” University of Galati, Fascicle I. Economics and Applied Informatics, 18:3, available from http://www.ann.ugal.ro/eco/Doc2012.3/Lut.pdf [Accessed 03 December 2018]

Macquarrie, J, (1973), Existentialism: An Introduction, Guide and Assessment, London: Penguin Books.

Martin, J E and Sinclair, R R (2007), A typology of the part-time workforce: Differences on job attitudes and turnover. Journal of Occupational and Organizational Psychology, 80:2, 301–319.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2021) A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6): 1-35.

Merriam, S. B. (2014). Qualitative Research: A Guide to Design and Implementation (Rev. and expanded ed.), Jossey Bass.

Mjoset, L, (2006), A Case Study of a Case Study, Strategies of Generalization and Specification in the Study of Israel as a Single Case, International Sociology, September 2006, 21, 5, 735–766

Morgan, D. L. (Ed.) (1993). Successful focus groups: Advancing the state of the art. SAGE Publications, Inc. https://www.doi.org/10.4135/9781483349008.

Mullin, J. R., (1982), Henry Ford and Field and Factory: An Analysis of the Ford Sponsored Village Industries - Experiment in Michigan, 1918-1941, Journal of the American Planning Association, 41.

Munaf, S (2009), Motivation, Performance and Satisfaction Among University Teachers: Comparing Public and Private Sectors in Pakistan and Malaysia, South Asian Journal of Management, Oct-Dec, 16: 4, 7-28.

Neuman, W. L. (1997), "Social Research Methods, Qualitative and Quantitative Approaches", Allyn & Bacon, Needham Heights, MA.

Palmer, M, Larkin, M, de Visser R & Fadden G, (2010), Developing an Interpretative Phenomenological Approach to Focus Group Data, Qualitative Research in Psychology, 7:2, 99-121.

Phillips, N. & Hardy, C., (2002), What is discourse analysis?, In Discourse analysis (pp. 2-17). SAGE Publications, Inc.

Ployhart, R. E., & Ward, A. K. (2011). The "quick start guide" for conducting and publishing longitudinal research. Journal of Business and Psychology, 26(4), pp. 413-422.

Potter, G. (2000), The Philosophy of Social Science: New Perspectives, Essex: Pearson Education.

Research Ethics Guidebook, (2015), The Research Ethics Guidebook: A Resource for Social Scientists. Available at: http://www.ethicsguidebook.ac.uk. (Accessed 9 January 2022).

Richardson, F.C and Fowers, B. J (1998), Interpretive Social Science: An Overview, American Behavioral Scientist, 1998, 41:4, 465-495

Robinson, D., Perryman, S. and Hayday, S. (2004), The Drivers of Employee Engagement, IES Report 408, Institute for Employment Studies, Available at: https://www.employment-studies.co.uk/system/files/resources/files/408.pdf [Accessed 04/05/2019]

Saunders, M, Lewis, P and Thornhill, A (2012), Research Methods for Business Students, 6th ed., Pearson Education Ltd, Harlow.

Schäffer, B. & Lieder, F. R. (2023) Distributed interpretation–teaching reconstructive methods in the social sciences supported by artificial intelligence. Journal of research on technology in education, 55(1): 111-124.

Smircich, L. (1983), Concepts of culture and organisational analysis, Administrative Science Quarterly, 28: 3, 339-358.

Smith, C (1996), Book Review on "Graham, L. 1995, On the Line at Subaru-Isuzu: The Japanese Model and the American Worker, Cornell University Press" Work Employment Society; 10; 394.

Stake, R.E. (2005), Qualitative case studies. In N.K. Denzin & Y.S. Lincoln (Eds.) The Sage handbook of qualitative research (3rd ed.), (pp. 443-466). Thousand Oaks, CA: Sage.

Strauss, A., & Corbin, J. (2014). Basics of qualitative research techniques. New York: Sage

Sulaiman, N.I.S & Burke, M., (2009), A case analysis of knowledge sharing implementation and job searching in Malaysia, International Journal of Information Management, 29, 321–325

Tashakkori, A. and Teddlie, C., (1998), Mixed methodology: combining qualitative and quantitative approaches, Thousand Oaks, Ca. Sage.

Watson, T.J. (2011) 'Ethnography, reality, and truth: the vital need for studies of "how things work" in organizations and management', Journal of Management Studies, 48:1, 202–217.

Wood, L. A., & Kroger, R. O. (2000), Doing discourse analysis: Methods for studying action in talk and text. Sage Publications, Inc.

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J. & Carroll, J. M. (2023) Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. arXiv preprint arXiv:2309.10771.

Electronic Medical Records of Women in an OB/GYN Context: A Design Justice Perspective

Ayushi Tandon[1], Silvia Masiero[2]
[1]Trinity College Dublin
[2]University of Oslo

Completed Research

Abstract

The notion of *design justice* captures both the way injustice can be directly designed into technology, and the role of human agency in countering it by justice-informed design. This paper applies a design justice approach to an object, Electronic Medical Records (EMRs) in the obstetrician/gynaecological (OB/GYN) context, which closely concerns the production and visualisation of health data. Through qualitative research on OB/GYN EMRs in four Indian hospitals, we discover a reality where a "normal" patient is seen as the default, and deviations from it require substantial workarounds on the doctors' side. Refusing digital health views originating in Western biomedical systems, the use of EMRs by doctors in our study points to three byways for design justice: working towards intersectionality-informed design; designing systems that allow users to inscribe their data preferences into digital health systems; and co-designing digital health solutions. All three byways, illustrated through our field data, offer implications for IS research on design justice.

Keywords: Electronic Medical Records; design justice; women; digital health.

1. Introduction

Data justice has recently garnered scholarly attention within the field of Information Systems (IS) in relation to digital identity systems, social networking platforms, and more topics connected to justice in data production and use (Akbari & Masiero, 2023; Zamani & Diaz Andrade, 2024; Masiero, 2023a; Vannini et al., 2024). With the notion of *data justice* we refer, with Taylor (2017: 1), to "fairness in the way people are visualised, represented and treated as a result of their production of digital data". Researchers of data justice in IS highlight the need for a more equitable approach to data ordering, which includes data collection, storage and processing (Hoefsloot et al., 2022; Masiero, 2023b). In particular data justice scholars advocate for a shift away from techno-solutionism to guide the development of digital systems, in order to imagine solutions that embody the fairness that Taylor's (2017) definition of data justice reflects.

This perspective aligns closely with research on design justice. In her book 'Design Justice: Community-Led Practices to Build the Worlds We Need', Costanza-Chock (2020) notes how the term *design justice* stems from a collective, the Design Justice Network, which participated in the workshop 'Generating Shared Principles for Design Justice' at the Allied Media Conference in 2015. As conceived by the network, design justice aims to capture both the ways injustice can be designed into technology, and the role of human agency in proposing alternatives to it. Research suggests that by integrating a design justice framework into the study of digital systems, we can better understand and address the ethical implications of data practices and work towards a more equitable digital future (Costanza-Chock 2018, 2020; Khene & Masiero, 2022).

Against this backdrop, in this study we focus on Electronic Medical Records (EMRs) to inform a praxis of design justice in the field of digital health. The central object of digital health research (Nielsen & Sahay, 2022; Bardhan et al., 2025), EMRs, has been touted to enable clinicians to quickly identify problems by reviewing patient history on computers, formulate diagnoses and suggest treatment plans (Edwards et al., 2008). The features designed into EMR systems, as materialities of the IT system (Faraj & Azad, 2012), interact with human users, namely doctors, to produce instances and processes of EMR use. Researchers study these interactions to understand the implications of EMR use. Following a similar process, this study examines EMR system use across various obstetrics and gynaecology (OB/GYN) ambulatory care centres by different doctors. We use a design justice approach to analyse and suggest alternative digital health designs.

To do so, we draw on a qualitative study of OB/GYN medical records in four Indian hospitals. The observational data on activities around EMR systems indicate whether a feature is used by doctors during their OB/GYN consultation routines across sites. Further, the design limitations of EMRs are revealed through non-use and interviews with doctors indicating that documentation features in EMR do not align with their expectations, experiences, or perceived ability to deliver the needed care to patients. In this respect, a central contribution of this study is that EMR systems are designed to support documentation and reinforce the implicit definition of a “normal” patient: that is an independent agentic individual, with biomedically quantifiable concerns, and health literate. By contrast, in cases of “*non-normal*” patients, doctors in ambulatory consultations deployed their own styles to create and utilise features in EMR systems. We tease out “*non-normal*” patient by taking design justice approach to analyse data.

The alternatives used by doctors involved- sometimes not using EMR as intended by design, such that they could accommodate the needs of patients not imagined during the design of EMRs. For instance, doctors carved out possibilities for data privacy (as expected by patient) and patient’s care giver involvement. Our findings illustrate instances where doctors challenged dominant medical data recording practices by taking notes that made patient information visible and actionable for themselves and their caregivers. These doctors, by occasionally deviating from Western biomedical documentation practices and challenging their dominance as the only form of expertise, opted for alternatives that empower patients and prioritize their value systems. We abstract and discuss how these doctors' innovative uses of EMRs reveal different byways for design justice in digital health. Based on our findings, we articulate three such byways. These can be defined as: 1) working towards intersectionality-informed design, i.e. unapologetically prioritising the needs of populations living in patriarchal and culturally diverse societies; 2) questioning Western biomedical canons and designing simple systems that allow users to inscribe their data preferences into digital health systems; and 3) situating design possibilities in local context by co-designing digital health solutions that give control to doctors and possibly sensitise designers to scenarios that they may not have imagined. While amenable to further interrogation, these design possibilities suggest pragmatic steps for digital health designers to advance a form of justice that could reverse the implications of unjust EMRs, and more broadly digital health systems.

This paper is structured as follows. We first offer a background section on design justice and its relevance when investigating issues of unfairness and harm connected to technology. The next section illustrates our methodology, centered on qualitative data collection in four hospitals in India. The following section illuminates our findings, starting from quantification of the details of a “normal” patient to its problematisation, as well as the workarounds adopted by the doctors in response to it. In the discussion, we explore three byways of design justice that, problematising a Western view of biomedical systems, present the doctors with new routes to the enactment of design justice.

2. Background

Previous studies within the Design Justice domain have highlighted how design processes can reproduce inequities across various contexts, from computational to urban design (Scheuerman et al., 2021; Piazzoni et al., 2024). Scholars have employed intersectional, speculative, and queer-feminist lenses to analyse systems that perpetuate systemic inequities and inequalities (Costanza-Chock 2020; Floegel & Costello, 2022), advocating for a more just approach to system design. A design justice perspective contrasts with the “dark side” view of IT which upholds those unjust features, potentially leading to harm on subjects, are peripheral and unintended consequences of technologies otherwise designed for social and commercial good (Masiero, 2023a). In opposition to that, a design justice view contributes the idea that injustice can be *directly embedded* in the IT systems as their features and is to be tackled as an inherent component of their design rather than a peripheral one (Costanza-Chock, 2020: 11-15). Such a perspective reveals how IT systems can embody unjust features and illuminates strategies for identifying and mitigating these biases in technologies that impact people's representation in data. Fundamentally, the Design Justice paradigm emphasises the need to examine both the material and non-material aspects of technology (Faulkner & Runde, 2019) to achieve a design approach rooted in liberatory epistemologies (Vannini et al., 2024).

Previous research has shown that the design and subsequent arrangement and relationships among material and non-material components of IT can be studied by examining technology-in-use at its point of application (Baiyere et al., 2023; Gabriel, 2008; Orlikowski, 2000). This suggests that IT design and its use are interconnected, and by observing technology-in-use as a process where users, along with non-human elements, create instances of use and non-use, we can not only analyse but also question the underlying assumptions in its design (Sandberg & Tsoukas, 2011). More so, as previous researchers have shown that IT use processes materialise and become integrated into IT use practices (Sarker et al., 2012; Barley, 1986). Thus, the inherent design rationale underlying the IT artifact significantly impacts the distribution of opportunities and benefits from IT use among diverse user groups.

With this in mind, we note that design justice has informed a particular type of studies, centered on unfair and, in some cases, outright harmful outcomes induced by IT. One class of these outcomes comes from technologies of digital identity, which convert human beings into machine-readable data. Such technologies are associated to the purpose of development objectives, derived by the ability to make people "visible" to the state, or to any organisation in charge of welfare and assistance (Dahan & Gelb, 2015). And at the same time, evidence on the implementation of digital identity programmes on vulnerable populations illustrates the opposite: converted into digital data, entitled users end up being excluded from social protection schemes, profiled by police authorities, and even put at effective risk of statelessness (Chaudhuri, 2021; Cheesman, 2022; Masiero, 2024). By eliciting the design principles implicit into technology, for instance the platform properties that induce exclusion of entitled users (Chaudhuri, 2022), a design justice view enables understanding injustice as directly inscribed in the artefact and its features.

As a result, a design justice perspective enriches understanding of harmful outcomes in IT through at least two routes. First, in opposition to "dark side" approaches, it affords illuminating injustice as embedded into the artefact, rather than reducing it to an unintended feature. Second, this enables researchers to envision routes to combat injustice by working toward the (re)design of artifacts for that specific purpose. Building on this theoretical foundation, this study adopts a design justice framework to examine the use of EMR systems designed to digitise patient medical records.

3. Method

This study is informed by the following research question: *how do OB/GYN doctors utilise EMR systems in their daily practice, and what are the limitations of these systems in supporting quality care?* Rather than considering EMR usage as a binary phenomenon (meaning: used or not used), we aim to investigate the nuanced spectrum of feature-level usage by healthcare providers. Thus, we collected data on EMR use during OB/GYN ambulatory consultations at four different hospitals. This study was conducted in a southern state of India, across urban and rural contexts.

The study covered two large, urban hospitals located in cities with significant knowledge-based industries, one urban hospital in a city primarily driven by the textile industry, and one smaller, rural hospital serving a predominantly agricultural community. The four research sites exhibited varying degrees of EMR usage policies and support for OB/GYN modules. While one large urban hospital mandated EMR use and provided a dedicated OB/GYN module, another site advised EMR use in outpatient settings and their EMR lacked specific OB/GYN features. The remaining two sites had mixed approaches, with both mandating EMR use for certain tasks but one lacking an OB/GYN module and the other having a dedicated OB/GYN module.

Data collection involved a combination of observational and interview methods, with interviews conducted with eight OB/GYN doctors (Rubin & Rubin, 2011). The doctors interviewed and observed had varying levels of experience with medical practice, ranging from less than three to over twenty years. While some sites had a mix of senior and junior doctors, others primarily relied on the expertise of senior OB/GYN consultants. On average, for each doctor, we conducted three interviews, ranging from 5 to 53 minutes in duration. We spent approximately fifteen days with each doctor at the hospital's OB/GYN department. Observations were conducted in various physical locations, including the hospital lobby, consultation rooms, cafeteria, IT department, and patient waiting areas. Based on these observations, doctors were asked specific consulting scenario questions aligned with medical training principles. All audio recordings were transcribed, including translations of interviews conducted in Hindi. Handwritten field notes and details of informal discussions recorded in the diary were digitised. Transcripts were reviewed multiple times. All data was imported into Atlas.Ti 8.0 software for data management and structured memo writing.

We conducted thematic analysis in order to make sense of the central themes of engagement of EMRs by our central respondents, namely doctors (Braun & Clarke, 2014; Williamson, 2017). The data was coded line-by-line for interviews and paragraph-by-paragraph for observations or incidents, as appropriate. Incident-by-incident coding was employed for observational data, as it had already undergone selection and interpretation during the recording process.

It appeared that doctors, during both interviews and direct observation of EMR usage during consultations, frequently referenced activities, protocols, characteristics, and expectations related to patient care. We initially categorised these references based on the specific EMR sections used, patient needs, and doctor expectations. Subsequently, we grouped these activities and expectations into themes that highlighted the tension between biomedical quantification in EMR and resistance to quantification. Similarly, we identified instances where doctors accommodated patients' needs versus those requiring more individualised care. These groupings were associated with underlying basic concepts such as using standardised templates, employing medical terminology, valuing patient culture, and understanding patient privacy concerns. Finally, we synthesised these grouped themes and their associated concepts into the following organising categories: Doctor-Led Adaptation and Resistance towards EMR Systems and Quantifying the Details of the "Normal" Patient in EMR. These two organising categories represent the global theme of variations in EMR usage observed among the doctors. We have thematically organised an presented the findings in the next section.

This research study received ethical clearance from the Institutional Review Board (IRB) at institute of first author. The IRB committee reviewed the research procedure, interview protocol, and sample questionnaires for both doctors. Additional details on obtaining consent from participants were provided, including individual written or verbal consent from patients for observing individual consultation sessions. The study plan and IRB approval number were also submitted to all hospitals' human resource department before fieldwork commenced. All participants' names referred in the finding section of this study are pseudonyms. They were given the choice to suggest pseudonyms for themselves during the fieldwork to avoid researcher bias.

4. Findings

4.1 Quantifying the Details of the “Normal” Patient in EMR

Our research revealed that most doctors described the documentation process during consultation as a standardised, rule-based practice, and emphasised the importance of following established protocols and medical guidelines. This approach, rooted in medical training, informed their use of EMR for record-keeping. They framed the process of recording diagnosis and treatment plan for patients as focusing on identifying specific diseases or conditions based on observable symptoms and diagnostic tests. Doctors especially emphasised the structured and detailed documentation in EMRs. They also mentioned that adhering to standardised practices and language is important to ensure consistency in records in EMR systems. This focus on standardised documentation inadvertently reinforced who is a “normal” patient or how a patient should be so that their usable EMRs are created. This idealised "normal" patient is one who easily fits the structured protocol of inquiring about symptoms, enabling doctors to make diagnoses and suggest treatment plans in line with medical training. The “normal” patient in EMR is one who can provide detailed, clinically relevant information upon request, facilitating structured documentation in biomedical language. Below we elaborate on the “normal” patient as reinforced in EMR.

4.1.1 The “Normal” Patient

The documentation process in EMR, as outlined by the doctors, involves a series of standardised steps, including a detailed medical history shared by patients. Most doctors mentioned first asking patients[1] to describe their condition, the presence or absence of symptoms, disease, or infections. They shared that EMR’s emphasis on sequentially documenting information about menstrual history, obstetric history, surgical history, and family history highlights the ideal framing of patient as someone who can systematically share clinically relevant details. For instance, Dr Manya[2] working at super speciality hospital in tier one city told:

> *“First, we write the* ***basic complaint****, then we write the* ***history*** *of those complaints. Whether they have any issue, just the history of that complaint, how it started and when it started, how it started, and how they got to know about it. Then we ask* ***for menstrual history****. Then we ask for* ***obstetric history****. Then we ask for history, and* ***past history*** *includes all the* ***medical examinations*** *and the results like tuberculosis- all those related diseases. Then we ask for* ***personal history****, which includes the eating habit or [alcohol] drinking habit, any differences in family, any changes in appetite changes in bowel or bladder, and constipation. Then we have* ***family history*** *and any*

[1] Doctors when describing consultation or any other details were using ‘patient’ as a label for people coming for consulting, only when they described something which was not directly related to medical aspects they labelled them as wife, mother, women, IT company employees.

[2] The name of doctor is anonymised as we are using pseudonyms

> ***genetic conditions*** *which are there in the family, and that's it. And then we write the* ***examination.*** *Then we do the* ***physical examination****."*[3]

Similarly, Dr Neeta, working at corporate hospital in tier one city told that they inquire and note in EMR about pre-existing health conditions like thyroid disorders, diabetes, and cardiovascular disease. She also acknowledged the importance of a comprehensive social history, alongside the medical history, for effective consultation.

> *"So, we have a slot for* ***menstrual history [in EMR].*** *I write in there; for married patients, we write there. Then we have* ***contraceptive history,*** *and then any* ***personal medical*** *and* ***surgical history*** *like you know somebody has thyroid, high BP already so we write about it, any surgery-like appendix and something anything can occur inside the body, and it has some relationship. Later, I might find it difficult because she had an appendix removal surgery. So, we want to know that history, surgical history, and any past surgery, and that is what [is] medical and surgical. Then we ask about* ***family history****. Do your mother or father have any disease, particularly blood pressure, thyroid, and diabetes. So, we ask about that and then we ask whether she is* ***allergic*** *to any particular thing or a drug,* ***that is the history we take****."*

We found that the emphasis on diagnostic tests and laboratory results in EMRs mandated doctors to prioritise objective sequential data over subjective unstructured experiences. For instance, when consulting a pregnant woman for the first time, doctors across various hospitals, including those in rural areas, inquired about obstetric history and the last menstrual period. However, before prescribing any treatment plan or follow-up dates, they routinely ordered ultrasound examinations. While this information is considered biomedically crucial for scheduling antenatal screenings and monitoring foetal development, it can pose financial burdens for women from low-income families, potentially discouraging them from returning for subsequent consultations.

[3] Emphasis in all quotes has been added from the authors.

Doctors shared the requirement for detailed documentation in EMRs also lead to a decreased reliance on patient-reported information and more reliance on diagnostic tests. This according to them is particularly concerning for marginalised and vulnerable patients who may face language barriers, cultural differences, or limited health literacy, potentially hindering their ability to accurately articulate their health-related concerns and details.

Overall, the design of EMRs is found to prioritise the collection of standardised data from patients. We find that idealised "normal" patient profile, as implied by EMR design, assumes an independent, agentic individual with a clear and well-articulated health concern. This assumption, as we discuss later, overlooks the barriers faced by marginalised and vulnerable patients, limiting their access to quality healthcare.

4.1.2 Quantification as Priority

Doctors' descriptions of the consultation process involved a detailed medical history, physical examination, followed by diagnostic tests leading to the treatment plan. The documentation of these details in EMRs often involved the use of standardised templates and medical terminology. As Dr Lilee working in corporate hospital shared that:

> *"The format will be there according to the* ***medical thing****, as for* ***a pregnant woman, this is a format for OB****.* ***Gynae cases,*** *then there is another format for gynae and everything else [other formats]. A* ***format is there for everything****."*

For new patients, doctors invested significant time in assigning a medical label to their symptoms and details shared by them and determining an initial treatment plan. Patients were primarily categorised in two groups based on their presenting concerns, with a focus on maternal and foetal health for obstetric patients and reproductive health issues for gynaecological patients. For instance, when a patient presented with concerns related to pregnancy or childbirth, they were categorised as an obstetric case. The focus of care in such cases was on maternal well-being and foetal development, involving regular antenatal check-ups and monitoring. For instance, Dr Sandhya described that for pregnant women meticulous documentation of vital signs in EMR, such as blood pressure and foetal heart rate was considered crucial for monitoring maternal and foetal health.

> *"Generally, pregnancy cases**, vitals are important** because we check the weight how much weight gain has happened, if weight gain is not adequate then we will see, as **generally in pregnancy weight is very critical. And BP is critical**. Generally, in pregnancy, BP will be high. So, BP and weight are more critical"*

Conversely, patients presenting with symptoms or complaints related to the female reproductive system were categorised as gynaecological cases. The diagnostic process for these patients involved a detailed medical history and physical examination, with a focus on identifying specific conditions such as menstrual disorders, sexually transmitted infections, or infertility. This categorisation of patients into distinct groups, based on their presenting concerns, reflects the universal, linear and structured model of care, which controls information about them in EMR. Overall, we find that doctors laid a strong emphasis on abstracting information from patient into label for EMR, particularly within the framework of obstetric and gynaecological biomedical care. In line with this, doctors also categorised women visiting for consultation as either new cases (first-time consultations) or follow-up patients, as explained by Dr Neeta. Thus, categorisation of women determined the specific EMR templates used by doctors, the information gathered, and the retrieval process. Dr Neeta also described how information in EMR was used during follow-up visits of patients.

> *"**This is a follow-up case. I already know the history of the patient,** and we have a proforma for follow up cases, and we **have proforma** for pregnant women. When the first time somebody walks in, we have a **different proforma**, and for the follow-ups, we have different proforma. So, for the **first time** when she comes, that **contains everything anything which is related to the women**, whatever I need anything which we need to ask a lady, all of those will be there."*

> *"But when a patient comes again for follow up, after one month, then, **I will go to previous records, my first thing is on computer.** I will check everything.*

> *Then I will go on from there itself. Then I check BP is entered by the nurse, and if not, it is taken again. Then only we check and examine her. We check previous weight, if there is a weight gain, and if there is any BP change. ...* ***So the next time when somebody visits, for follow up, the notes will be very short notes****."*

This theme indicates that the documentation of consultation process in EMRs often involved the use of standardised templates and medical terminology. Next, we discuss that by prioritising quantifiable data documentation practices, EMRs overshadow the broader context of women's illness and overall health, including their social, and emotional well-being. Thus, doctors had to carve out possibilities of using EMR systems differently and take ownership of such variations in the use of EMR.

4.2 Doctor-Led Adaptation and Resistance towards EMR systems

Our further analysis revealed that while EMR systems are designed to facilitate data-driven clinical decision-making, the doctors in this study demonstrated several instances of resistance against this dominant quantification paradigm. In this section, we describe instances that capture how doctors had to rely on non-standardised methods (i.e. not using EMRs as intended by design) of information gathering and documentation to address the complexities of patient care. Furthermore, doctors often engaged in informal conversations with patients, discussing personal experiences, cultural beliefs, and social factors that influenced their health. These conversations, while not always formally documented, provided valuable insights into the patient's context and helped inform the development of personalised care plans. Additionally, findings suggest that doctors navigate the limitations of EMR systems by adapting their documentation practices. These adaptations highlight potential for discrepancies between the documented record and the full complexity of the patient encounter, but our analysis of documentation processes also provides the opportunity to recommend byways for more just EMR design.

4.2.1 Caring for Non-Normative Cases

During our observation of consultation sessions, we found that doctors often engaged in the use of EMRs in ways that diverged from conventional design expectations. These instances occurred when doctors provided care to patients who did not fit the "normal" patient profile, such as those with limited abilities to share information, describe symptoms, or follow treatment plans due to cultural, financial, or other reasons. Additionally, some patients presented with queries that fell outside the boundaries of clinical language, prompting doctors to document information outside the formal EMR system. For instance, we found that doctors often engaged in detailed discussions with patients about their dietary requirements and specific nutritional needs. When consulting pregnant women, especially in tier-one and tier-two cities, common advice included the benefits of consuming milk for calcium, incorporating green leafy vegetables and fruits into the diet, and avoiding fast food. One such observation is provided below, where Dr Vani, consulting a patient at a multi-specialty hospital in a tier-one city, first advised the patient to consume dry fruits regularly and then, after a pause, recommended custard apple, a fruit high in calcium. Interestingly, while the calcium supplement was mentioned in the EMR, the specific dietary recommendations were not formally documented.

> *The doctor opened a new page on the consultation file and wrote a diagnosis, followed by the prescribed medicine name on the same page [EMR record showed calcium supplement in prescription]. While the doctor was noting, she also asked the woman about her daily schedule and daily household work.* ***The woman explained various problems with the family business. The woman also shared how she's been cooking meals daily for the whole family. Listening to her reaction, the doctor instructed her to eat custard apple (seasonal fruit) everyday, earlier doctor has advised her to eat calcium supplements along with dry fruit.***

In another instance, a patient visited for a postoperative follow-up at a rural multi-speciality hospital, expressing concern about bathing and potential water exposure to her stitches. The attending doctor, Dr Rita after reviewing the patient's medical record and noting the sufficient time elapsed since surgery, advised her to bathe. Specific instructions were provided regarding the application of disinfectant to the stitches and post-bath cleaning. A prescription for disinfectant ointment was also added to the patient's EMR. A second doctor present during the consultation further elaborated on the post-operative care instructions, demonstrating the correct technique for drying the stitches to patient. This doctor also explained the researcher, the local custom of

"Nahan," where women traditionally avoid bathing for a few days post-childbirth. This observation underscores the importance of considering cultural factors and individual patient needs. It also highlights the limitation of EMR systems in capturing such nuanced interactions, as there may not be a specific field or template to document these details. Dr Rita and Dr Sarita both expressed concern that such information, since not explicitly recorded, might be overlooked by future healthcare providers who may not be aware of local customs or have access to relevant contextual information within the EMR. Additionally, some doctors expressed hesitation in assigning definitive diagnoses, particularly for younger patients. For instance, Dr Reena, working at a specialty women's hospital in a tier-two town, avoided labelling young girls with Polycystic Ovary Syndrome (PCOD) while documenting in EMR, instead using terms like "query PCOD" to reflect uncertainty. This cautious approach highlights the potential limitations of EMR design relying solely on standardised diagnostic and uniform documentation practices vis-a-vis the relevance of considering individual patient factors.

> *"**Unless** I am hundred percent **sure of the diagnosis**, I would **mention query PCO** in the record this is for my thing, this is for my own reference. She may not have PCO, so **we don't label them unless, we are sure."***

We also found that treatment and care plans varied based on geographical location and the perceived socioeconomic status of the patient. For instance, doctors in multi-specialty hospitals in rural towns often suggested iron- and protein-rich foods, such as green leafy vegetables. In some cases, patients requested oral supplements as an alternative to dietary modifications, particularly when they could not afford to consume fresh vegetables daily. It is important to note that most of these discussions took place verbally and were not documented in the EMR. Only when patients explicitly requested written documentation were these recommendations recorded, either as free text entries or handwritten notes on EMR printouts. Sometimes, doctors used the plan of care section of the EMR to document specific care requirements, but in some cases, as explained by Dr Sandhya, they would write the care plan and related instructions directly on physical copies of the patient's records.

> *“We don’t* ***usually document the minor things*** *which we tell. There are several things which we tell orally, which we do not document, like exercises, diet habits. Because Indians have a lot of issues in pregnancy. [such as] We should eat that, and we should not eat that.* ***It is nothing like that in the medical literature so we cannot write those things because you do not have medical support for that****.”*

4.2.2 Averting the Quantification

Doctors shared that they often had to inquire about patients' social and personal lives, including their work-life balance, family dynamics, and exposure to stress or violence. They believe that this approach to patient care allows them to better understand the broader context of their patients' health and well-being. However, EMR systems often lack the necessary fields and templates to adequately document such sensitive information. This raises concerns about patient privacy and confidentiality, as some patients may be hesitant to share personal details if they believe they could be accessed by unauthorised individuals. As Dr Ashok, working at a women's specialty hospital in a tier-two city, mentioned during an interview, privacy concerns raised by patients can be a significant challenge in creating comprehensive EMR records.

> *“Sometimes, it is a second marriage, so* ***they don’t want us to mention their first marriage.*** *Husband and wife know, but maybe not everybody knows, so they don’t want us to mention that this was marriage second marriage. But in the record, we have to mention that, like in the first marriage, she has two pregnancies and then something happened and then she got married again, and this time again they are trying, so there are problems. Like this,* ***we have to mention everything for our understanding, and we need to have the records, but they don’t want those details to be on their file****.”*

Doctors themselves acknowledged the delicate balance between the need to document important information for future reference and the risk of compromising patient privacy. To address this tension, some doctors opted to document sensitive information on physical records or in less accessible sections of the EMR. This practice highlights the limitations of current EMR systems in adequately addressing the complex and nuanced aspects of patient care.

During interviews, doctors mentioned that in the case of unmarried or sexually active women, they inquired about vaginal hygiene, menstrual hygiene, and contraceptive use. Notably, there was no dedicated template in the EMR system for documenting these discussions with unmarried women, which could potentially compromise their privacy. Instead, doctors often used the follow-up note template or resorted to handwritten notes. As Dr Lilee pointed out during an interview, since contraceptive prescriptions for unmarried women were often not explicitly documented in the EMR, patients had to rely on their own records and inform doctors about their contraceptive use during future consultations.

> *"Suppose if periods are not regularised, as periods need to be regular after they stop the tablets, so then* ***a patient has to come to follow up or if tablets are exhausted, and they have to come for follow up."***

This highlights the limitations of standardised EMR templates in capturing the nuances of patient-doctor interactions, particularly when it comes to discussing sensitive topics like sexual health and personal hygiene. These adaptations also indicate potential limitations of a one-size-fits-all approach in EMR design and the ongoing negotiation between standardised documentation and the complexities of individual patient experiences.

Overall, this theme suggests how doctors adapted their use of EMR systems to accommodate patients who did not fit the "normal" patient profile. Notably, doctors employed various strategies to bridge the gap between standardised EMR templates and the need for individualised care plans. This included using free text entries, handwritten notes on printouts, and prioritising verbal explanations during consultations. However, their reliance on non-standardised methods highlights the need for alternative approaches to EMR design that can better accommodate the diverse needs of patients and healthcare providers.

5. Discussion: Byways to Design Justice

Overall, our findings highlight a potential limitation of EMR systems, which often assume a uniform patient population and may not be adequately designed to support the contextual nature of individualised care. These findings align with previous research in human computer interaction and sociology of health disciplines. We find that doctors navigate the challenges of caring for non-normative patients, they frequently employ creative workarounds and informal documentation practices to ensure comprehensive care (Berg, 1999; Safadi & Faraj, 2010; Park et al., 2015; Blijleven et al., 2017). While EMR systems can provide valuable information about a patient's medical history and treatment plan, they may not always capture the full range of cultural, social, and personal factors that influence patient care (Berg, 2000). Overall, we can say that EMRs, designed with normative assumptions about patients, can facilitate efficient data entry and retrieval, but they may not always capture the nuances of patient-doctor interactions and the personalized nature of care plans.

In this study, a design justice approach (Costanza-Chock 2020) enabled us to question the design of current EMR systems. This includes questions of health equity, such as: 1) Should we define patients as uniform individual knowing that they are also members of society and community? 2) Which concerns of patients we prioritise, and can we overlook that they have to navigate their daily lives besides managing health? and 3) Who are the beneficiaries of digital records or who do we design for or with? Additionally, we noted the conflict between biomedical values versus the lived realities of patients' lives (what values do we encode and reproduce in the EMR design?). Our findings align with previous research demonstrating how societal norms and limited healthcare access marginalize women, hindering their ability to seek care for reproductive health issues (Garg et al., 2012; 2016; Sivakami & Rai, 2019; Tandon, 2020). The intersection of low literacy and socioeconomic status further exacerbates these challenges, particularly making women part of marginalized patient communities (Desai, 2016, Tandon et al., 2019).

Learning from the technology-in-use described in this paper, we propose three byways that health IT designers can undertake to help mitigate the oppression of marginalised patients, particularly women. The "byway" metaphor encapsulates the tension between dominant paths and less-travelled byways that individuals adopt when their intended destinations do not fall on main trails. These three design byways suggest deviations from dominant biomedical protocols to better address the specific contextual needs of marginalized and underrepresented patients.

1) *Staying with the Marginalised and Designing for Inter-sectional Patient Experiences*. Our findings bring forth potential opportunities for designers and researchers; most importantly expressed through a need to move beyond a medical problem-solving defined only by physiological aspects of patient's bodies. Consequently, we suggest that one byway for just design practice could be to embrace that which is perceived to be marginal need in society around OB/GYN health consultation of women and give it a privileged status in design of EMR systems.

2*) Seeking Alternatives to Datafication originating in Biomedical Dominance*. Our findings suggest the designer to broaden their perspective on the world. Thus, instead of reproducing an already dominant biomedical paradigm where doctor is expert, the designer is expected to value alternative paradigms around women healthcare needs, which are more aligned with the future they want to impact**.**

3) Collectively Imagining and Codesigning with Doctors. Our findings also suggest the importance of working with doctors to understand their documentation practice with and without EMR use. This would allow designers to explore the pluralism of technological use cases and realities that exist in ongoing pasts and alternative nows. Not just designing for them but attuning our sensibilities towards the workflow that they adopt under different circumstances.

But if this is the case, then there is no one-size-fits-all solution to the challenges associated with EMR use, as highlighted in our findings. Instead, we advocate for a design praxis rooted in in-depth observation of existing, just design practices, extraction of generalized principles, and their adaptation to hyper-contextualised digital health interventions. By following this iterative process, we believe it is possible to operationalize a more equitable and inclusive approach to health IT design.

References

Akbari, A., & Masiero, S. (2023). Critical ICT4D: The need for a paradigm change. In *IFIP Joint Working Conference on the Future of Digital Work: The Challenge of Inequality* (pp. 350-355). Cham: Springer Nature Switzerland.

Baiyere, A., Grover, V., Lyytinen, K. J., Woerner, S., & Gupta, A. (2023). Digital "x"—Charting a path for digital-themed research. *Information Systems Research*, 34(2), 463-486.

Bardhan, I., Kohli, R., Oborn, E., Mishra, A., Tan, C. H., Tremblay, M. C., & Sarker, S. (2025). Human-Centric Information Systems Research on the Digital Future of Healthcare. *Information Systems Research,* 0(0), 1-20.

Barley, S. R. (1986). Technology as an occasion for structuring: Evidence from observations of CT scanners and the social order of radiology departments. *Administrative Science Quarterly*, 78-108.

Berg, M. (1999). Patient care information systems and health care work: a sociotechnical approach. *International Journal of Medical Informatics*, 55(2), 87-101.

Berg, M. (2000). Lessons from a dinosaur: Mediating IS research through an analysis of the medical record. In *Organizational and Social Perspectives on Information Technology: IFIP TC8 WG8. 2 International Working Conference on the Social and Organizational Perspective on Research and Practice in Information Technology June 9–11, 2000, Aalborg, Denmark* (pp. 487-504). Boston, MA: Springer US.

Blijleven, V., Koelemeijer, K., Wetzels, M., & Jaspers, M. (2017). Workarounds emerging from electronic health record system usage: consequences for patient safety, effectiveness of care, and efficiency of care. *JMIR human factors*, 4(4), e7978.

Braun, V., & Clarke, V. (2014). What can "thematic analysis" offer health and wellbeing researchers?. *International journal of qualitative studies on health and well-being*, 9(1), 26152.

Chaudhuri, B. (2022). Programmed welfare: An ethnographic account of algorithmic practices in the Public Distribution System in India. *New Media & Society*, 24(4), 887-902.

Chaudhuri, B. (2021). Distant, opaque and seamful: seeing the state through the workings of Aadhaar in India. *Information Technology for Development*, 27(1), 37-49.

Cheesman, M. (2022). Self-sovereignty for refugees? The contested horizons of digital identity. *Geopolitics*, 27(1), 134-159.

Chen, J. X., McDonald, A., Zou, Y., Tseng, E., Roundy, K. A., Tamersoy, A., & Dell, N. (2022). Trauma-informed computing: Towards safer technology experiences for all. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1-20).

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1), 3-21.

Costanza-Chock, S. (2018). Design justice: Towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society*.

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Dahan, M., & Gelb, A. (2015). The role of identification in the post-2015 development agenda. World Bank, Open Knowledge Repository, available at https://openknowledge.worldbank.org/bitstream/handle/10986/22513/The0role0of0id050development0agenda.pdf.

Desai, S. (2016). Pragmatic prevention, permanent solution: Women's experiences with hysterectomy in rural India. *Social Science & Medicine, 151*, 11–18.

Edwards, P. J., Moloney, K. P., Jacko, J. A., & Sainfort, F. (2008). Evaluating usability of a commercial electronic health record: A case study. *International Journal of human-Computer studies*, 66(10), 718-728.

Faraj, S., & Azad, B. (2012). The materiality of technology: An affordance perspective. *Materiality and organizing: Social interaction in a technological world*, *237*(1), 237-258.

Faulkner, P., & Runde, J. (2019). Theorizing the Digital Object. *MIS Quarterly*, 43(4), 1279-1302.

Floegel, D., & Costello, K. L. (2022). Methods for a feminist technoscience of information practice: Design justice and speculative futurities. *Journal of the Association for Information Science and Technology*, 73(4), 625-634.

Gabriel, Y. (2008). Against the tyranny of PowerPoint: Technology-in-use and technology abuse. *Organization Studies*, 29(2), 255-276.

Garg, R., Goyal, S., & Gupta, S. (2012). India moves towards menstrual hygiene: subsidized sanitary napkins for rural adolescent girls—issues and challenges. *Maternal and Child Health Journal,* 16(4), 767–774.

Hoefsloot, F. I., Jimenez, A., Martinez, J., Miranda Sara, L., & Pfeffer, K. (2022). Eliciting design principles using a data justice framework for participatory urban water governance observatories. *Information Technology for Development*, 28(3), 617-638.

Khene, C., & Masiero, S. (2022). From research to action: The practice of decolonizing ICT4D. *Information Technology for Development*, 28(3), 443-450.

Masiero, S. (2024). Unfair ID. London: Sage.

Masiero, S. (2023a). Digital Identity Platforms: A Data Justice Perspective. *Hawaii International Conference on System Sciences 2023 (HICSS-56)*.

Masiero, S. (2023b). Decolonising critical information systems research: A subaltern approach. *Information Systems Journal*, 33(2), 299-323.

Nielsen, P., & Sahay, S. (2022). A critical review of the role of technology and context in digital health research. *Digital health*, 8, 1-10.

Orlikowski, W. J. (2000). Using technology and constituting structures: A practice lens for studying technology in organizations. *Organization Science*, 11(4), 404-428.

Park, S. Y., Chen, Y., & Rudkin, S. (2015). Technological and organizational adaptation of EMR implementation in an emergency department. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 22(1), 1-24.

Piazzoni, F., Poe, J., & Santi, E. (2024). What design for urban design justice?. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 17(3), 379-400.

Rubin, H. J., & Rubin, I. S. (2011). *Qualitative interviewing: The art of hearing data*. London: Sage.

Safadi, H., & Faraj, S. (2010). The role of workarounds during an opensource electronic medical record system implementation. International Conference of Infromation Systems (ICIS) 2010.

Sandberg, J., & Tsoukas, H. (2011). Grasping the logic of practice: Theorizing through practical rationality. *Academy of Management Review*, 36(2), 338-360.

Sarker, S., Sarker, S., Sahaym, A., & Bjørn-Andersen, N. (2012). Exploring value cocreation in relationships between an ERP vendor and its partners: a revelatory case study. *MIS Quarterly*, 36(1), 317-338.

Scheuerman, M. K., Hanna, A., & Denton, E. (2021). Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-37.

Sivakami, M., & Rai, S. (2019). What do we know about sexual and reproductive health of adolescents and youth in India: A synthesis of literature. *In Health and Wellbeing of India's Young People* (pp. 121–156). Springer.

Tandon, A., Kandathil, G., Deodhar, S., & Mathur, N. (2019). Electronic records of obstetrics and gynecology encounter: beyond professional logics of health care. In *Proceedings of the 10th Indian Conference on Human-Computer Interaction* (pp. 1-14).

Tandon, A. (2020). Practical Affordance: EMR Use Within Outpatient Consulting on Women's Health. In *The Future of Digital Work: The Challenge of Inequality: IFIP WG 8.2, 9.1, 9.4 Joint Working Conference, IFIPJWC 2020, Hyderabad, India, December 10–11, 2020, Proceedings* (pp. 180-193). Springer International Publishing.

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, 4(2), 1-14.

Vannini, S., Tandon, A., & Masiero, S. (2024). Feminist and queer approaches to ICT4D: Imagining and enacting liberation. *Information Technology for Development*, 30(2), 195-208.

Williamson, K., & Johanson, G. (Eds.). (2017). *Research methods: Information, systems, and contexts*. Chandos Publishing.

Zamani, E. D., & Díaz Andrade, A. (2024). Understanding local social processes in ICT4D research. *Information Technology for Development*, 30(3), 351-353.

An attention-based view of generative AI technologies in business

Panos Panagiotopoulos
Queen Mary University of London

Muhammad Afzal
Queen Mary University of London

Matthew Forshaw
Newcastle University and the Alan Turing Institute

Completed research

Abstract

Generative AI (GenAI) technologies have captured the attention of business leaders in a powerful way due to their transformational potential. To understand leadership attention to GenAI technologies, we develop a framework based on the attention-based view of the firm. A survey with 419 IT decision-makers integrates aspects of situated attention (external environment, issue characteristics) and the structural distribution of attention (slack resources, organisational data readiness). The findings reveal that industry peer pressure is a key driver of both leadership attention and the anticipated benefits of GenAI. The study underlines the importance of understanding how attention is allocated to strategic priorities in information systems research. It further improves our knowledge of the emerging transition from hype to reality in the early stages of business use of AI technologies.

Keywords: Generative AI, Organisational data readiness, Issue framing, Slack resources, IT adoption, Survey research, Structural equation modelling.

1.0 Introduction

Organisational decision makers are often under considerable pressure to interpret the potential of disruptive digital technologies and revaluate strategic priorities (Ghawe & Chan, 2022; Swanson & Ramiller, 1997). Generative artificial intelligence (GenAI) applications amplify these challenges due to their distinctly transformative nature in automating human input across a large variety of value creation activities (Feuerriegel et al., 2024). As a result, GenAI sets a ‘business no longer as usual’ context where successful organisations can be empowered to create operational efficiencies and redesign their practices (Chowdhury et al., 2024). Even in modest implementation scenarios, GenAI poses a complex set of decisions with implications for resource allocation, upskilling the workforce and assessing changing industry dynamics.

The stakes of GenAI are indeed too high for business leaders to ignore. Adoption surveys like Deloitte (2024)'s State of Generative AI in the Enterprise and Cisco (2023) AI Readiness Index indicate how intensively GenAI has captured the attention of business leaders by revealing a mix of excitement and experimentation. In realistic terms, achieving organisational benefits like efficiency, productivity and reduced costs at a large scale can be hard to prove. In their search for successful applications, organisations have focused on identifying initial GenAI use cases that will demonstrate sufficient value for further investments to scale up (Deloitte, 2024). Accordingly, research has started to unpack the behavioural aspects of GenAI adoption by exploring what determines managers' responses and appraisal of potential investments (Cao et al., 2021; Queiroz et al., 2024; Suseno et al., 2022).

This study aims to examine how business leaders engage with the implementation of GenAI by revisiting the critical role of *attention* in information technology decisions. We develop a theoretical framework and set of hypotheses based on the attention-based view of the firm (Gavetti et al., 2012; Ocasio, 1997; Ocasio et al., 2018). This theory brings forward a behavioural understanding of organisational processes and outcomes as the result of how managers channel their attention and act on specific issues (Brielmaier & Friesl, 2023). The main principle is that organisational leaders are boundedly rational and respond to competing demands based on how their selective attention is generated and distributed by information flows in and around their organisations (Dutton & Ashford, 1993; Gavetti et al., 2012). As Ocasio (1997) notes, attention is situated in the sense that "*what issues and answers decision-makers focus on, and what they do, depends on the particular context or situation they find themselves in*" (Ocasio, 1997, p. 188). Accordingly, organisational leaders have to consider competing priorities, limited resources and other constraints to effectively direct their attention when making decisions.

By seeking to explain organisational attention instead of directly measuring adoption, the study aims to improve our understanding of what drives the early stages of business use of generative AI. This perspective has not been sufficiently explored in information systems research, which more commonly relates organisational adoption decisions to environmental, organisational and technological characteristics. However, attention limits to emerging issues and potential solutions can explain the absence of behaviours such as not engaging with new technologies despite their transformative potential. On the contrary, excessive attentional engagement – or more commonly, the 'hype' – attributed to technologies like GenAI can be a primary driver that shapes strategic priorities, investment decisions and organisational outcomes.

The study presents the findings of a survey with 419 IT decision-makers commissioned via the market research company YouGov. First, in the next section, we develop an attention-based framework that brings together aspects of situated attention (pressures from the external environment, issue characteristics) and the structural distribution of attention (slack resources, organisational data readiness). The findings indicate that industry peer pressure is the most influential factor driving both leadership attention and the expected benefits of GenAI. This was followed by available resources and – to a lesser extent – organisational data readiness and the anticipated benefits of GenAI. Differences in business size, as well as geographic disparities, were found to affect these relationships.

We discuss how an attention-based understanding of business use of GenAI contributes to previous information systems research drawing on behavioural theories (e.g. Arnott & Gao, 2021; Dong et al., 2021; Salge et al., 2015). Finally, we recommend that industry-level organising initiatives can alleviate peer pressure and support more informed investment decisions in implementing GenAI technologies.

2.0 Theoretical development and hypotheses

Herbert Simon challenged the notion of rational decision-making, arguing that bounded rationality stems from limited attentional capacity constrained by resources and time (Simon, 1955). This led to the behavioural theory of the firm which posits that human decision-makers, constrained by attentional limits, cannot consider all possible options together and, therefore, do not always make optimal choices (Cyert & March, 1963). To manage these constraints, organisations establish appropriate structures to allocate and regulate attention, thus shaping decision-making behaviours within firms. Building on this approach, Ocasio (1997) proposed the attention-based view (ABV) positing that attention is the process by which organisational decision-makers notice, encode, interpret, and focus their time and effort on specific issues and answers.

Attention allocation is grounded in three processes. First, an individual's focus of attention is directed to a specific or limited set of issues and alternatives, influencing individual actions (Ocasio, 1997; Ocasio et al., 2018). Second, this focus of attention is shaped by the situational context – constructed by the organisation and influenced by the broader environment – at a given moment (Ocasio, 1997; Ocasio et al., 2018). Third, the latter, in turn, depends on how the organisation creates and regulates situations through its social, economic, and cultural structures (Ocasio, 1997; Ocasio et al., 2018). There are four organisational-level regulators, namely structural positions, rules of the game, resources, and key players – that collectively guide the focus of organisational decision-making.

We argue that the ABV provides a robust framework connecting individual cognitive factors and the organisational context to explain how leadership attention is directed towards emerging technologies like GenAI. When organisations adapt to new opportunities, first they must become aware of external stimuli; this is the initial focus of attention that ABV emphasises. The more attention given to particular stimuli, the more likely resources and managerial support will be allocated to address them. This is particularly relevant for GenAI that requires recognising its strategic potential (Pan et al., 2024). The ABV also emphasises that decision-makers prioritise issues based on alignment with the firm's structure and resources, focusing on those seen as feasible and urgent. In the context of GenAI, this means assessing its relevance for strategic fit, resources, and competitive urgency. Thus, ABV can explain why leadership attention to GenAI may vary based on situational contexts – such as competitive pressures – which influence the relevance and urgency of investing in GenAI. Finally, the ABV underscores how leaders may overlook critical external stimuli if they appear beyond their immediate focus. This bounded attention can make firms vulnerable, as overlooking technologies like GenAI can lead to missed opportunities, increasing the risk of falling behind competitors.

As demonstrated by Cai & Canales (2024), the configuration of these different drivers of attention can significantly illustrate the complexity involved in allocating attention to innovations. As a starting point in the context of GenAI, we distinguish between situated attention and the structural distribution of attention and bring together the research framework and seven hypotheses shown in Figure 1.

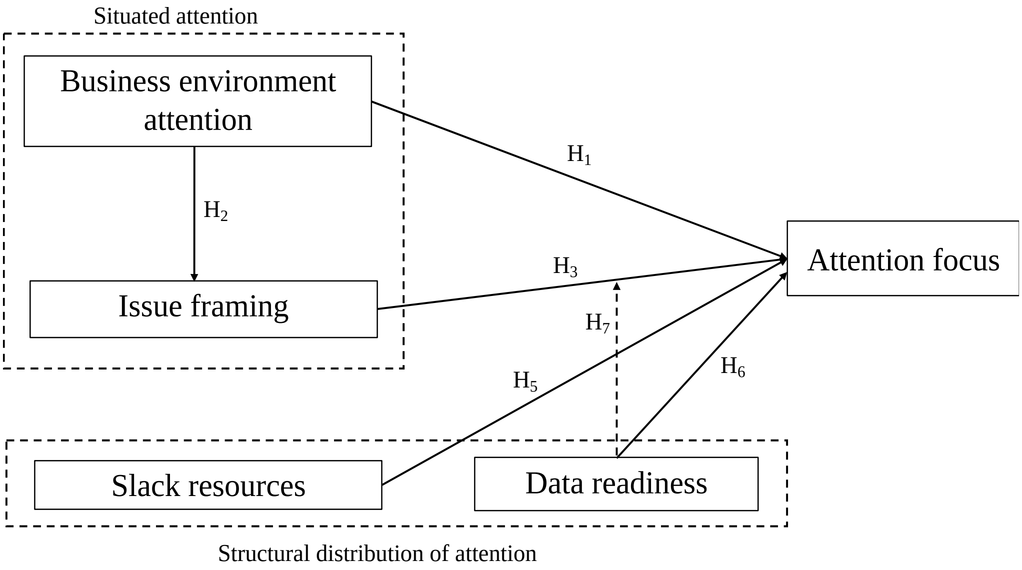


Figure 1. Research framework

2.1 Situated attention

The ABV emphasises that organisations are situated within a dynamic landscape where regulatory, competitive, and technological forces determine the salience of strategic issues, guiding the allocation of leadership attention towards specific challenges and opportunities (Ocasio, 1997). Alternatively, the ABV highlights the environmental cues that are pivotal in shaping the issues and alternatives that organisational leaders perceive as worthy of attention.

Research that draws on the ABV framework consistently demonstrates how environmental pressures can steer leadership focus (e.g. Eggers & Kaplan, 2008; Ghobadian et al., 2022; Papanikolaou & Schmidt, 2022). Specifically, in response to competitive intensity, regulatory shifts, or industry velocity, firms often direct attention toward adaptive strategies that align with these external pressures (McCann & Bahl, 2017; Nadkarni & Barr, 2008). For instance, McCann & Bahl (2017) found that informal competition heightened attention to new product development as an adaptive response, while Nadkarni & Barr (2008) observed that industry velocity shaped the speed and direction of strategic adjustments.

The rapidly advancing technological landscape of GenAI presents both challenges and opportunities, directing leadership attention toward this transformative technology as a source of competitive advantage (Enholm et al., 2022; Queiroz et al., 2024). The external environment’s influence – through factors such as the rapid evolution of AI capabilities and competitive pressures to adopt innovative technologies – can capture and sustain leadership's attention on generative AI (Enholm et al., 2022; Queiroz et al., 2024). Consistent with the ABV perspective, we propose that the business environment, shaped by technological dynamism and competitive intensity, will likely shape leadership focus on generative AI as firms seek to enhance their competitive positioning and adapt to an increasingly AI-driven marketplace. Thus, we hypothesise:

Hypothesis H_1: Attention from the business environment is associated with leadership attention to generative AI.

The ABV further emphasises that environmental embeddedness shapes not only the focus of attention but also how issues are interpreted within the firm. Issue framing is a process by which organisations present specific problems or opportunities in ways that align with organisational priorities, goals, or perceived threats. As Dutton et al. (2001) found, issues framed around valued goals and aligned with recognised organisational logics are more likely to capture leaders' attention. Within the ABV framework, issue framing becomes a mechanism for translating external pressures into internal focus by contextualising these pressures in ways that resonate with leaders and motivate action.

Businesses operating in dynamic and competitive environments are more likely to frame GenAI adoption as a strategic imperative (Pan et al., 2024). McCann & Bahl (2017) found that competitive pressures directed attention toward adaptive responses like new product development. Similarly, businesses under competitive or technological pressures may prioritise GenAI as crucial for sustaining competitive advantage, accelerating digital transformation, or mitigating the risks of falling behind. This framing process thus translates environmental cues into internal narratives, influencing leaders' perspectives. We, therefore, hypothesise:

Hypothesis H2: Attention from the business environment is associated with the issue framing of generative AI within the organisation.

Furthermore, issue framing serves as a crucial factor in determining whether specific issues attract and retain leaders' attention. Research demonstrates that the salience and presentation of an issue can significantly influence its ability to engage top management (Cornelissen & Werner, 2014). For instance Gorgijevski et al. (2019) highlighted that presentation tactics and issue bundling can enhance the appeal of subsidiary initiatives for managers. These findings suggest that when the expected impacts of GenAI are considered strategically valuable, necessary for competitive positioning, or essential for operational efficiency, leaders are more likely to signal their strategic priority (Pan et al., 2024). Specific framing approaches can enhance the appeal of GenAI to leadership by presenting it as a solution aligned with organisational needs, such as cost reduction or innovation. When issues are framed in ways that emphasise urgency, value, and strategic fit, they are more likely to resonate with decision-makers (Cavanagh et al., 2023). Thus, we hypothesise:

Hypothesis H3: The issue framing of generative AI within an organisation is associated with leadership attention.

Grounded in the ABV, the organisation’s business environment shapes leadership attention indirectly by framing GenAI as a strategic issue. Issue framing acts as a mediating mechanism that translates external pressures into salient, actionable organisational priorities. When GenAI adoption is framed in terms of competitive advantage, operational efficiency, or industry demands, it bridges the gap between external pressures and internal strategic focus. This mediation aligns with the ABV’s concept of situated attention, whereby environmental context and organisational framing work together to direct managerial attention. We hypothesise:

Hypothesis H4: The issue framing of generative AI within the organisation, mediates the relationship between attention from the business environment and leadership attention.

2.2 Structural distribution of attention

Slack resources constitute resources that exceed the firm’s immediate operational needs, enabling organisations to explore new opportunities, innovations and investments (Ang & Straub, 1998; Daniel et al., 2004). According to the ABV, leadership attention is a limited and highly valuable resource that tends to be directed toward issues deemed critical or resource-intensive (Ocasio, 1997). Slack resources can influence leadership attention by creating organisational ‘breathing room’, allowing decision-makers to implement strategic initiatives or respond to emerging issues that may otherwise be overshadowed by immediate operational demands (Cyert & March, 1963).

Research demonstrates that organisations with greater optionality of resources are more likely to engage in exploratory activities, such as technological innovation or research and development, as these resources enable them to absorb risks associated with such initiatives (Cai & Canales, 2024; Tsai-Lin et al., 2024). Nadkarni and Barr (2008) and Thosuwanchot and Lee (2024) highlight that resource availability allows businesses to respond proactively in dynamic environments, as they have the necessary reserves to allocate resources toward adaptive initiatives. In the case of GenAI, slack resources make it more feasible to explore this rapidly evolving technology by supporting training, piloting, and other resource-intensive activities (Du et al., 2016; Li et al., 2018; Svahn et al., 2017).

Organisations with slack resources may develop dedicated structures, such as innovation teams or specialised budgets, that specifically encourage attention to exploratory technologies like GenAI. Slack resources thus provide not only the capacity but also the organisational justification to frame GenAI as a viable area for attention, as they reduce the opportunity costs of diverting attention from other activities. Therefore, echoing ABV’s perspective that resource structures shape attention focus, slack resources enhance an organisation’s capacity to engage in strategic exploration, allowing leadership to devote attention to emerging technologies without compromising immediate operational needs. We hypothesise:

Hypothesis H5: The availability of slack resources in organisation is associated with leadership attention to generative AI.

As GenAI has to be integrated into existing organisational processes, data readiness emerges as a critical proxy of GenAI’s realistic implementation potential. Data readiness refers to the infrastructure, skills, processes and policies in place to collect, manage, and analyse organisational data (Abraham et al., 2019; Khatri & Brown, 2010). According to the ABV, attention is often directed toward issues that can be effectively operationalised in light of existing resources and capabilities. Organisations with higher confidence in their data readiness, can be better placed to consider the realistic potential of implementing GenAI technologies. Accordingly, data readiness can facilitate the perception of GenAI as a feasible opportunity and influence the allocation of attention by leaders. In this sense, data readiness reduces the cognitive load associated with integrating GenAI and reduces the risks of rapid experimentation, which aligns well with the exploratory nature of GenAI (Oesterreich et al., 2022; Shamim et al., 2019). Thus, leadership is more likely to view generative AI as a strategic priority in organisations where data readiness is higher, as it aligns with the firm's existing capabilities, infrastructure, and strategic goals. In contrast, when data readiness is lower, leadership may perceive higher obstacles to implementation, reducing the likelihood of allocating attention. Thus, we hypothesise:

Hypothesis H6: Organisational data readiness is associated with its leadership attention to generative AI.

Besides the direct influence on leadership attention to GenAI, data readiness can moderate how the relationship between issue framing and leadership attention. Issue framing helps leadership to make sense of and prioritise emerging issues by presenting them in ways that resonate with organisational goals and values. However, for framing to capture attention successfully, it must be supported by the practical ability to act on the framed issue. Data readiness serves as an important contextual factor in the ABV framework, enhancing the effectiveness of framing efforts by reinforcing the feasibility of GenAI. Framing GenAI in terms of strategic benefits (e.g., productivity gains, cost reduction) is more likely to attract leadership attention when data readiness underpins a clear vision of practical application, reinforcing the framing and elevating GenAI as a strategic priority. This moderating effect enhances the likelihood that leaders will pursue data-centric initiatives. We hypothesise:

Hypothesis H7: Organisational data readiness moderates the relationship between the issue framing of generative AI and leadership attention to generative AI.

3.0 Research methodology

The study employs survey research methodology to test the proposed hypotheses. As shown in Table 1, an instrument was developed to operationalise the theoretical constructs under investigation in line with previous research. Given the scarcity of priori research on attentional structures, no validated instruments were available to measure the constructs of business environment attention, issue framing, and leadership attention focus. Therefore, a new instrument was developed, drawing on theoretical conceptualisations relevant to these constructs and framing items in the context of GenAI (Cai & Canales, 2024).

Constructs	Items	References
Business environment attention	1. Organisations in our industry are actively experimenting with generative AI applications. 2. Being successful in using generative AI can change the dynamics of our industry. 3. Organisations that do not invest in generative AI risk being left behind.	Conceptualised from Brielmaier & Friesl (2023)
Issue framing	1. Generative AI applications can reduce costs. 2. Generative AI applications can improve productivity. 3. Generative AI can help us operate more efficiently.	Conceptualised from Brielmaier & Friesl (2023)
Slack resources	1. Compared with peers, our organisation has more available resources to invest in new initiatives. 2. Our organisation is facing tighter budget conditions than 3 years ago. 3. Our organisation has resources that can be used to fund new initiatives at short notice.	Adapted from Ang & Straub (1998) and De Luca & Atuahene-Gima (2007)
Data readiness	1. My organisation can manage its data responsibly and securely. 2. The data in my organisation are of sufficient quality to support our operational and strategic goals. 3. Employees in my organisation are confident about their data literacy and training needs.	Conceptualised from Abraham et al. (2019) and Khatri & Brown (2010)
Attention focus	1. Leaders have communicated the potential of generative AI as a strategic priority. 2. Leaders have identified the skills and training needs to use generative AI applications. 3. Leaders have established initiatives like working groups, pilot projects and training to support the use of generative AI applications.	Conceptualised from Cai & Canales (2024)

Table 1. Survey instrument

Business environment attention was operationalised through three items: the first assessed the industry's current level of GenAI adoption and openness, the second gauged perceptions of GenAI's potential impact on competitive advantage and industry dynamics, and the third measured the perceived urgency and necessity of GenAI for maintaining competitive parity. *Issue framing* was operationalised through three items designed to capture perspectives on GenAI's organisational value with specific mention to cost reduction (item 1), productivity enhancement (item 2), and operational efficiency (item 3). *Slack resources* was adapted using items from Ang & Straub (1998) and De Luca & Atuahene-Gima (2007).

Attention focus was measured using three items reflecting key aspects of strategic prioritisation: the first assessed leaders' communication of GenAI as a strategic priority, aligning the workforce with long-term goals related to GenAI integration, the second gauged leaders' identification of specific skills and training requirements, highlighting a proactive approach to building GenAI competencies, and the third measured leadership's commitment to action-oriented support, including launching initiatives and pilot projects and dedicating resource allocation to foster generative AI integration within the organisational culture.

All items were measured on a seven (7)-point Likert scale ranging from strongly disagree (1) to strongly agree (7) to impart sufficient discriminatory power to differentiate between respondents. The instrument was pilot-tested with a sample of 100 respondents from the target population to confirm that the items effectively measured the intended theoretical constructs and were clearly understood, ensuring clarity and response reliability.

Control variables were selected based on prior research showing that larger firms with greater financial, technological, and human resources may be more successful in implementing advanced technologies (Lin et al., 2022). Accordingly, firm size, annual turnover, and age were included as controls.

To accurately capture insights on organisational practices within the target population, we followed good practice in survey research focusing on employees with managerial decision-making roles, as they are likely to have a deeper understanding of attention allocation and strategic decision-making within the organisation (Malik et al., 2024). Given the study's context, senior IT professionals with decision-making authority were selected as the primary respondents. Participation required respondents to confirm their decision-making roles with managerial or supervisory responsibilities, ensuring that only qualified individuals proceeded with the survey. The survey was shared with the market research company YouGov, who then distributed it to the target population through their bespoke platform of UK business decision makers (YouGov, 2024).

A total of 491 valid responses were obtained. Table 1 presents the sample demographics. Most respondents are male (79%), and a substantial portion of the sample is highly experienced, with 72% over the age of 35. Approximately 62% of respondents are from SMEs, while 38% represent large businesses. Additionally, around 76% of respondents are from companies established for over 10 years, and 72% work in organisations with an annual turnover of at least £1 million. Notably, the sample is broadly representative, encompassing respondents from a wide array of industries and featuring nearly equal representation of firms located within and outside of London.

An a priori power analysis conducted using G*Power determined that the sample size is sufficient to detect correlations as small as 0.03 with a statistical power of 80% at a significance level of 5%. Furthermore, applying the more conservative inverse square root method, the sample is adequate to detect a minimum path coefficient in the range of 0.11 to 0.12 with 80% power at a 1% significance level (Hair et al., 2022).

Data analysis was based on partial least squares structural equation modelling (PLS-SEM), chosen for its causal-predictive approach and its capability to assess complex structural and measurement models without the constraints of distributional assumptions or sample size limitations (Hair et al., 2022). PLS-SEM analysis was conducted using SmartPLS 4.0, with a maximum of 300 iterations, bootstrapping set to 5,000 samples, and a stop criterion of 10^7.

Characteristics	**Categories**	**n**	**%**
Gender	Female	97	23%
	Male	322	77%
	Prefer not to say	0	0%
Age	Below 35 years	120	29%
	36 - 45 years	113	27%
	Above 46 years	186	45%
Company age	Up to 5 years	40	10%
	5 – 10 years	58	14%
	10 – 20 years	107	26%
	20 – 35 years	86	21%
	Over 35 years	122	29%
Size (full-time equivalent staff)	Small (10-49 employees)	151	36%
	Medium (50-249 employees)	110	26%
	Large (250+ employees)	158	38%
	Don't know / Not sure	4	1%
Company annual turnover	First year of trading	1	<1%
	Less than £1 million	77	18%
	£1 million to £9.9 million	103	25%
	£10 million or more	198	47%
	Don't know	20	5%
	Prefer not to say	20	5%
Type of industries	IT and telecoms	91	22%
	Finance and accounting	60	14%
	Manufacturing	59	14%
	Construction	54	13%
	Retail	22	5%
	Hospitality and leisure	22	5%
	Medical and health services	21	5%
	Media / marketing / advertising / PR and sales	16	4%
	Education	16	4%
	Others	58	14%
Geographical location	Outside London	240	57%
	London	179	43%
Total		419	

Table 2. Demographics of survey respondents and their organisations

4.0 Findings

The reliability and validity of the measurement model were examined following the Hair et al. (2022) criteria. Indicator reliability was assessed using factor loadings, cross-loadings, and Cronbach's alpha. Items with factor loadings above 0.5 on their respective constructs were retained, while others were removed. Factor loadings exceeded 0.7 for each construct, except for one item of *Slack resources*. Cross-loadings for each item on non-substantive constructs were at least 0.2 lower than on its primary construct. Composite reliability, representing internal consistency among items within constructs, was above the minimum threshold of 0.7 for all constructs. Convergent validity, the degree to which indicators of a latent variable measure the same construct, was confirmed using the average variance extracted (AVE) criterion. Discriminant validity, which assesses whether latent variables represent distinct constructs, was examined using HTMT ratios. The AVE for each construct exceeded the minimum threshold of 0.5, ranging between 0.68 to 0.90, and the HTMT ratios for all construct correlations were below the 0.85 threshold. Therefore, both convergent and discriminant validity were established.

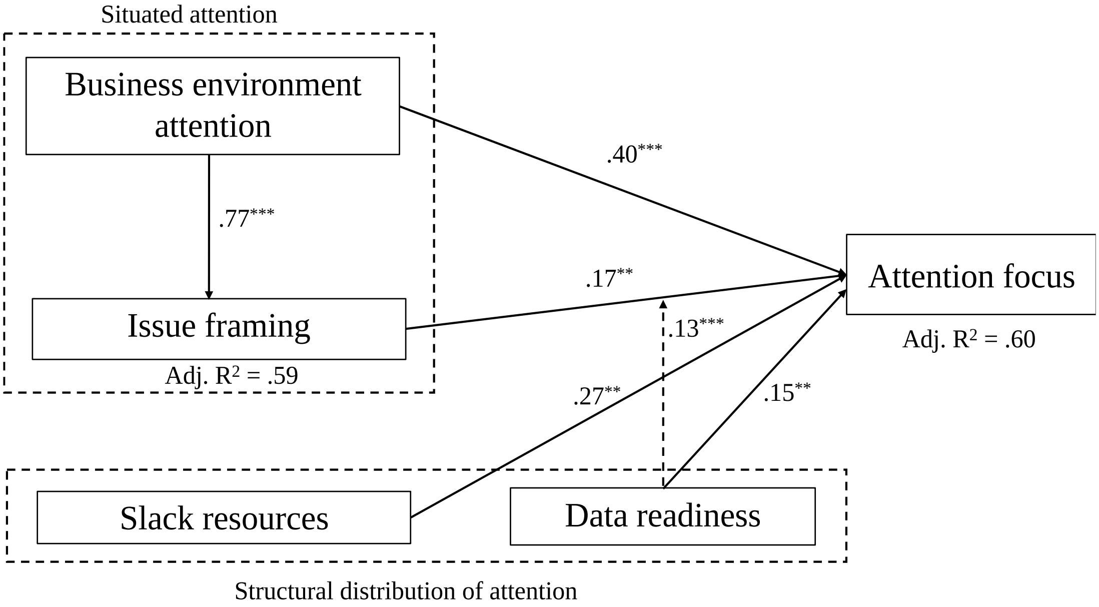


Figure 2. Results of PLS-SEM

The results of the PLS-SEM evaluation are shown in Figure 2 and summarised in Table 3. Overall, the model has strong in-sample predictive power for explaining the variation in *Attention focus*, as indicated by R-square values greater than 0.45. Q^2 values greater than 0.50 further indicate strong out-of-sample predictive accuracy.

The analysis of the situated attention variables, first, reveals that *Business environment attention* is significantly and positively associated with *Attention focus (b=0.40, p<0.001)*. *Business environment attention* is also significantly and positively associated with *Issue framing (b=0.77, p<0.001).* Thus, hypotheses H_1 and H_2 are supported. Furthermore, *Issue framing* is significantly and positively associated with *Attention focus (b=0.17, p=0.01)* and positively and significantly mediates the relationship between *Business environment attention* and *Attention focus (b=0.16, p<0.001)*. Thus, hypotheses H_3 and H_4 are also supported.

On the structural distribution of attention, *Slack resources* are significantly and positively associated with *Attention focus (b=0.27, p<0.001)*, providing support for hypothesis H_5. *Data readiness* is significantly and positively associated with *Attention focus* (*b=0.15, p<0.001*), which supports hypothesis H_6. Finally, *Data readiness* significantly and positively moderates the relationship between *Issue framing* and *Attention focus* (*b=0.13, p<.001)*, substantiating hypothesis H_7.

	Coefficients	***t*-statistics**	***p*-values**
Hypotheses:			
Business environment → Attention focus	.40	5.86	.00
Business environment → Issue framing	.77	27.23	.00
Issue framing → Attention focus	.17	2.55	.01
Slack resources → Attention focus	.27	4.73	.00
Data readiness → Attention focus	.15	2.68	.00
Data readiness x Issue framing → Attention focus	.13	3.94	.00
Indirect effects:			
Business environment → Issue framing → Attention focus	**.**16	4.12	.00
	$\mathbf{R^2}$	**Adj.** $\mathbf{R^2}$	$\mathbf{Q^2_{predict}}$
Issue framing	.60	.59	.59
Attention focus	.61	.60	.58
Total *N = 491*			

Table 3. **Results of hypotheses testing**

To examine variations in perceptions across different demographic groups based on firm size and geographical location, we conducted multi-group analyses using the Welch-Satterthwaite t-test. This analysis was performed to assess differences between small and medium-sized enterprises (SMEs) and large businesses, as well as between firms located inside and outside London. The analysis shows that the relationship between *Data readiness* and *Attention focus* does not hold for SMEs (t-test=0.41, *p<0.01*). Regarding the geographical location analysis, a significant difference emerges in the relationship between organizational *Data readiness* and *Attention focus*, with firms based outside London exhibiting a stronger association compared to those inside London. Notably, *Data readiness* is positively linked to *Attention focus* and further moderates the relationship between *Issue framing* and *Attention focus*, but only for firms located outside London.

5.0 Concluding remarks

Information systems research has begun to draw on theories under the umbrella of the behavioural view of the firm to examine how organisations make strategic IT investment decisions (e.g. Arnott & Gao, 2021; Dong et al., 2021; Salge et al., 2015). This study extends our understanding of organisational responses to the adoption of emerging technologies from an attention-based view. In particular, it responds to suggestions to examine the nature of situated attention in the strategic framing of key organisational issues (Brielmaier & Friesl, 2023). The behavioural challenges that GenAI poses for organisational decision-makers allowed us to conceptualise and measure attention in a highly visible context across a sample of diverse industries within the UK.

The key finding is that peer pressure and contextual influences simultaneously drive leadership attention and the expected benefits of GenAI. In other words, it is ‘hype’ that largely determines the extent to which business leaders focus their attention and signal priorities at the early stages of GenAI implementation. This illustrates the observations by Pan et al. (2024) about concerns of missing out and similar motivations that have led to significant investments in business AI technologies. Other drivers of attention focus, namely: slack resources, organisational data readiness and issue framing – consistent with the ABV – were also found to be significant. These variables – theorised from situated attention and the structural allocation of attention – confirm the theory’s key propositions in an information technology adoption context (Ocasio, 1997; Ocasio et al., 2018). The contribution of the ABV is to provide a new perspective that puts emphasis on attention flows within and outside the organisation. In line with Cai and Canales (2024), uncovering the nature of attention allocation to strategic IT issues like GenAI illustrates the pathways of different motivations that determine organisational responses. For example, we found that larger businesses pay more attention to their current state of data readiness (data management capabilities) while smaller businesses are even more intensively affected by situated attention.

In their practical translation, the findings highlight the importance of industry-level organising activities like Bridge AI initiative by Innovate UK (2024) to alleviate peer pressure and support more informed investment decisions. In particular, SMEs receive so much attention to GenAI from their environment – or competitive pressure – that the contribution of enabling factors like the state of data readiness and available resources become less important. This demonstrates that SMEs need special support in their efforts to implement GenAI more effectively.

Reflecting on the limitations that the study poses, we need to consider how the survey design and data collection influence the findings. Asking IT managers about the use of GenAI in their organisation assumes sufficient shared understanding, however perceptions of these emerging technologies may fundamentally vary between participants, especially since they represent different industries. Therefore, despite approaching GenAI as a unified technology that receives considerable attention, there can be variations in its practical understanding and implementation that the survey is not able to capture. Furthermore, variables from the ABV have been predominantly researched with secondary data sources in strategic management research (Brielmaier & Friesl, 2023). To mitigate this limitation, we extensively pretested and redeveloped the items in the questionnaire.

Finally, the single source of survey data from a reputable source demonstrates the initial ideas but presents constrains of generalisation. Additional studies can, for example, capture areas of organisational leadership other than IT and more contextualised operationalisations of *Issue framing* where the expected benefits of GenAI may represent a higher proportion of situated attention. Future data collection opportunities can also associate later stage GenAI implementation initiatives with attention focus and its drivers. As Pan et al. (2024) similarly suggest, it would be interesting to examine the impacts on certain industries based on their characteristics and how they are influenced by peer pressure dynamics.

References

Abraham, Rene, Johannes Schneider, and Jan vom Brocke. 2019. "Data governance: A conceptual framework, structured review, and research agenda." *International Journal of Information Management* 49:424-438.

Ang, Soon, and Detmar W. Straub. 1998. "Production and Transaction Economies and IS Outsourcing: A Study of the U. S. Banking Industry." *MIS Quarterly* 22 (4):535-552.

Arnott, David, and Shijia Gao. 2021. "Behavioral economics in information systems research: Critical analysis and research strategies." *Journal of Information Technology* 37 (1):80-117.

Brielmaier, Christoph, and Martin Friesl. 2023. "The attention-based view: Review and conceptual extension towards situated attention." *International Journal of Management Reviews* 25 (1):99-129.

Cai, Jing, and J. Ignacio Canales. 2024. "Attention Focus and Attention Framework: A Configuration Perspective of Attention to Innovation." *British Journal of Management* 35 (2):914-931.

Cao, Guangming, Yanqing Duan, John S. Edwards, and Yogesh K. Dwivedi. 2021. "Understanding managers' attitudes and behavioral intentions towards using artificial intelligence for organizational decision-making." *Technovation* 106:102312.

Cavanagh, Andrew, Paul Kalfadellis, and Susan Freeman. 2023. "Developing successful assumed autonomy-based initiatives: An attention-based view." *Global Strategy Journal* 13 (1):176-216.

Chowdhury, Soumyadeb, Pawan Budhwar, and Geoffrey Wood. 2024. "Generative Artificial Intelligence in Business: Towards a Strategic Human Resource Management Framework." *British Journal of Management* 35 (4):1680-1691.

Cisco. 2023. Cisco AI Readiness Index: Intentions outpacing abilities. Cisco.

Cornelissen, Joep P., and Mirjam D. Werner. 2014. "Putting Framing in Perspective: A Review of Framing and Frame Analysis across the Management and Organizational Literature." *The Academy of Management Annals* 8 (1):181-235.

Cyert, Richard, and James G. March. 1963. *A behavioral theory of the firm* Upper Saddle River, NJ: Prentice Hall.

Daniel, Francis, Franz T. Lohrke, Charles J. Fornaciari, and R. Andrew Turner. 2004. "Slack resources and firm performance: a meta-analysis." *Journal of Business Research* 57 (6):565-574.

De Luca, Luigi M., and Kwaku Atuahene-Gima. 2007. "Market Knowledge Dimensions and Cross-Functional Collaboration: Examining the Different Routes to Product Innovation Performance." *Journal of Marketing* 71 (1):95-112.

Deloitte. 2024. The State of Generative AI in the Enterprise: Moving from potential to performance. In *Generative AI Report*: Deloitte.

Dong, John Qi, Prasanna P. Karhade, Arun Rai, and Sean Xin Xu. 2021. "How Firms Make Information Technology Investment Decisions: Toward a Behavioral Agency Theory." *Journal of Management Information Systems* 38 (1):29-58.

Du, Wenyu Derek, Shan L Pan, and Jinsong Huang. 2016. "How a Latecomer Company Used IT to Redeploy Slack Resources." *MIS Quarterly Executive* 15 (3).

Dutton, Jane E., and Susan J. Ashford. 1993. "Selling Issues to Top Management." *The Academy of Management Review* 18 (3):397-428.

Dutton, Jane E., Susan J. Ashford, Regina M. O'Neill, and Katherine A. Lawrence. 2001. "Moves that Matter: Issue Selling and Organizational Change." *Academy of Management Journal* 44 (4):716-736.

Eggers, J. P., and Sarah Kaplan. 2008. "Cognition and Renewal: Comparing CEO and Organizational Effects on Incumbent Adaptation to Technical Change." *Organization Science* 20 (2):461-477.

Enholm, Ida Merete, Emmanouil Papagiannidis, Patrick Mikalef, and John Krogstie. 2022. "Artificial Intelligence and Business Value: a Literature Review." *Information Systems Frontiers* 24 (5):1709-1734.

Feuerriegel, Stefan, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. 2024. "Generative AI." *Business & Information Systems Engineering* 66 (1):111-126.

Gavetti, Giovanni, Henrich R. Greve, Daniel A. Levinthal, and William Ocasio. 2012. "The Behavioral Theory of the Firm: Assessment and Prospects." *Academy of Management Annals* 6 (1):1-40.

Ghawe, Ali S, and Yolande Chan. 2022. "Implementing Disruptive Technologies: What Have We Learned?" *Communications of the Association for Information Systems* 50 (1):36.

Ghobadian, Abby, Tian Han, Xuezhi Zhang, Nicholas O'Regan, Ciro Troise, Stefano Bresciani, and Vadake Narayanan. 2022. "COVID-19 Pandemic: The Interplay Between Firm Disruption and Managerial Attention Focus." *British Journal of Management* 33 (1):390-409.

Gorgijevski, Alexander, Christine Holmström Lind, and Katarina Lagerström. 2019. "Does proactivity matter? the importance of initiative selling tactics for headquarters acceptance of subsidiary initiatives." *Journal of International Management* 25 (4):100673.

Hair, Joseph F., G. Tomas M. Hult, Christian M. Ringle, and Marko Sarstedt. 2022. *A primer on partial least squares structural equation modeling (PLS-SEM)*. Third edition. ed. Los Angeles: SAGE.

Innovate UK. 2024. "Bridge AI: Bridging the gap in artificial intelligence" https://iuk-business-connect.org.uk/programme/bridgeai/ [accessed 02/11/2024].

Khatri, Vijay, and Carol V. Brown. 2010. "Designing data governance." *Commun. ACM* 53 (1):148–152. doi: 10.1145/1629175.1629210.

Li, Liang, Fang Su, Wei Zhang, and Ji-Ye Mao. 2018. "Digital transformation by SME entrepreneurs: A capability perspective." *Information Systems Journal* 28 (6):1129-1157.

Lin, Shunzhi, Jiabao Lin, Feiyun Han, and Xin Luo. 2022. "How big data analytics enables the alliance relationship stability of contract farming in the age of digital transformation." *Information & Management* 59 (6):103680.

Malik, Mohsin, Amir Andargoli, Roberto Chavez Clavijo, and Patrick Mikalef. 2024. "A relational view of how social capital contributes to effective digital transformation outcomes." *The Journal of Strategic Information Systems* 33 (2):101837.

McCann, Brian T., and Mona Bahl. 2017. "The influence of competition from informal firms on new product development." *Strategic Management Journal* 38 (7):1518-1535.

Nadkarni, Sucheta, and Pamela S. Barr. 2008. "Environmental context, managerial cognition, and strategic action: an integrated view." *Strategic Management Journal* 29 (13):1395-1427.

Ocasio, William. 1997. "Towards an Attention-Based View of the Firm." *Strategic Management Journal* 18:187-206.

Ocasio, William, Tomi Laamanen, and Eero Vaara. 2018. "Communication and attention dynamics: An attention-based view of strategic change." *Strategic Management Journal* 39 (1):155-167.

Oesterreich, Thuy Duong, Eduard Anton, and Frank Teuteberg. 2022. "What translates big data into business value? A meta-analysis of the impacts of business analytics on firm performance." *Information & Management* 59 (6):103685.

Pan, Shan-Ling, Rohit Nishant, Tuure Tuunanen, and Jyoti Choudrie. 2024. "Is AI a strategic IS? Reflections and opportunities for research." *The Journal of Strategic Information Systems* 33 (4):101866.

Papanikolaou, Dimitris, and Lawrence D. W. Schmidt. 2022. "Working Remotely and the Supply-Side Impact of COVID-19." *The Review of Asset Pricing Studies* 12 (1):53-111.

Queiroz, Magno, Abhijith Anand, and Aaron Baird. 2024. "Manager Appraisal of Artificial Intelligence Investments." *Journal of Management Information Systems* 41 (3):682-707.

Salge, Torsten Oliver, Rajiv Kohli, and Michael Barrett. 2015. "Investing in information systems: on the behavioral and institutional search mechanisms underpinning hospitals' is investment decisions." *MIS Quarterly*. 39 (1): 61–90.

Shamim, Saqib, Jing Zeng, Syed Muhammad Shariq, and Zaheer Khan. 2019. "Role of big data management in enhancing big data decision-making capability and quality among Chinese firms: A dynamic capabilities view." *Information & Management* 56 (6):103135.

Simon, Herbert A. 1955. "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics* 69 (1):99-118.

Suseno, Yuliani, Chiachi Chang, Marek Hudik, and Eddy S. Fang. 2022. "Beliefs, anxiety and change readiness for artificial intelligence adoption among human resource managers: the moderating role of high-performance work systems." *The International Journal of Human Resource Management* 33 (6):1209-1236.

Svahn, Fredrik, Lars Mathiassen, and Rikard Lindgren. 2017. "Embracing digital innovation in incumbent firms." *MIS quarterly* 41 (1):239-254.

Swanson, E. Burton, and Neil C. Ramiller. 1997. "The Organizing Vision in Information Systems Innovation." *Organization Science* 8 (5):458-474.

Thosuwanchot, Nongnapat, and Min Suk Lee. 2024. "Independent directors' ownership and CSR performance: the moderating roles of factors impacting directors' attention." *Journal of Strategy and Management* 17 (1):167-187.

Tsai-Lin, Tung-Fei, Ming-Huei Chen, Hui-Ru Chi, and Pei-Shan Chiang. 2024. "The impact of R&D organizational structure on developing technological capabilities and the moderation of R&D slack." *Journal of Organizational Change Management* 38(1): 158-181.

YouGov. 2024. "B2B Omnibus." https://business.yougov.com/product/realtime/b2b-omnibus [accessed 02/11/2024].

Integrating UDL and AI: A Reflexive Account of the Digital Maieutic Project Team

Eleni Tzouramani
University of the West of Scotland, UK
Charis Manousou
University of the West of Scotland, UK
Theofilos Tzanidis
University of the West of Scotland, UK

Research In progress

Abstract

This paper offers a reflexive account of the Digital Maieutic Project which integrates Universal Design for Learning (UDL) principles with artificial intelligence (AI) to create adaptive and inclusive educational environments. 'Socrates', an AI chatbot designed to facilitate dialogic learning through the Socratic method and to apply the UDL principles of multiple means of representation, engagement and action/expression. Adopting Bell and Willmott's (2020) model of constitutive, epistemic and disruptive reflexivity the project team reflected on ethical considerations, power dynamics and interdisciplinary challenges. The team addressed tensions inherent in integrating AI within UDL frameworks such as technological constraints in delivering multimodal content, ethical dilemmas regarding student autonomy and the need for cultural relevance in AI interactions. Early feedback shows that 'Socrates' improves student engagement and administrative efficiency, but questions remain about defining and measuring success in AI-driven education. This paper contributes to the broader discourse on sociomateriality of AI in higher education.

Keywords: AI in education, Inclusive Education, Multimodal Learning, Reflexivity.

1.0 Introduction

This paper is a reflexive account of our ongoing engagement with the Digital Maieutic Project. The Digital Maieutic Project applies Universal Design for Learning (UDL) principles to integrate artificial intelligence (AI) in higher education and develop adaptive and inclusive learning environments. The main part of this project is the development and use of 'Socrates', an AI chatbot that combines Socratic questioning with UDL's core principles of multiple means of representation, engagement and action/expression (CAST, 2024). Designed to accommodate diverse cognitive, sensory and motivational needs through flexible and reflexive pedagogical approaches, 'Socrates encourages dialogic, student centred learning. Socrates supports both module administration and module learning by engaging students in dialogue prompting deeper inquiry and concept understanding. The project further integrates multimodal AI interactions and integrates AI across curriculum design, assessment and teaching practice. Our aim is not to just implement AI but to reflect on pedagogical and ethical implications and continuously adapt our approach through reflexivity in order to develop inclusive, adaptive learning environments where technological innovation is balanced with pedagogic integrity..

By integrating UDL principles, the project aims to create an educational experience that engaging, inclusive and accessible, aligning with neuroscience perspectives that address accessibility as a fundamental right rather than a privilege, guiding educators to adopt inclusive and reflective teaching practices (Proctor, Dalton & Grisham, 2007; Rao, Ok & Bryant, 2014). In adopting UDL principles, we reframe the role of educators as facilitators of adaptive learning environments, moving away from a traditional teacher-centred model of education, where educators are the primary sources of knowledge, towards a more student-centred, inclusive approach that customises educational experiences to diverse learner needs. The Digital Maieutic project is implemented across four university modules, Digital Communications Project (Information Systems), Digital Marketing Practice (Marketing), Organisational Behaviour (Management) and Inclusive Leadership Practices (Education), utilising dialogic interactions to support students with both administrative information and content comprehension. The project impact will be evaluated trough student engagement metrics, achievement and qualitative feedback from focus groups.

The integration of AI in educational contexts is both promising and contentious. On the one hand, there is the potential of AI offering individualised learning experiences through adaptive systems and chatbots which can transform the learning experience (Chen et al., 2020). At the same time, the implementation and use of AI in academia raises ethical concerns about issues of fairness, bias and the potential for dehumanization (Al-Amoudi, 2022; Etzioni & Etzioni, 2017; Nguyen et al. 2022). In the Digital Maieutic project, we seek to navigate these tensions by positioning AI within a UDL framework that values both student autonomy and critical engagement. We aim to adopt a balanced approach which would address the need for ethical AI, fostering an environment in which learners are not passive recipients but active participants within a continuously adaptive, technology-mediated learning environment.

To ground the project in ethical integrity the project team engages in systematic reflexive practice to continuously examine and refine 'Socrates' interactions and application. Reflexivity is embedded in the project's methodology as a critical lens for examining our assumptions, practices and outcomes. Following Bell and Willmott's (2020) framework of constitutive, epistemic and disruptive reflexivity, the team evaluated the sociomaterial, ethical and epistemic dimensions of AI integration in higher education. Reflexivity is not seen as a prescriptive framework but as an iterative process, enabling continuous adaptation to emergent ethical and practical challenges. By reflecting on team dynamics, design choices and on the interplay of human and non-human actors, the project seeks to establish a model for ethical and inclusive AI integration in educational contexts.

The objectives of this paper are: 1) to explore how reflexivity informs our (the project team's) efforts to design and implement 'Socrates' according to UDL principles. 2) To analyse the way in which the three UDL principles of engagement, representation and action/expression are applied through AI. 3) To evaluate the ethical and pedagogical implications of using AI to foster inclusive learning environments. In this way, the paper contributes to the broader discourse of AI in education, demonstrating how reflexivity can support the alignment of AI with UDL principles. Moreover, this paper situates the Digital Maieutic project within the evolving field of educational sociomateriality (Orlikowski & Scott 2008), acknowledging the entanglement of human and non-human actors in shaping educational practice.

2.0 Methodological Framework

2.1 Reflexivity approach

Reflexivity as a methodological approach includes multiple dimensions. It involves critically examining the assumptions and values underpinning research practices while acknowledging how these are shaped by and contribute to broader social, material and epistemic contexts (Bell and Willmott, 2020). Reflexivity also requires an awareness of the influence of language, ideologies and power dynamics on the construction and interpretation of knowledge, highlighting how research is embedded in and reproduces social interests and structures (Alvesson, Hardy and Harley, 2008). Reflexivity also engages with the ethical and political dimensions of research, requiring researchers to critically reflect on their positionality and the implications of their actions within embodies and contextualised research processes (Bell and Willmott, 2020). Together, these aspects emphasise the iterative and critical nature of reflexivity as a tool for navigating the complexities of knowledge production.

In the context of the Digital Maieutic Project, reflexivity is integral in guiding design and decision making processes as well as continuously aligning with UDL principles and ethical considerations. Reflexivity and especially team reflexivity can act as a catalyst for innovation by encouraging a culture of critical engagement and knowledge sharing as well as iterative learning processes and collaborative creativity (Hoegl & Parboteeah, 2006; Schippers, West & Dawson, 2015). As this project is new to the team, we decided to allocate the time for reflexivity to allow us to navigate complex interdependencies and continuously adapt to evolving contexts (Dryden-Palmer, Parshuram & Berta, 2020).

Although aware of Hoegl and Parboteeah's (2006) conclusion that reflexivity enhances team effectiveness but there is an inconsistency in improving efficiency due to the significant time and resources required to engage in the reflective process, we value Steen's (2021) 'slow innovation', seeing reflexivity as a counterbalance to the pressures of rapid innovation, allowing our team to deliberate effectively and address systemic complexities. We are also aware of the advantages of reflexivity varying across project phases as well as the risk for shifts in focus and scope creep (Hoegl and Parboteeah, 2006) but we hope will be addressing this as the project progresses. Regarding the emotional dynamics of the team, we align with Bieler et al.'s (2020) collaborative reflexivity as a distributed process developing the emotional and relational dimensions of our team. Following Hartmann et al.'s (2020) suggestions, we find that joy enhances our reflexive capabilities and we make an effort to include levity in our interactions, strengthening our mutual relationships and a positive emotional climate in the team.

Drawing on Bell and Willmott's (2020) conception of reflexivity, this paper adopts a multidimensional approach, integrating constitutive, epistemic and disruptive reflexivity. Constitutive reflexivity examines the foundational assumptions of the project design and implementation and the ways in which these assumptions influence and are influenced by UDL principles and technological/practical choices. Within this we challenged our own assumptions and potential biases and limitations in our approach e.g. What does ‘providing multiple means of representation’ mean? Is the dialogic method as inclusive as we think it is? Are we favouring specific forms of engagement over others? Epistemic reflexivity examines the knowledge claims and practices shaping the project, assessing the validity and inclusivity of practice implementation. Here we reflect on our assumption that the Maieutic will enhance students’ understanding of concepts however, in our effort to make the chatbot interactive and interesting while maintaining accessibility, are we oversimplifying complex concepts? Is the over reliance on text limiting the range of representation? Are different types of prompts, e.g. image or audio prompts, needed to address wider range of needs? We also spend time discussing issues of privacy and transparency but mostly whether this type of personalisation of the learning experience might make students too reliant on AI to provide answers and limit their interest in exploring other sources. Disruptive reflexivity challenges bias in embedded practices and explores alternative possibilities. It mostly comes up in critical moments of disruption where through reflective dialogue we challenged our practice and started exploring alternatives. Specifically, we challenged the biases in our design of ‘Socrates’. Would the particular type of dialogue and the material it consists of limit thinking diversity? Would it lead to perceived ‘correct’ answers that might not be universally applicable or openminded? Most importantly, although our overall efforts aim to engage the whole person, this is a digital modality based on text, image or video, unable to engage kinaesthetic ways of learning.

We also include in our approach, insights from Schon’s (1983) concept of reflective practice which emphasises the iterative relationship between reflection and action, advocating for ‘reflection-in-action’ as a means of adapting to emergent challenges. This principle aligns with the project’s dynamic design process where we continuously engage in reflexive dialogue to discuss the project and address emergent issues. Reflexivity, allows us to critically asses show design decisions influence and are influenced by broader educational, social and technical considerations. Our reflective practice also focuses on ethical accountability, guiding us in navigating tensions between technological constraints and the imperatives of UDL principles.

With regards to data collection for the reflective piece of this work, we draw on observation, meeting note and reflective debrief data that we have started to systematically collect. We record our observations on how we engaged with reflexive practice to address challenges such as the generation of multimodal content and ethical risks in data collection. We meet once or twice weekly and in between meetings we exchange notes on Microsoft Teams. Every Friday, we meet for an hour to reflect on the project's process and support each other during challenges. As the project is still in an initial stage, our collected reflective data are basic. Utilising and thematic analysis approach guided by the principles of reflexivity. We analyse patterns and tensions on the project's technical design, alignment with UDL principles and ethical standards.

As our dataset grows, we plan to keep reviewing and validating themes and perhaps incorporate extra reflective methods such as a reflective journal.

2.2 Ethical considerations

Ethical accountability is central to the Digital Maieutic Project, integrating UDL principles with AI technology with equity, transparency and care. Engaging in constitutive, epistemic and disruptive reflexivity (Bell and Willmott, 2020), we continuously check and evaluate our decisions and their implications for privacy, inclusivity and autonomy. Through this reflexive process, inherent in this project, we critically examine how each part of the project aligns with ethical and pedagogical goals. Regular reflexive discussions among team members allowed for critical examination of the impact of each part of the project on student privacy, autonomy and inclusivity. We worked on our ethics submission via Google Docs collaboratively while also exchanging messages on the project's Microsoft Teams chat.

The overall Digital Maieutic Project adopts a proactive approach to ethical research and implementation. All student participants are fully informed bout the use of AI in their modules, participation is voluntary with informed consent obtained before engagement. Alternatives are provided for students who prefer not to use AI tools. Data are anonymised at the point of transcription with identifying information removed before analysis. All collected data are stored securely in password protected files, accessible only to the project team and will be deleted one year after the project's completion. To ensure confidentiality, pseudonyms will be used in any published outcomes that include participant quotes. Transparency is ensured by informing students about the exploratory nature of the project and by providing multiple feedback channels to voice their reflection and concerns. Focus groups include debriefing sessions to address any emotional or intellectual challenges that might arise. Module Evaluation Questionnaire (MEQ) results and Student-Staff Liaison Group (SSLG) feedback is anonymised before any use.

Our reflexive practice tends to reveal tensions as well as potential challenges associated with engaging students in AI mediated learning. We are still in the process of addressing these challenges and our experience so far shows that reflexivity is invaluable for ensuring our project is based in ethical accountability. Most importantly, the project is guided by an ethics of care and critical accountability (Bell and Willmott, 2020) that extends beyond procedural ethics. Discussions of sensitive topics during focus groups and in class are conducted with care, addressing potential emotional effects through debriefing sessions and supportive follow-ups. Continuous and systematic reflexivity of the project team are embedded in the process as intentional strategies to promote fairness, inclusivity and autonomy in learning.

3.0 Context of the project

3.1 Multiple means of representation

The UDL principle of multiple means of representation advocates for diverse methods of content delivery to support the different sensory and cognitive needs of learners (CAST, 2024). This principle challenges traditional one size fits all educational models by suggesting that knowledge is more accessible when presented through multiple means, fostering a richer, more inclusive learning environment (Rao, Ok & Bryant, 2014). Within the Digital Maieutic project, the AI chatbot Socrates is designed to present information dynamically across textual, visual and interactive formats, transforming passive learning into a more engaged dialogic process while the maieutic process provides support for scaffolding (Vygotsky, 1978) of learning by giving adaptive responses, offering tailored made content and enabling a more personalised learning experience. At the scaffolding process, learners are given structures that support them in progressing beyond their immediate capabilities (Vygotsky, 1978). Doing this through multiple means of representation such as text, visual images or more interactive elements, can help support different learning pathways according to individual learners sensory preferences and processing styles (Meyer, Rose & Gordon, 2014). Based on the neuroscience of learning, UDL is supported by research showing that multimodal engagement allows learners to approach material from multiple cognitive angles therefore activating various neural networks and enhancing memory retention (CAST, 2024).

Combining UDL’s principle of representation and the Socratic method, our chatbot enables dialogic interaction, encouraging students to critically question, reinterpret and actively construct meaning from the material. Contrary to traditional static formats, this approach promotes ‘critical reflexivity’ (Cunliffe, 2004) by encouraging learners to continuously reflect on their understanding, reevaluating and deepening their engagement with content. Through its different representational formats, ‘Socrates’ supports a reflexive learning process that doesn’t only provide information but also prompts students to question and analyse in a process of active critical inquiry instead of just passive content consumption. Learners are called to reinterpret and question the response and in this way the engagement with the module material becomes deeper. This approach aligns with a critical pedagogical approach where knowledge is co-constructed through dialogue rather than just transmitted from teacher to student (Freire, 2017). In this way, it challenges traditional, hierarchical, educator led models towards more student-centred, dialogic learning environments where the educator takes on a supportive, facilitator role. The ‘Socrates’ chatbot is a start in working towards providing all students, regardless of background or prior knowledge, the tools to engage meaningfully with the module content.

Working in a widening participation university, the project team is consciously and continuously trying to reimagine the learning experience and support all students as active participants in their learning journey. The principle of representation allows for improving inclusivity through offering content in textual, visual and interactive formats, surpassing traditional models that favour text based learning which could potentially disadvantage students with different learning styles and sensory preferences. Based on Nissenbaum’s (2004) theory of ‘contextual integrity’, where equitable access to information requires sensitivity to contextual needs, we tried to develop Socrates with care for privacy and balancing this with ensuring that information is available to all and accessible in formats that respect and value individual learner differences.

While this is an introductory project, we have enlisted the help of students in adopting a reflexive approach where we continuously evaluate how representational choices affect student engagement and learning outcomes. We adopt a stance of assuming ethical accountability in knowledge production (Bell & Willmott, 2020) where we critically examine the influence of AI representations on learner autonomy and agency. We, the project team, have asked for continuous feedback from students to help us audit Socrates's interactions, the different representation formats and checking against UDL principles and ethical considerations.

3.2 Multiples means of Engagement

Engagement as a principle of UDL emphasises the importance of motivation, interest and autonomy in shaping meaningful learning experiences. In the Digital Maieutic project, engagement is enhanced through interactive, student driven dialogue with 'Socrates'. Allowing students to explore the module content autonomously, ask questions and approach topics at their own pace and level of understanding, Socrates transforms the learning process from passive reception to active inquiry. Through questioning, 'Socrates' prompts students to critically evaluate their responses, fostering a continuous reflexive process that engages both cognitive and affective dimensions. Socrates allowing students to take ownership of the learning, creates a sense of autonomy. Deci and Ryan's (2000) Self Determination Theory identifies autonomy, competence and relatedness as key drivers for motivation. Through the interactive learning space provided by Socrates, students become co-creators of knowledge, enhancing their engagement and sense of agency.

From a reflexivity perspective, engagement is not only about attention or involvement but also about fostering a critical, self-reflective approach to learning and seeing learning as a 'critically reflexive practice' (Cunliffe 2004). Students are called to question, analyse and interpret information actively. By encouraging students to question assumptions and explore diverse perspectives 'Socrates' supports an engagement that goes beyond the surface into a deeper, critically engaged learning experience. This engagement alights with critical pedagogical principles where learning is an active, co-constructive process and students are empowered to think independently and critically (Freire 1970).

From a sociomaterial perspective (Orlikowski &Scott, 2008) these action and expression pathways are not simply student choices but constitutively entangled expressions shaped by Socrates' affordances. Each mode of response, whether analytical, reflective or exploratory, gains meaning through its relational entanglement with the chatbot's prompts. Student and 'Socrates' responses mutually influence and co-constitute each other in a dynamic interrelation that shapes the learning process.

Engagement becomes more attainable within flexible, transformative experiences rather than in traditional teaching formats. 'Socrates' adopts a conversational approach that engages learners in a dialogic process and allows for autonomous, private and 'safe' explorations of the module content. Drawing on Nissenbaum's (2004) concept of 'contextual integrity', where inclusivity in engagement requires sensitivity to each student's unique context, needs and motivations, this process engages personal motivations and interests on the material by allowing learners to ask their own questions and explore content in their own time and space. In line with UDL values, this process values the diverse ways in which students are inspired to learn.

3.3 Multiple Means of Action and Expression

UDL's focus on varied forms of action and expression is integrated in the digital maieutic project by promoting student autonomy in the ways in which they interact with the content and they utilise AI to complete their assessment or parts of their assessment as instructed. This UDL principle challenges traditional restrictive assessment frameworks valuing instead flexibility and inclusivity learners' engagement with the module content. Learner autonomy as well as agency and choice encourage learners to engage in ways that align with their strengths and preferences, reinforcing their sense of competence and motivation (Deci & Ryan, 2000). For this part, 'Socrates' currently faces limitations as its emphasis on text does not support other forms of action and expression such as receiving prompts in image, audio or video form and responding accordingly. A more advanced system would provide multimodal capabilities, allowing students to engage and respond through diverse formats. Such tensions within sociomaterial systems can be seen as opportunities for innovation and exploring links with other systems may help address this.

4.0 Preliminary reflexive thematic analysis.

Recognising that thematic analysis is not a purely mechanical or formulaic process but instead, it requires deep engagement with data and a sensitivity to research complexities (Braun & Clarke, 2021; Braun & Clarke, 2023), we present this section as a preliminary analysis at this early stage of the project.

Drawing on Bell and Willmott's (2020) approach to reflexivity, our team recognised the importance of reflexivity and looking at how our interactions, assumptions and power dynamics shape the project. Working across disciplines and on an innovative project, requires us to continuously navigate and negotiate our differences in epistemological perspectives, professional priorities and technical expertise, as Mauthner and Doucet (2003) suggest, knowledge construction is never neutral but it is deeply embedded in personal, institutional and social contexts. To maintain responsiveness and transparency we have adopted both structured and unstructured communication routines including weekly meetings and at least three informal discussions via Teams chat each week. This combination of formal and more ad hoc communication helps us engage in reflexivity in an iterative but dynamic way (Braun and Clarke, 2023).

One challenge that emerged early during our discussions about the development of Socrates was the tension between the need for multimodal means and technical constraints. Although we prioritised aligning the project with UDL principles, the limitations of technology (Bray et al., 2024), project timeline and scope require continuous dialogue and reconciliation. These tensions continue to emerge in our discussions and inspired by Gabriel's (2015) ideas of embracing tensions as productive forces that catalyse innovation, we welcome them and consciously integrate them into our decision making. Specifically, we see value in the tensions about technological limitations and the development of inclusive means of engagement as opportunities to explore innovative solutions that balance technological constraints with pedagogical expectations. In the same spirit, we make an effort to allow for imagination in our reflective practice (Gabriel 2015), moving beyond self-criticism towards envisioning alternative possibilities.

Representation is proving to be one of the most challenging UDL systems to integrate within the constraints of an AI system. Reflexive analysis highlighted the limitations of 'Socrates's' reliance on text which excludes students with visual or auditory learning preferences. Here, Gabriel's (2008) critique of technological determinism helped us examine the limitations of technology and our ability to provide multimodal learning. We explored creative solutions such as embedding multimedia content, diagrams, videos and images within the chatbot's responses. Despite some initial promising outcomes, the development of multimodal responses remains challenging. Most of our creative solutions require significant technical and conceptual effort as well as a significant investment in time to develop videos and graphs and infographics. Trainor and Bundon's (2021) critique reminded us that even these solutions are not neutral but they are shaped by the cultural and technological biases embedded in the tools and resources we use. Our dialogical reflections further revealed the issue of cultural and contextual relevance in responses and across disciplines where we are now in the process of adapting examples and case studies to reflect the diversity of our participants.

Trial feedback from students suggests that Socrates is having a positive impact on student engagement. They highlight the ease with which they can interact not only with the module content but also with the administrative parts such as accessing submission deadlines. These early responses affirm our confidence in the project and 'Socrates' as both a pedagogical and administration support system. At this stage, it is still too early to make inferences about any impact on conceptual understanding. While engagement seems to be improving, through reflexivity we approach these results cautiously, acknowledging the complexity of measuring engagement in relation to learning outcomes during the early stages of the project. One important question that has come up is how we define and assess the success of AI learning tools in terms of the complexity of defining success, assessing it across disciplines and across time as well as its alignment with our ethical commitments.

5.0 Conclusion

By critically evaluating assumptions, practices and outcomes through constitutive, epistemic and disruptive reflexivity, the project team works to address tensions between UDL principles and technological constraints within a process of iterative learning and ethical accountability. Although preliminary student responses are positive, challenges persist especially in developing multimodal content, ensuring cultural relevance and balancing technological limitations with diverse learner needs. The project contributes to the broader discourse on AI in education by offering a model for aligning AI systems with inclusive pedagogical frameworks and adopting reflexivity as its basis for iterative learning and ethical accountability.

6.0 References

Al-Amoudi, I. (2022) Are post-human technologies dehumanizing? Human enhancement and artificial intelligence in contemporary societies. Journal of Critical Realism, 21(5): 516–538.

Alvesson, M., Hardy, C. & Harley, B. (2008) Reflecting on reflexivity: reflexive textual practices in organization and management theory. Journal of management studies, 45(3): 480-501.

Bell, E. & Willmott, H. (2020) Ethics, politics and embodied imagination in crafting scientific knowledge. Human Relations, 73: 1366-1387.

Bieler, P., Bister, M. D., Hauer, J., Klausner, M., Niewöhner, J., Schmid, C. & Von Peter, S. (2021) Distributing reflexivity through co-laborative ethnography. Journal of Contemporary Ethnography, 50(1): 77-98.

Braun, V. & Clarke, V. (2023) Is thematic analysis used well in health psychology? A critical review of published research, with recommendations for quality practice and reporting. Health Psychology Review, 17(4): 695-718.

Braun, V. & Clarke, V. (2021) One size fits all? What counts as quality practice in (reflexive) thematic analysis?. Qualitative research in psychology, 18(3): 328-352.

Bray, A., Devitt, A., Banks, J., Sanchez-Fuentes, S., Sandoval, M., Riviou, K., Byrne, D., Flood, M., Reale, J. and Terrenzio, S. (2024) What next for Universal Design for Learning? A systematic literature review of technology in UDL implementations at second level. British Journal of Educational Technology, 55: 113-138.

CAST (2024) UDL Guidelines. Available at: https://udlguidelines.cast.org/

Chen, L., Chen, P. & Lin, Z. (2020) Artificial Intelligence in Education: A Review. Ieee Access, 8: 75264-75278.

Craig, S.L., Smith, S.J. & Frey, B.B. (2019) Professional development with universal design for learning: supporting teachers as learners to increase the implementation of UDL. Professional Development in Education, 48(1): 22-37.

Cunliffe, A. L. (2004) On becoming a critically reflexive practitioner. Journal of Management Education, 28: 407-426.

Deci, E. L. & Ryan, R. M. (2000) The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. Psychological inquiry, 11(4): 227-268.

Dryden-Palmer, K.D., Parshuram, C.S. & Berta, W.B. (2020) Context, complexity and process in the implementation of evidence-based innovation: a realist informed review. BMC Health Services Research, 20: 1-15.

Etzioni, A. & Etzioni, O. (2017) Incorporating Ethics into Artificial Intelligence. The Journal of Ethics, 21(4): 403-4018.

Freire, P. (2017) Pedagogy of the oppressed. (13th Ed) London: Penguin Modern Classics.

Gabriel, Y. (2015) Reflexivity and beyond – a plea for imagination in qualitative research methodology. Qualitative Research in Organizations and Management: An International Journal, 10(4): 332-336.

Gabriel, Y. (2008) Against the tyranny of PowerPoint – Technology-in-use and technology abuse. Organization Studies, 29(2): 255-276.

Hartmann, S., Weiss, M., Hoegl, M. & Carmeli, A. (2021). How does an emotional culture of joy cultivate team resilience? A sociocognitive perspective. Journal of Organizational Behavior, 42(3): 313-331.

Hoegl, M. & Parboteeah, K.P. (2006) Team reflexivity in innovative projects. R&d Management, 36(2): 113-125.

Mauthner, N. S. & Doucet, A. (2003) Reflexive accounts and accounts of reflexivity in qualitative data analysis. Sociology, 37(3): 413-431.

Meyer, A., Rose, D.H. & Gordon, D. (2014) Universal design for learning: Theory and Practice. Wakefield, MA: CAST Professional Publishing.

Nguyen, A., Ngo, H.N., Hong, Y., Dang, B., Nguyen, B.P.T. (2022) Ethical principles for artificial intelligence in education. Education and Information Technologies 28: 4221–4241.

Nissenbaum, H. (2004) Privacy as contextual integrity. Washington Law Review, 79(1): 119-157.

Orlikowski, W. & Scott, S. (2008) Sociomateriality: challenging the separation of technology, work and organization. The Academy of Management Annals, 2(1): 433-474.

Proctor, P.C., Dalton, B. & Grisham, D.L. (2007) Scaffolding English Language Learners and Struggling Readers in Universal Literacy Environment With Embedded Strategy Instruction and Vocabulary Support. Journal of Literacy Research, 39(1): 71-93.

Rao, K., Ok, M.W. & Bryant, B.R. (2014) A Review of Research on Universal Design Educational Models. Remedial and Special Education, 35(3): 153-166.

Rogers-Shaw, C., Carr-Chellman, D.J. & Choi, J. (2018) Universal Design for Learning: guidelines for accessible online instruction. Adult Learning, 29(1): 20-31.

Rose, D.H. & Meyer, A. (2006) A practical reader in Universal Design for Learning. Harvard Education Press.

Schippers, M.C., West, M.A. & Dawson, J.F. (2015) Team reflexivity and innovation: The moderating role of team context. Journal of Management, 41(3): 769-788.

Schön, D. A. (1983). The reflective practitioner: How professionals think in action. London: Temple Smith.

Steen, M. (2021) Slow Innovation: The need for reflexivity in Responsible Innovation (RI). Journal of Responsible Innovation, 8(2): 254–260.

Trainor, L. R. & Bundon, A. (2021) Developing the craft: Reflexive accounts of doing reflexive thematic analysis. Qualitative research in sport, exercise and health, 13(5): 705-726.

Vygotsky, L.S. (1978) Mind in Society: The development of higher psychological processes. Harvard University Press.

Advancing Pedagogical Innovation and the Research-Teaching Nexus with the Cognitive AI Framework in a Global Context

Colin Fu
School of Management
UCL

Athina Ioannou
Surrey Business School
University of Surrey

Effie Kelana Chia
Surrey Business School
University of Surrey

Research In progress

Abstract (around 150 words)

This study critically examines the impact of integrating ChatGPT, a generative AI tool, into global education systems, highlighting both its transformative potential and the intricate challenges it presents. Utilizing the Cognitive AI Framework developed by Fu (2023), the research explores how ChatGPT can enhance personalized learning and operational efficiency while navigating ethical, cultural, and regional obstacles. Findings reveal that while ChatGPT supports differentiated learning and encourages self-directed study, its reliance on user data and susceptibility to algorithmic bias raise significant ethical concerns. Regional disparities, such as stringent data privacy laws in Europe and limited digital infrastructure in parts of Asia and Africa, further complicate its adoption. The study emphasizes the necessity of culturally adaptive AI solutions and ethical literacy programs to mitigate these challenges. By applying the Cognitive AI Framework, educators and policymakers can align AI integration with diverse educational values, promoting equitable access and fostering responsible use. This research underscores the critical need for a balanced approach to AI adoption, ensuring technological innovation does not exacerbate existing educational inequalities.

Keywords: AI in Education, ChatGPT, Cognitive AI Framework, Ethical AI, Global Education, PRME

1.0 Introduction

The rapid integration of Artificial Intelligence (AI) into education is transforming traditional learning and teaching models globally (World Economic Forum, 2024).

Initially focusing on automating administrative tasks such as grading multiple-choice assessments and analysing administrative data, AI’s educational applications have now expanded into more complex areas, offering tools for personalized learning, adaptive assessments, and data-driven insights into student performance. Among the most significant developments in AI for education is the emergence of generative AI tools, such as ChatGPT by OpenAI, among others, which provides real-time, human-like responses. This technology has become a pivotal tool in education, enabling students and educators to interact with information in novel ways, from virtual tutoring to aiding classroom discussions and content creation (Kohnke, Moorhouse & Zou, 2023). However, while ChatGPT presents exciting possibilities for individualized instruction and support, its role in education is not without controversy, as it raises concerns about data privacy, ethics, and over-reliance on AI for learning (Baidoo-Anu & Owusu Ansah, 2023).

This study explores ChatGPT’s impact on global education by focusing on both the benefits it brings and the challenges it presents, particularly concerning equity and cultural adaptability. On the one hand, ChatGPT has demonstrated significant potential to support differentiated learning, adjusting to students’ individual needs and encouraging self-directed study. Yet, AI’s reliance on user data and potential for bias poses ethical challenges that are further complicated by diverse regional norms and regulatory landscapes. For example, regions with stringent data privacy laws, such as Europe, may approach AI use cautiously, while areas with less digital infrastructure may see ChatGPT as a tool for bridging educational gaps (Zhou, Jurafsky & Hashimoto, 2023). This research seeks to understand how ChatGPT’s adoption varies across educational contexts and the extent to which it can support inclusive, equitable education globally.

To address these issues, the study applies the Cognitive AI Framework developed by Fu (2023), a model that provides structured guidance for integrating AI ethically and thoughtfully in diverse educational settings. This framework is designed to promote responsible AI usage through stages that encourage exploration, engagement, critical examination, and ethical reflection. By framing ChatGPT within this model, the study investigates how educators and policymakers can make AI a constructive force in education, balancing technological innovation with ethical integrity.

The objectives of this research are twofold: to evaluate the advantages of ChatGPT for enhancing personalized learning and operational efficiency, and to analyze the ethical, cultural, and regional challenges its use entails. The findings aim to guide educators and policymakers in adopting AI thoughtfully, ensuring it aligns with regional educational values and fosters equitable access to learning opportunities worldwide.

2.0 Literature Review

Prior literature on Artificial Intelligence (AI) in education underscores its transformative potential for both teaching and learning. AI-powered tools, including adaptive learning platforms and virtual tutors, offer powerful benefits such as personalized instruction, immediate feedback, and enhanced student engagement (Kohnke, Moorhouse & Zou, 2023). ChatGPT, in particular, is part of a new wave of generative AI that can respond dynamically to students' queries and engage them in conversation, offering explanations and guidance around the clock. This makes it a valuable supplement to traditional education, where teacher availability may be limited, and personalized support is often resource-intensive (Adeshola & Adepoju, 2023).

Despite these benefits, the integration of AI in education also presents significant ethical challenges, particularly concerning data privacy, algorithmic bias, and the digital divide. Data privacy is a primary concern, as AI systems like ChatGPT rely on substantial user data to function effectively. This data collection raises concerns about confidentiality, consent, and potential misuse, especially in regions without robust regulatory protections (Korkmaz, Cemal Aktürk & Talan, 2023). Additionally, AI systems are vulnerable to algorithmic bias, as they often reflect the biases present in their training data. This can lead to culturally insensitive or inaccurate responses, disproportionately affecting students from underrepresented backgrounds and reinforcing existing inequalities (Zhou, Jurafsky & Hashimoto, 2023). The digital divide further complicates AI's role in education, as students in low-resource regions may lack access to the digital infrastructure needed to benefit from AI-enhanced learning (Abbas, Jam & Khan, 2024). These issues highlight the need for culturally adaptive AI solutions that address the needs and values of diverse educational settings.

To guide ethical AI adoption in education, Fu (2023) introduced the Cognitive AI Framework, which emphasizes a balanced, structured approach to AI use. The framework consists of five stages: *explore*, *engage*, *examine*, *formulate*, and *reflect*.

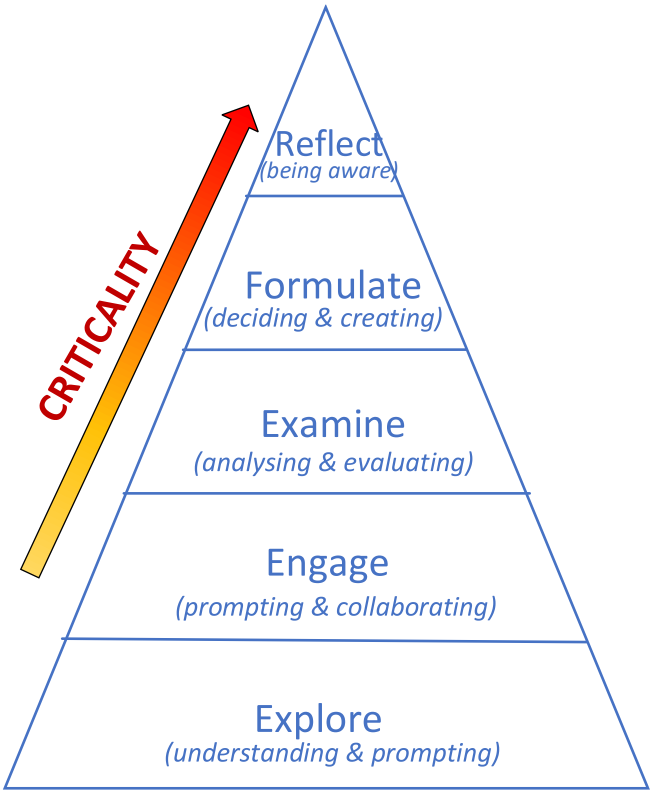


Figure 1.

Cognitive AI Framework (Fu, 2023)

The *explore* stage fosters understanding of questions posed to generative AI, identifies knowledge gaps, and encourages critical inquiry, laying the groundwork for informed and thoughtful AI integration (Krause and Stolzenburg, 2024). The *engage* stage involves collaborating with generative AI by continuously refining and updating questions based on its outputs to achieve optimal outcomes. The *examine* stage encourages users to evaluate AI outputs critically, considering factors like relevance, accuracy, and cultural sensitivity. This stage is essential for identifying potential biases within AI responses, helping educators to mitigate the impact of biased data (Rospigliosi, 2023). The *formulate* stage emphasizes designing and developing innovative uses of generative AI to create ethical, robust, and value-aligned outputs. Students and educators are encouraged to use AI creatively, incorporating it into innovative learning strategies that enhance, rather than replace, traditional methods, fostering creativity and advancing knowledge within an ethical framework. Finally, the *reflect* stage promotes ongoing ethical evaluation, prompting users to consider the broader social, cultural, and ethical implications of AI in education. This reflection is particularly relevant in regions with limited digital literacy, where the risks associated with rapid AI adoption may not yet be fully understood (Fu, 2023).

By structuring AI integration around these stages, the Cognitive AI Framework supports an approach that respects cultural diversity and fosters responsible AI use. Its emphasis on critical examination and reflection aligns with theoretical models such as the Technology Acceptance Model (TAM) and the Diffusion of Innovations (DOI), while expanding them to address ethical concerns uniquely pertinent to AI in education (Davis, 1989; Rogers, 2003). The framework's adaptability makes it a valuable tool for educators and policymakers worldwide as they work to integrate AI responsibly into diverse educational landscapes.

3.0 Methodology and Analysis

This study employed a mixed-methods approach to explore ChatGPT's role in education, utilizing a global survey to gather direct insights from educators, students, and AI professionals. The survey was designed to collect quantitative data on participants' familiarity with and perceptions of ChatGPT, as well as qualitative data on its benefits, challenges, and ethical concerns. This methodology was framed by the Cognitive AI Framework, which guided data collection and analysis to ensure a structured, ethical, and culturally sensitive approach.

3.1 Survey Design

The survey included both closed-ended and open-ended questions, covering four key areas:

- Familiarity and experience with ChatGPT.
- Perceived benefits, such as personalized learning and operational efficiency.
- Challenges, including ethical concerns related to data privacy and algorithmic bias.
- Cultural relevance, focusing on how ChatGPT aligns with regional educational norms.

The survey was hosted on Qualtrics, a secure platform compliant with data privacy standards, and distributed globally through professional networks, educational institutions, and social media channels. Targeted outreach was employed to ensure balanced representation across regions, particularly in underrepresented areas. Over the one-month data collection period, responses were monitored, and additional outreach was conducted as needed to achieve diversity in participants.

Survey responses were analysed using a combination of quantitative and qualitative techniques:

- Quantitative data, derived from closed-ended questions, was processed using descriptive and inferential statistical methods to identify trends and correlations.
- Qualitative data from open-ended responses underwent thematic analysis to uncover recurring themes and insights, particularly regarding ethical concerns and cultural relevance.

3.3 Application of the Cognitive AI Framework

The Cognitive AI Framework, developed by Fu (2023), structures the research design, guiding data collection and analysis around five stages: Explore, Engage, Examine, Formulate, and Reflect. Each stage emphasizes a specific aspect of ethical and cultural evaluation in AI integration, ensuring that the research approach is aligned with the study's objectives.

The Cognitive AI Framework structured the study across five simplified stages, ensuring ethical and culturally relevant integration of AI into education:

1. *Explore:* This stage involved assessing participants' general understanding and familiarity with how they use ChatGPT for knowledge acquiring. Survey questions in this stage focused on capturing baseline knowledge and setting the foundation for deeper engagement.
2. *Engage:* Participants shared their experiences of using ChatGPT in educational settings, detailing specific applications such as tutoring, administrative tasks, and content creation. This stage highlighted practical use cases and challenges.
3. *Examine:* Critical evaluation of ChatGPT's outputs formed the core of this stage. Participants reflected on issues like reliability, bias, and cultural appropriateness, enabling an understanding of how these factors influence the tool's effectiveness.
4. *Formulate:* In this creative stage, responses explored how ChatGPT could be used innovatively to enhance learning. Ideas included integrating ChatGPT into collaborative projects and adapting it for region-specific needs.
5. *Reflect:* This final stage emphasized ongoing ethical considerations. Participants addressed broader implications, including data privacy, digital equity, and cultural sensitivity, helping to shape recommendations for responsible AI adoption in education.

By aligning each stage of the Cognitive AI Framework with distinct elements of data collection and analysis, this study provided a holistic and ethically grounded exploration of ChatGPT's role in education. This structured approach facilitated a comprehensive understanding of AI’s impact across diverse educational settings, effectively balancing quantitative data with culturally sensitive, qualitative insights.

4.0 Preliminary Findings

The preliminary findings from the global survey reveal a nuanced and diverse landscape of perceptions and attitudes toward ChatGPT in education, shaped by regional infrastructure, regulatory frameworks, and cultural contexts. Across all regions, ChatGPT was acknowledged for its potential to enhance learning efficiency and provide personalized educational support. Participants highlighted its applications in virtual tutoring, administrative assistance, and content creation, underscoring the tool’s versatility and ability to complement traditional teaching methods. However, perceptions of its benefits and challenges varied significantly across regions.

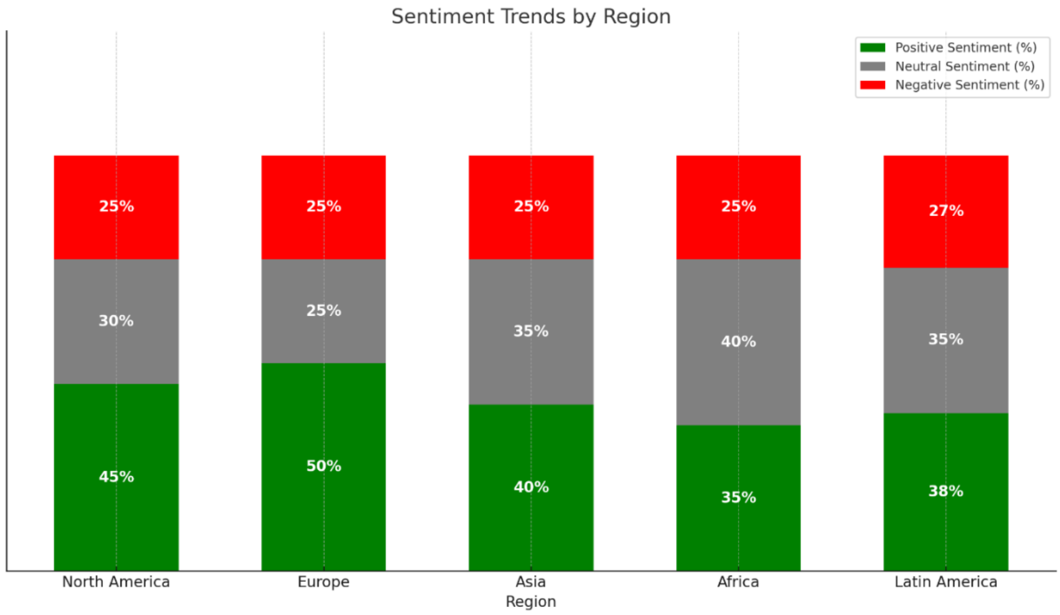


Figure 2.

Sentiment Analysis by Regions

In Europe, respondents expressed optimism about ChatGPT’s integration into education. Many educators viewed the tool as a valuable resource for personalized instruction and operational efficiency, citing its ability to reduce workload and facilitate student engagement. This positive perception was supported by regulatory frameworks, such as the General Data Protection Regulation (GDPR) in the UK and equivalence in

Europe, and an emphasis on AI literacy, which fostered confidence in ethical and responsible AI use (Baidoo-Anu & Owusu Ansah, 2023).

Contrastingly, responses from South Asia and parts of Africa highlighted significant challenges. Participants in these regions reported limited access to digital infrastructure and concerns about the digital divide, which restricts the effective adoption of AI technologies. The lack of adequate connectivity and resources emerged as a significant barrier to leveraging ChatGPT for equitable education. Moreover, cultural relevance was a recurring theme, as many educators noted that ChatGPT's reliance on Western-centric training data often led to outputs that were misaligned with local cultural and educational norms (Zhou, Jurafsky & Hashimoto, 2023). This underscores the need for region-specific adaptations of AI tools to ensure inclusivity and cultural appropriateness. Respondents from North America shared Europe's optimism but were more cautious about privacy and academic integrity concerns. The presence of strong data protection laws and ethical safeguards has contributed to higher confidence in AI adoption but has also raised expectations for transparency and accountability in AI tools like ChatGPT (Kamalov, Calonge & Gurrib, 2023).

Across all regions, participants identified the importance of integrating ChatGPT into education in a way that balances technological innovation with ethical and cultural considerations. Issues such as algorithmic bias, data privacy, and over-reliance on AI were frequently raised, highlighting the need for ongoing evaluation and adaptation.

6.0 Discussion and Implications

The application of the Cognitive AI Framework structured the analysis and interpretation of these findings, revealing key insights and actionable implications.

In the *explore* stage, the study uncovered varying levels of familiarity with ChatGPT across regions, emphasizing the importance of foundational AI literacy programs to ensure equitable access to its benefits. The *engage* stage provided insights into ChatGPT's current applications, with positive feedback from regions with robust digital infrastructure, contrasting with accessibility challenges in underserved areas. The *examine* stage highlighted the critical need to address issues of reliability, cultural relevance, and algorithmic bias. Participants consistently emphasized that ChatGPT's outputs must be evaluated for accuracy and contextual appropriateness, particularly in non-Western regions where cultural misalignment was more pronounced. These findings underscore the importance of adapting AI tools to local needs, fostering trust, and promoting their effective use. In the *formulate* stage, participants suggested innovative uses of ChatGPT, such as creating culturally tailored educational materials and enabling collaborative learning experiences. These responses demonstrated ChatGPT's potential to support creativity and collaboration in education when aligned with regional priorities.The *reflect* stage revealed broader concerns about the ethical implications of AI in education, including privacy, equity, and data governance. Participants from regions with limited regulatory protections highlighted the urgent need for policies that address these challenges, while respondents from regions with stronger frameworks stressed the importance of maintaining high ethical standards as AI adoption grows.

Overall, the findings affirm ChatGPT's potential to transform education but emphasize the need for region-specific adaptations and ethical oversight. Policymakers and educators must address the challenges of digital equity, cultural relevance, and ethical literacy to ensure that AI tools like ChatGPT support inclusive, equitable, and meaningful learning experiences worldwide.

7.0 Conclusion and Further Research

This study underscores the transformative potential of ChatGPT in global education, while also revealing the intricate challenges that must be addressed to realize its full promise. Central to this research is the Cognitive AI Framework, illuminating how AI adoption can be aligned with diverse educational, ethical, and cultural imperatives, offering a replicable model for future explorations of AI in education.

Future research should expand on this study by conducting longitudinal assessments of ChatGPT's impact on educational outcomes, particularly in fostering critical thinking, creativity, and teacher-student interactions. Additionally, in-depth regional studies are needed to explore how AI literacy and cultural adaptations influence the effectiveness of AI tools. These studies should aim to refine and localize the Cognitive AI Framework, ensuring its applicability across diverse educational settings.

References

Abbas, M., Jam, F.A. and Khan, T.I. (2024) 'Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students', *International Journal of Educational Technology in Higher Education,* 21(1). doi: https://doi.org/10.1186/s41239-024-00444-7.

Adeshola, I. and Adepoju, A.P. (2023) 'The opportunities and challenges of ChatGPT in education', *Interactive Learning Environments,* pp. 1–14. doi: https://doi.org/10.1080/10494820.2023.2253858.

Baidoo-Anu, D. and Owusu Ansah, L. (2023) 'Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning', *Journal of AI*, 7(1), pp. 52 - 62. doi: https://doi.org/10.61969/jai.1337500

Davis, F.D. (1989) 'Perceived usefulness, perceived ease of use, and user acceptance of information technology', *MIS Quarterly,* 13(3), pp. 319-340. doi: https://doi.org/10.2307/249008.

Fu, C. (2023) Revolutionize The Education Paradigm in the Generative AI Era. In: QS Higher Ed Summit Aisa Pacific 2023. 2023. Kuala Lumpur, Malaysia, 7 - 9 November, 2023. QS, London

Kamalov, F., Calonge, D.S. and Gurrib, I. (2023) 'New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution', *Sustainability* 2023, *15*(16), 12451. doi: https://doi.org/10.3390/su151612451

Kohnke, L., Moorhouse, B.L. and Zou, D. (2023) 'ChatGPT for language teaching and learning', *RELC Journal*, 54(2), pp. 537–550. doi: https://doi.org/10.1177/00336882231162868.

Korkmaz, A., Cemal Aktürk, M. and Talan, T. (2023) 'Analyzing the User's Sentiments of ChatGPT Using Twitter Data', *Iraqi Journal for Computer Science and Mathematics*, pp. 202–214. doi: https://doi.org/10.52866/ijcsm.2023.02.02.018.

Krause, S. and Stolzenburg, F. (2024) 'Commonsense reasoning and explainable artificial intelligence using large language models', *Communications in Computer and Information Science*, pp. 302–319. doi: https://doi.org/10.1007/978-3-031-50396-2_17.

Rogers, E.M. (2003) *Diffusion of Innovations*. 5th edn. New York: Free Press.

Rospigliosi, P.A. (2023) 'Artificial intelligence in teaching and learning: what questions should we ask of ChatGPT?', *Interactive Learning Environments*, 31(1), pp. 1–3. doi: https://doi.org/10.1080/10494820.2023.2180191.

World Economic Forum (2024). *Education and Skills : The future of learning: How AI is revolutionizing education 4.0* [online] https://www.weforum.org/stories/2024/04/future-learning-ai-revolutionizing-education-4-0/ [Accessed 14 Nov 2024]

Zhou, K., Jurafsky, D. and Hashimoto, T. (2023) 'Navigating the grey area: Expressions of overconfidence and uncertainty in language models', *arXiv preprint arXiv:2302.13439*. Available at: https://doi.org/10.48550/ARXIV.2302.13439.

Precursors of Master Data Quality Issues across Enterprise Systems

Kim Keith
University of Cape Town

Lisa Seymour
University of Cape Town

Completed Research

Abstract

The data quality of master data, and data governance, has been under researched despite the increase in importance and relevancy to technologies such as AI and for data-driven decision making. To understand why there is a disjoint between the need for accurate master data and the lack of attention that it has been given in research and organisations, a systematic literature review was conducted for relevant Information Systems papers over the last ten years. The Work Systems Framework was used to interpret possible precursors of master data quality issues found from 56 relevant articles. Most of the precursors were found to be related to information, and infrastructure. This may indicate that there are important precursors of master data quality that may or may not be entirely in the control of the work system itself.

Keywords: Data Governance, Data Quality, Enterprise Systems, Master Data, Master Data Management, Work Systems Framework

1.0 Introduction

Data has become a first-class citizen that is akin to code (Whang et al., 2023). Two reasons for this, include firstly the immense pressure for organisations to make accurate decisions in a timely manner (Elragal & Elgendy, 2024). Yet confident reporting and decision-making requires trusting institutional data (Wende, 2007). Secondly, the increase of Artificial Intelligence (AI) usage (Revilla et al., 2023). As AI use increases, there is acknowledgement that machine learning algorithms can't operate efficiently without good data. Yet data in the real world rarely possess the necessary data quality (Whang et al., 2023). While organisations are often keen to experiment with AI, they do not allocate the prerequisite time to ready the organisational data for it (Janssen et al., 2020).

High data quality is an essential requirement for data-driven decision making (Jia et al., 2015). Data quality is a multifaceted phenomenon that encompasses features such as accessibility, accuracy, believability, completeness, consistency, fitness for use, interpretability, relevancy, security and timeliness (Janssen et al., 2017; Pipino et al., 2002; Schäffer & Stelzer, 2017; Strong et al., 1997; Wand & Wang, 1996). In the absence of quality data, user confidence in report results is jeopardised (Smith & McKeen, 2008). Data integration into Enterprise Systems (ES) is integral to addressing data quality issues within organisations but faces the challenges from ineffective organisational processes, lack of management support and poor data management (Haug et al., 2023).

ES are large-scale packaged software systems that are implemented to integrate the business processes in organisations (Bhattacharya et al., 2010; Shang & Seddon, 2002). Formerly ES were mainly referred to as Enterprise Resource Planning (ERP) systems but, more recently, ES go beyond traditional ERP functionality as ES functionality encompasses most business processes and management functions of an organisation and includes Customer Relationship Management (CRM), Supply Chain Management (SCM) as well as other large scale organisational systems (Lech, 2024; Roztocki et al., 2021).

Master data forms an fundamental connection between ES and business processes because it is core in creating one version of the truth as all of an organisation's transactions are created against master data (Hannila et al., 2022). Master data is critical data about core organisational activities that are elemental to the continuation processes within the organisation and rarely change (Schäffer & Stelzer, 2017; Spruit & Pietzka, 2015). Master Data Management (MDM) is a common practice used by organisations to improve their data quality (Haug et al., 2023). MDM relies on data integration to share accurate data throughout an organisation (Vilminko-Heikkinen & Pekkola, 2017). The creation of rules to aid data quality fall under the ambit of Data Quality Management (DQM) which involves the collection, organisation, storage and presentation of high-quality data (Wende, 2007). Master Data Management (MDM) encompasses the DQM of inconsistent master data from different internal and external sources (Loser et al., 2004). Organisation-wide accountabilities for DQM can be implemented via Data Governance which takes the organisational strategy and legal requirements into account (Wende, 2007). Data Governance concerns the provisioning of decision-making responsibilities within an organisation (Otto, 2011).

Common issues in the maintenance of high-quality master data include: A lack of clarity regarding who is accountable or responsible for MDM at an organisation; insufficient or an absence of policies and procedures related to MDM; and the lack of support from management (Vilminko-Heikkinen & Pekkola, 2019). Since organisations rely on data which are fundamental in every transaction (Hannila et al., 2022), the volumes of data are increasing (Haneem et al., 2019) and the need for data quality is vital in current business environments (Patel et al., 2024), it has become increasingly more important to address master data issues in ES. To address the issues of the data quality of master data, a systematic, stand-alone, literature review was undertaken to understand the precursors of poor data quality across ES, leading to the research question: **What leads to master data quality issues across enterprise systems?**

The next section of this paper details a systematic, stand-alone literature review of the precursors of master data quality issues. These issues are categorised using the nine elements from the Work Systems Framework (Alter, 2013) and then discussed in the findings section. Finally, the conclusion summarises the findings and indicates shortcomings of the research and further research to pursue.

2.0 Literature Review Method

Literature reviews are valuable to the IS community, for researchers who would like to commence a research project and to classify what has been produced in the literature (Alhassan et al., 2018; Rowe, 2014). The perspective of this literature review is to address an emerging issue that would benefit from theoretical explanation (Webster & Watson, 2002). In this case, the precursors of master data quality issues that could be described and better understood through the Work Systems Framework (WSF) by Alter (2013) as an established framework that has previously been used to describe data as an organisational product (Hasan & Legner, 2023). In this way, the researchers aim to show knowledge of the literature in the field of study that they are aiming to contribute to (Ngwenyama, 2019).

According to Okoli (2015), a rigorous stand-alone literature review must be systematic in aligning with the methodological approach, explicit in explaining the way in which it was done, comprehensive in including all relevant articles and reproducible for others to be able to conduct a review using the same steps (Okoli, 2015). Okoli (2015) presents an eight-step methodology for conducting a stand-alone literature review. These steps are:

1. Identify the purpose: to identify the purpose of the review and what the goals of the review are so that the readers are clear about what the review intends to do.
2. Draft protocol and train the team: this entails all authors of the review to agree on how the review will be conducted.
3. Apply practical screen: this step involves screening appropriate literature for inclusion or exclusion from the review. Reasons need to be provided regarding the inclusion or exclusions of literature.
4. Search for literature: the method of searching for appropriate literature needs to be explicit in a way that readers can be assured of the comprehensiveness of the search.
5. Extract data: appropriate information needs to be extracted from the literature.
6. Appraise quality: this step is also referred to as screening for exclusion where the researchers need to be transparent about the criteria they are using to exclude any literature where there is insufficient quality, depending on the research methodology they are using.
7. Synthesize studies: this analysis step includes amalgamating facts gleaned from the appropriate literature. This analysis can be qualitative, quantitative or both.
8. Write the review: the results of the literature review need to be written in a way that the results can be independently reproduced by other researchers (Okoli, 2015).

2.1 Identify the Purpose

The aim of the systematic literature review was to look for information around the key concepts of the research question (Templier & Paré, 2015). These key concepts were the exact terms “data quality” and “master data”. Although master data quality is not clearly defined, it can be derived by combining the concepts, namely master data and data quality (Schäffer & Leyh, 2017).

2.2 Draft Protocol and Train the Team

The authors of the paper agreed to the coding that would be used, including the categorisation of the codes using the WSF.

2.3 Apply Practical Screen

From the research question, the terms “Data Quality” and “Master Data” were core terms needing to guide the screening process. ES was not used as a specific search term because, the term “Enterprise Systems” was not always specifically included in research papers when referring to systems such as CRM, SCM, ERP etc. ES is a more recent term encompassing other systems such as ERP etc. (Lech, 2024; Roztocki et al., 2021). To demonstrate a command of the literature in this field of study (Ngwenyama, 2019), a comprehensive search would be required. LitBaskets XL (extra-large) searches over 154 essential IS journals which would be a comprehensive search in the field of IS. A researcher using the same tool would be guaranteed to get the same results, and this would assist the reproducibility of the results recommended by Okoli (2015).

2.4 Search for Literature

The search was conducted on 31 May 2024, is outlined in Figure 1 and is based on the research procedure by Günther et al. (2017).

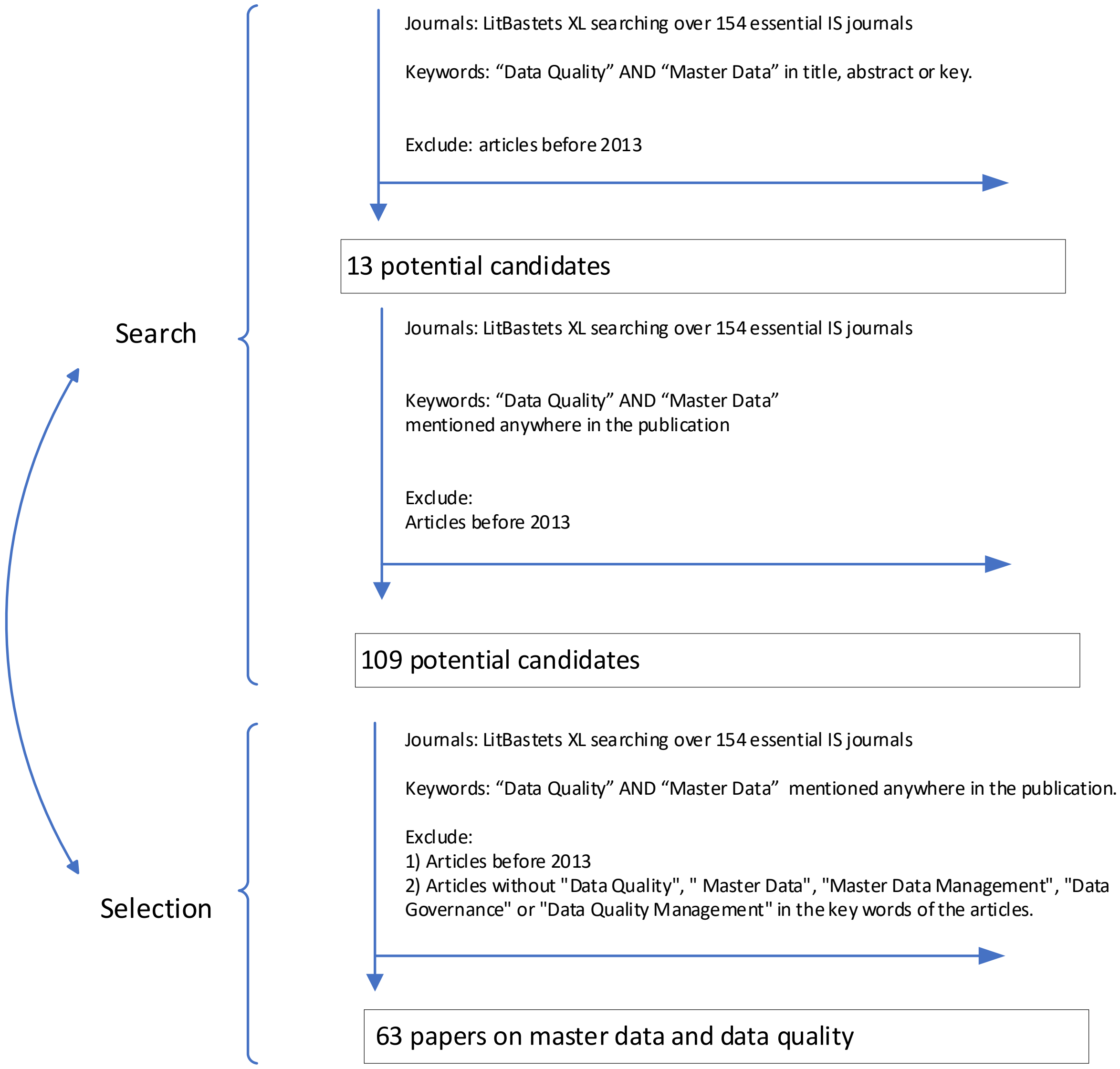

Search
Journals: LitBastets XL searching over 154 essential IS journals
Keywords: “Data Quality” AND “Master Data” in title, abstract or key.
Exclude: articles before 2013
13 potential candidates
Journals: LitBastets XL searching over 154 essential IS journals
Keywords: “Data Quality” AND “Master Data” mentioned anywhere in the publication
Exclude:
Articles before 2013
109 potential candidates
Selection
Journals: LitBastets XL searching over 154 essential IS journals
Keywords: “Data Quality” AND “Master Data” mentioned anywhere in the publication.
Exclude:
1) Articles before 2013
2) Articles without "Data Quality", " Master Data", "Master Data Management", "Data Governance" or "Data Quality Management" in the key words of the articles.
63 papers on master data and data quality

Figure 1: Research Procedure

The inclusion and exclusion of sources needs to be transparent in order for the literature review to be credible (Okoli, 2015). Initially, a search was conducted through Litbaskets XL (extra-large) in Scopus for both exact terms “Data Quality” and “Master Data” in the title, abstract or key of the papers in the Litbaskets. This resulted in only 13 papers. The search code was then manipulated to include “ALL” fields instead of “TITLE-ABS-KEY” to increase the results. This resulted in 109 results but, in order to make sure that these words were not only in the references or other non-significant parts of the papers, one the exact words “Data Quality”, “Master Data”, “Master Data Management”, “Data Governance” and “Data Quality Management” needed to feature in the key words. Certain terms, such as Data Governance (Goel et al., 2024; Walsh et al., 2022) and MDM (Silvola et al., 2016) have been under researched, only recently gaining attention. For this reason, the researcher limited the search to the last ten years of articles. The amended search query, omitting the specific journal source IDs is as follows:

> *ALL ("Master Data" AND "Data Quality") AND PUBYEAR > 2013 AND (LitBaskets XL list) AND (LIMIT-TO (EXACTKEYWORD , "Data Quality") OR LIMIT-TO (EXACTKEYWORD , "Master Data Management") OR LIMIT-TO (EXACTKEYWORD , "Master Data") OR LIMIT-TO (EXACTKEYWORD , "Data Governance") OR LIMIT-TO (EXACTKEYWORD , "Data Quality Management")).*

2.5 Extract Data

All 63 papers were added to NVivo and analysed to ascertain what kind of master data quality issues were being experienced in ES in alignment to the research question.

2.6 Appraise Quality

This step, which is also referred to as screening for exclusion (Okoli, 2015), was conducted to ascertain which articles would be excluded. Articles that did not reveal any codes relevant to the precursors of master data quality issues were excluded for further analysis. This reduced the number of papers to 56, listed in Table 1.

Source Title	Articles
ACM Transactions on Database Systems	1
Computers in Human Behavior	1
Data and Knowledge Engineering	2
Decision Support Systems	1

Source Title	Articles
Electronic Markets	2
Enterprise Information Systems	1
Expert Systems with Applications	1
IEEE Intelligent Systems	1
IFIP Advances in Information and Communication Technology	4
Industrial Management and Data Systems	1
Information Processing and Management	1
Information Sciences	1
Information Systems	3
Information Systems Frontiers	2
Information Systems Management	2
International Journal of Business Information Systems	3
International Journal of Information Management	4
International Journal of Medical Informatics	1
Journal of Computer Information Systems	1
Journal of Data and Information Quality	7
Journal of Decision Systems	2
Journal of Enterprise Information Management	4
Journal of Information Science	1
Journal of Information Technology Management	1
Journal of Systems and Software	2
Lecture Notes in Business Information Processing	2
Lecture Notes in Information Systems and Organisation	1
Personal and Ubiquitous Computing	1
VLDB Journal	1
World Wide Web	1
Grand Total	56

Table 1. Papers Reporting Precursors of Master Data Quality Issues

2.7 Synthesize studies

The literature review needs to include analytical critique of theory while synthesising (Okoli, 2015). According to Alter (2013), a work system is a useful way to think about, and analyse, systems in organisations. The WSF is a framework that can be used to describe and analyse an organisational Information Technology (IT) work system through the understanding of the nine elements that form the basis of it (Alter, 2013). The WSF, illustrated in Figure 2, is an appropriate and established lens for data research as it covers the creation of useful products or services for consumers of data that has been collected from different sources (Hasan & Legner, 2023).

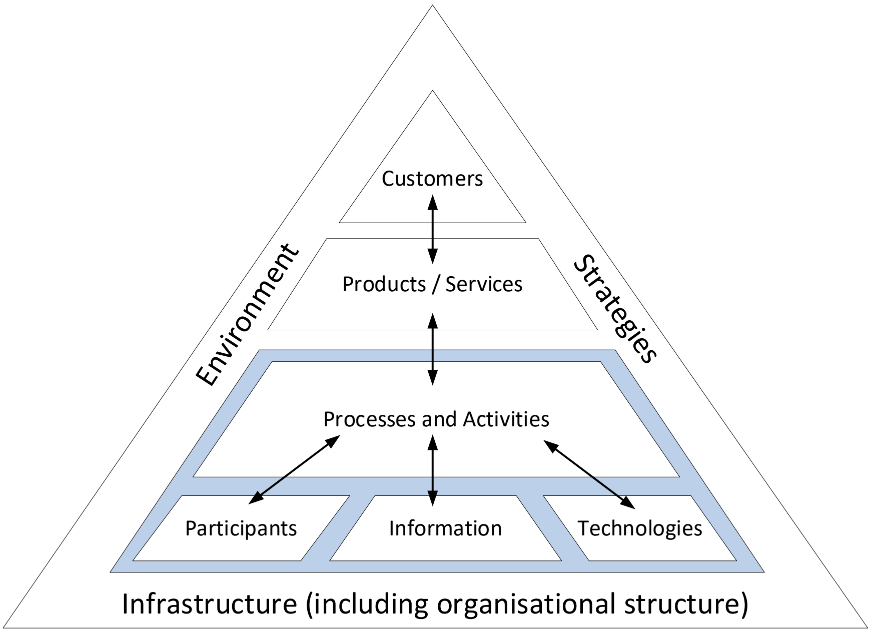


Figure 2: Adapted from the Work Systems Framework (Alter, 2013, p. 78)

In the shaded section, the *Processes and Activities*, *Participants*, *Information*, and *Technologies* are seen as core to the work system (Alter, 2013). *Processes and Activities* form the foundation for the production of products and services for customers and are the first elements of a work system (Lindgren et al., 2021). *Processes* may not be well defined but can be viewed from a current performative perspective (Alter, 2013). *Participants* can be both users and non-users of IT systems who are responsible for the work within the work system (Alter, 2013). *Information* is created in all work systems and, in the context of a work system the distinction between data and information is not relevant as well as whether the data is computerised or not (Alter, 2013). In the case of ES, this would be particularly relating to master data. *Technologies* form the basis of almost all operational work systems and include tools used by participants as well as fully automated agents (Alter, 2013). Categories from the literature coded to these elements were internal to the work system.

Products/Services are the indication of a work system's effectiveness and refer to the products available for the *Customers*, who can also be *Participants* in the work system, if they are recipients of the *Products/Services* (Alter, 2013). *Environment* refers to stakeholders, policies and procedures as well as the cultural, technical, regulatory, political, historic or demographic space that the work system works within and can affect its performance (Alter, 2013). *Infrastructure* refers to the technical, information and human resources that are used inside the work system but are managed and shared outside of it with other work systems (Alter, 2013). *Strategies* can exist at various hierarchical levels from the enterprise to the department and to the work system and should ideally be in alignment and supported by the work system although they may be inconsistent with the understanding or beliefs of key stakeholders (Alter, 2013). These elements external to the work system relate to the categories that are indirectly involved in creating the master data quality issues.

The researchers looked at each element of the WSF and allocated each code to the most appropriate category. Specific care was taken to distinguish which codes were internal and external to the work system. For example, categories relating to the organisational structure were placed in the *Infrastructure* element and not the *Participants* element because, in the context of master data quality, employees within the work system work in departments or units from the organisational structure, where the department or unit managers are not part of the work system and the employees can also be involved in other work or projects external to the work system. The definition of *Infrastructure* caters for this scenario i.e. "relevant human, information, and technical resources that are used by the work system but are managed outside of it and are shared with other work systems" Alter (2013, p. 81).

2.8 Write the review

The writing up of the review proceeds in the next section with the codes emerging from the literature organised into the nine basic elements of the WSF.

3.0 Findings

The codes were allocated to relevant sections of the WSF. The number of papers mentioning each issue are reflected in Table 2 and expanded on in this section.

WSF Element	Issues From Literature	Number of Papers Mentioned
Processes and Activities	Increase in complexity	30
	Inefficient organisational processes	5
	Manual data capturing	4
	Deadlines and timelines	3

WSF Element	Issues From Literature	Number of Papers Mentioned
	Inadequate application of business rules	1
Participants	Skills of the users	27
	Lack of data ownership	13
	User participation	1
Information	Source data issues	46
	Increasing data volumes	36
	Insufficient naming, data standards, classification or semantics	34
	Intangibility of data	11
Technologies	Inadequate technological infrastructure or software applications	24
	Lack of data management automation	5
Products/Services	Data analysis and reporting requirements	21
Customers	Incorrect information from self-reporting	1
Environment	Organisational customs and culture	22
	Lack of importance given to data quality	14
	Ineffective communication	9
	Lack of, or unclear, roles and responsibilities	8
	Need for institutional knowledge	8
Infrastructure	Data from disparate sources	40
	Silos	22
	Lack of a global understanding of data or processes across an organisation	17
	Need for management support	12
	Increase in software applications	10
	Resource constraints	7
	Organisational structure challenges	6
	Organisational size	3
Strategies	Lack of data governance	16
	Lack of strategic alignment	4
	Lack of data quality management strategy	3

Table 2. Codes from literature organised according to the WSF

3.1 Processes and activities

Processes and activities from the literature include five codes. *Increase in complexity* was the most prevalent issue in this category, specifically mentioned in 30 of the papers. Different formats (Haneem et al., 2019), heterogenous technologies (Karkošková, 2023), rigid/inflexible software applications (Rohani & Yusof, 2023) and multiple uses and users of data (Silvola et al., 2016) seemed to be the precursors for the complexity. *Inefficient organisational processes* covered issues where existing processes need to be changed in order to fit in with new solutions (Patel et al., 2024), or are increasingly causing errors (Xu, 2015). *Manual data capturing* indicated that end users are capturing data in different ways (Rohani & Yusof, 2023) or making typing errors when capturing data (De et al., 2016). *Deadlines and timelines* refers to the pressure that the participants were under to complete system implementations despite not having enough time to do so thoroughly (Patel et al., 2024) or where the participants capturing data are under time pressure and make mistakes in doing so (Liu et al., 2020). *Inadequate application of business rules* referred to generating business rules at the incorrect level of granularity to be useful (Caballero et al., 2022). The latter three codes were found infrequently.

3.2 Participants

The *Participants* element had three codes. *Skills of the users* was the most frequently cited code, indicating that users, especially novices, often don't possess the required knowledge to enter data correctly (Liu et al., 2020). *Lack of data ownership* shows that organisations assumed the responsibility of the data quality of master data should reside in the IT department (Mlangeni & Ruhode, 2017) although IT were not process or master data owners (Vilminko-Heikkinen & Pekkola, 2017). *User participation* was infrequently mentioned as an issue related to user feedback and users taking part in data governance initiatives (Jiang et al., 2023).

3.3 Information

Information included four codes that occurred frequently throughout the literature. *Source data issues* included missing data (Mezzanzanica et al., 2015a), incomplete data (Liu et al., 2020), inaccessible data (Wibisono et al., 2023), redundant data (Lohmer et al., 2021), data duplication (Vilminko-Heikkinen & Pekkola, 2019), inaccurate data (Abraham et al., 2019), inconsistent data (Patel et al., 2024), untimeliness of data (Silvola et al., 2019) and data not fit for purpose (Silvola et al., 2016).

Increasing data volumes was a widespread issue mentioned by 36 papers because of the wide use of software applications (Mezzanzanica et al., 2015b), and the rise of big data (De et al., 2016) in a way that organisations can't manage it (Vilminko-Heikkinen & Pekkola, 2019). *Insufficient naming, data standards, classifications or semantics* includes data that are poorly defined (Mlangeni & Ruhode, 2017), lack consistency in naming standards (Patel et al., 2024), where there are differing standards from the data sources (Vilminko-Heikkinen & Pekkola, 2017) or where there are issues with the semantics of the data (Mezzanzanica et al., 2015a). *Intangibility of data* ranges from issues were there was difficulty assigning monetary value to data (Hannila et al., 2022), the general abstract nature of data (Silvola et al., 2016) and the problem where data quality is controlled by qualitative parameters (Bodendorf & Franke, 2024).

3.4 Technologies

Technologies had two codes. *Inadequate technological infrastructure* was often uncovered when trying to integrate data and the integration technologies were outdated or inadequate (Wibisono et al., 2023), there were too many source applications (Karkošková, 2023), the software applications did not function adequately (Wibisono et al., 2023) or there were issues relating to the lack of skills of the people who needed to integrate the data (Valencia-Parra et al., 2021). *Lack of data management automation* covered the lack of a pre-existing master data management software application (Vilminko-Heikkinen & Pekkola, 2019) or the automation of data management tools in general (Wibisono et al., 2023).

3.5 Products/services

Products/services in the context of master data quality issues revolved around *Data analysis and reporting requirements* because there is a significant concern that the lack of quality can impact reporting and lead to incorrect decision-making (Hannila et al., 2022), including unsuccessful implementation of analytics capabilities (Liu et al., 2020) and barriers to finding, analysing and publishing data (Abraham et al., 2023).

3.6 Customers

Customers was not an element of the WSF that was widely covered, only being picked up in one paper, but *Incorrect information from self-reporting* was mentioned where respondents did not understand the context of the information requested from them, were not open about filling in the required information or moved around too frequently to track them down for the required data (Doharta et al., 2018).

3.7 Environment

Environment included five codes. *Organisational customers and culture* include where employees perform in a certain way that is not transparent to others (Wibisono et al., 2023), where there is tension between a new practice and the existing organisational culture (Vilminko-Heikkinen & Pekkola, 2019), where different organisations with their own established practices have been merged into one and collaboration is required (Mlangeni & Ruhode, 2017), also within one organisation where a new organisational practice is required (Vilminko-Heikkinen & Pekkola, 2017), where there is a collective action issue where employees sabotage a collective goal to rather realise their own, or group-specific, short-term benefits (Sæbø et al., 2020), employees are used to relying on their experience or intuition to make decisions (Hannila et al., 2022), employees are used to working within established silos (Hannila et al., 2020), where individuals believe that the data belongs to them and that they will lose control of the data if others have access to it (Walsh et al., 2022), where there are power issues when switching existing responsibilities from one individual to another, or due to general human resource issues (Foidl et al., 2024).

Lack of importance given to data quality included where management needed to be convinced that resources should be allocated to improving data quality (Benfeldt et al., 2020), there was a lack of academic research or empirical data available for data quality (Karpischek et al., 2014), there was a lack of awareness of existing data quality issues (Spruit & Pietzka, 2015), there was an existing lack of data literacy amongst employees (Jiang et al., 2023) or there was a lack of enthusiasm for data quality initiatives (Vilminko-Heikkinen & Pekkola, 2017).

Ineffective communication arose when there was a breakdown of communication the business and the IT department when implementing systems (Liu et al., 2020), formal communication channels were infrequently used during MDM initiatives (Vilminko-Heikkinen & Pekkola, 2017) and where there were communication gaps between various stakeholders during software implementations (Rohani & Yusof, 2023).

Lack of, or unclear, roles and responsibilities arose where there were no existing, dedicated roles and responsibilities for data quality (Doharta et al., 2018), data ownership and responsibilities are ambiguous (Vilminko-Heikkinen & Pekkola, 2019), there is uncertainty around whether or not data ownership should be allocated to IT, a business function or a separate governance structure (Abraham et al., 2019), there is the added complication of adding new data roles and responsibilities when introducing data governance (Karkošková, 2023) or where there are many stakeholders required to make data decisions (Vilminko-Heikkinen & Pekkola, 2017).

Need for institutional knowledge indicates that users need to understand data in a new context when merging data from different sources (Glowalla & Sunyaev, 2014), there are differences in understanding of key terms (Spruit & Pietzka, 2015), new data categories need to be created (Vilminko-Heikkinen & Pekkola, 2017), business units need to share data in order to determine the affect that the data will have on the entire organisation (Spruit & Pietzka, 2015) and expertise in a specific domain was required to interpret the data so that the correct results could be achieved (Huang et al., 2020).

3.8 Infrastructure

Infrastructure contains eight codes. *Data from disparate sources* covers the creation of a single point of reference for master data from many existing software applications (Al-Ruithe et al., 2019) that may be heterogenous, outdated and fragmented (Benfeldt et al., 2020), sometimes even incorporating data from spreadsheets and paper-based records requiring data capture or from self-reporting (Wibisono et al., 2023), the integration of which is a significant challenge to ensuring data quality and data management (Liu et al., 2020). *Silos* can occur from fragmented (Al-Ruithe et al., 2019) or isolated (Hannila et al., 2019) software applications, where employees preferred to work in existing groups (Haug et al., 2023) and they hinder decision-making and data sharing (Wibisono et al., 2023). L*ack of a global understanding of data or processes across an organisation* seems to occur where employees lack organisational-wide knowledge and they create business rules that are worthless for the organisation (Caballero et al., 2022), where someone with an organisational-wide understanding of master data is required but only employees with individual understandings are available (Benfeldt et al., 2020), that there are different levels of granularity of the data that needed (Vilminko-Heikkinen & Pekkola, 2017) and, in general, that more research is required for master data to be used throughout an organisation by connecting effectively with ES solution-specific data (Hannila et al., 2020).

Need for management support, sometimes specifically relating to top management support (Doharta et al., 2018) or at different management levels (Vilminko-Heikkinen

& Pekkola, 2017), is seen as an important factor in influencing the data quality within an organisation, from the point of data quality management (Liu et al., 2020), decision-making when investing in data quality initiatives (Vilminko-Heikkinen & Pekkola, 2019) or where passionate employees dedicated to data quality initiatives lack management support and eventually lose their passion (Benfeldt et al., 2020). I*ncrease in software applications* occurs where there is data in many different software applications that makes the management of the data within those applications difficult (Vilminko-Heikkinen & Pekkola, 2019). This can happen time as software applications are accumulated within an organisation (Benfeldt et al., 2020) and then the linking of the data becomes a complex task (Al-Ruithe et al., 2019).

Resource constraints covers human resources and the time needed to complete tasks when implementing software applications (Liu et al., 2020) where data quality issues are not detected (Silvola et al., 2016), where management doesn't prioritise MDM implementations (Haneem et al., 2019) or where resources aren't allocated because tasks aren't well enough understood (Vilminko-Heikkinen & Pekkola, 2017). *Organisational structure challenges* include where data is dispersed at various levels within an organisation (Mlangeni & Ruhode, 2017) or the organisational structure is complicated (Haneem et al., 2019) and the set-up of the organisation needs to be understood because it affected MDM (Spruit & Pietzka, 2015). *Organisational size* affected master data quality issues in terms of the responsibilities assigned to employees (Karkošková, 2023) as well as how well master data quality is understood (Vilminko-Heikkinen & Pekkola, 2017) especially where the integration of master data from different departments or units is required (Spruit & Pietzka, 2015).

3.9 Strategies

Strategies included three codes. *Lack of data governance* indicates that a lack of data governance may diminish the perceived value of data quality (Bodendorf & Franke, 2024), isn't holistic enough for the entire organisation (Hannila et al., 2020) or that the increasing volume of data makes the need for governance over the data more important (Hannila et al., 2019), although there are organisations that lacked data-governance at an organisational-wide level completely (Hannila et al., 2019). *Lack of strategic alignment* indicated that there is no alignment between the organisational strategy, the business processes and IT software application (Hannila et al., 2019), there was a general lack of an organisational-wide MDM strategy (Myung, 2016) although strategy, business processes and master data was seen as vital for organisational performance (Silvola et al., 2019).

3.10 Summary

The WSF was found to be a useful framework to describe the precursors of master data quality issues. Each category was carefully allocated to an element of the WSF based on the definitions of the elements. Particularly useful was the distinction between the elements internal and external to the work system, so that there's clarification between precursors emanating from the work system itself, or issues where there are external factors attributing to these issues. *Information* and *Infrastructure* were the elements with the largest number of categories emerging from the literature, as one internal element, and one external element respectively.

4.0 Conclusion

The data quality of master data is important for institutional trust and to support decision making, however, there are many precursors of master data quality issues that are affecting ES. Using the WSF, a holistic organisational perspective on master data quality was obtained, from an internal and external view of a work system. A recommendation for further research would be to expand the geographical areas or types of organisations investigated for the precursors of master data quality issues. Only one paper was written based on issues found in Africa (Mlangeni & Ruhode, 2017), and described a specific case relating to the merging in organisations, whereas three papers studied universities (Guerra-García et al., 2023; Mlangeni & Ruhode, 2017; Wibisono et al., 2023). The authors intend to explore a case from Africa and in a university context in future work.

Although the WSF was useful in describing precursors, relationships between precursors would be useful to further interrogate and is therefore a limitation of this research. For example, the increasing number of software applications may have been caused by software applications being implemented in silos, this may lead to an increase in data, the increase in complexity of data management and/or the reason why employees are unclear about data ownership roles and responsibilities. Yet this cannot be determined through the WSF alone. It may be useful for both practitioners and academics to further understand the precursors of master data quality issues by understanding which mechanisms cause them in an appropriate case study.

References

Abraham, R., Schneider, J. and vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. International Journal of Information Management, 49 424-438.

Abraham, R., Schneider, J. and vom Brocke, J. (2023). A taxonomy of data governance decision domains in data marketplaces. Electronic Markets, 33 Article 22.

Al-Ruithe, M., Benkhelifa, E. and Hameed, K. (2019). A systematic literature review of data governance and cloud data governance. Personal and Ubiquitous Computing, 23 839-859.

Alhassan, I., Sammon, D. and Daly, M. (2018). Data governance activities: a comparison between scientific and practice-oriented literature. Journal of Enterprise Information Management, 31 300-316.

Alter, S. (2013). Work system theory: overview of core concepts, extensions, and challenges for the future. Journal of the Association for Information Systems, 72.

Benfeldt, O., Persson, J. S. and Madsen, S. (2020). Data Governance as a Collective Action Problem. Information Systems Frontiers, 22 299-313.

Bhattacharya, P. J., Seddon, P. B. and Scheepers, R. (2010). Enabling strategic transformations with enterprise systems: Beyond operational efficiency. MIS quarterly, 34 731-756.

Bodendorf, F. and Franke, J. (2024). What is the business value of your data? A multi-perspective empirical study on monetary valuation factors and methods for data governance. Data and Knowledge Engineering, 149 Article 102242.

Caballero, I., Gualo, F., Rodríguez, M. and Piattini, M. (2022). BR4DQ: A methodology for grouping business rules for data quality evaluation. Information Systems, 109 Article 102058.

De, S., Hu, Y., Meduri, V. V., Chen, Y. and Kambhampati, S. (2016). BayesWipe: A scalable probabilistic framework for improving data quality. Journal of Data and Information Quality, 8 Article 5.

Doharta, I. A., Hidayanto, A. N., Budi, N. F. A., Samik Ibrahim, R. M. and Solikin, S. (2018). Framework for prioritizing solutions in overcoming data quality problems using analytic hierarchy process (AHP). Journal of Information Technology Management, 10 27-40.

Elragal, A., & Elgendy, N. (2024). A data-driven decision-making readiness assessment model: The case of a Swedish food manufacturer. Decision Analytics Journal, 10 Article 100405.

Foidl, H., Golendukhina, V., Ramler, R. and Felderer, M. (2024). Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers. Journal of Systems and Software, 207 Article 111855.

Glowalla, P., & Sunyaev, A. (2014). ERP system fit – An explorative task and data quality perspective. Journal of Enterprise Information Management, 27 668-686.

Goel, K., Martin, N. and ter Hofstede, A. (2024). Demystifying data governance for process mining: Insights from a Delphi study. Information and Management, 61 Article 103973.

Guerra-García, C., Nikiforova, A., Jiménez, S., Perez-Gonzalez, H. G., Ramírez-Torres, M. and Ontañon-García, L. (2023). ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design. Data and Knowledge Engineering, 145 Article 102152.

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M. and Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. The Journal of Strategic Information Systems, 26 191-209.

Haneem, F., Kama, N., Taskin, N., Pauleen, D. and Abu Bakar, N. A. (2019). Determinants of master data management adoption by local government organizations: An empirical study. International Journal of Information Management, 45 25-43.

Hannila, H., Koskinen, J., Harkonen, J. and Haapasalo, H. (2019). Product-level profitability. Journal of Enterprise Information Management, 33 214-237.

Hannila, H., Kuula, S., Harkonen, J. and Haapasalo, H. (2020). Digitalisation of a company decision-making system: a concept for data-driven and fact-based product portfolio management. Journal of Decision Systems, 31 258-279.

Hannila, H., Silvola, R., Harkonen, J. and Haapasalo, H. (2022). Data-driven Begins with DATA; Potential of Data Assets. Journal of Computer Information Systems, 62 29-38.

Hasan, M. R. and Legner, C. (2023). Decoding Data Products–Through the Lens of Work System Theory. In Proceedings of Forty-Fourth International Conference on Information Systems (ICIS), AIS, Hyderabad, India.

Haug, A., Staskiewicz, A. M. and Hvam, L. (2023). Strategies for Master Data Management: A Case Study of an International Hearing Healthcare Company. Information Systems Frontiers, 25 1903-1923.

Huang, Y., Milani, M. and Chiang, F. (2020). Privacy-aware data cleaning-as-a-service. Information Systems, 94 Article 101608.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S. and Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. Government information quarterly, 37 Article 101493.

Janssen, M., van der Voort, H. and Wahyudi, A. (2017). Factors influencing big data decision-making quality. Journal of Business Research, 70 338-345.

Jia, L., Hall, D. and Song, J. (2015). The conceptualization of data-driven decision making capability. In Proceedings of Twenty-first Americas Conference on Information Systems, Puerto Rico.

Jiang, G., Cai, X., Feng, X. and Liu, W. (2023). Effect of data environment and cognitive ability on participants' attitude towards data governance. Journal of Information Science, 49 740-761.

Karkošková, S. (2023). Data Governance Model To Enhance Data Quality In Financial Institutions. Information Systems Management, 40 90-110.

Karpischek, S., Michahelles, F. and Fleisch, E. (2014). Detecting incorrect product names in online sources for product master data. Electronic Markets, 24 151-160.

Lech, P. (2024). Enterprise Systems implementation projects: waterfall, agile or hybrid? In Proceedings of Thirtieth Americas Conference on Information Systems (AMCIS), AIS, Salt Lake City, USA.

Lindgren, I., Melin, U. and Sæbø, Ø. (2021). What is E-Government? Introducing a Work System Framework for Understanding E-Government. Communications of the Association for Information Systems, 48 503-522.

Liu, C., Zowghi, D. and Talaei-Khoei, A. (2020). An empirical study of the antecedents of data completeness in electronic medical records. International Journal of Information Management, 50 155-170.

Lohmer, J., Bohlen, L. and Lasch, R. (2021). Blockchain-Based Master Data Management in Supply Chains: A Design Science Study. In Proceedings of IFIP Advances in Information and Communication Technology (APMS), IFIP, Nantes, France.

Loser, C., Legner, C. and Gizanis, D. (2004). Master Data Management for Collaborative Service Processes. In Proceedings of the 1st International Conference on Service Systems and Service Management (ICSSSM'04), IEEE, Beijing, China.

Mezzanzanica, M., Boselli, R., Cesarini, M. and Mercorio, F. (2015a). A model-based approach for developing data cleansing solutions. Journal of Data and Information Quality, 5 13-28.

Mezzanzanica, M., Boselli, R., Cesarini, M. and Mercorio, F. (2015b). A model-based evaluation of data quality activities in KDD. Information Processing and Management, 51 144-166.

Mlangeni, T. and Ruhode, E. (2017). Data governance: A challenge for merged and collaborating institutions in developing countries. In Proceedings of IFIP Advances in Information and Communication Technology, IFIP, Yogyakarta, Indonesia.

Myung, S. (2016). Master data management in PLM for the enterprise scope. In Proceedings of IFIP Advances in Information and Communication Technology, IFIP, Doha, Qatar.

Ngwenyama, O. (2019). The Ten Basic Claims of Information Systems Research: An Approach to Interrogating Validity Claims in Scientific Argumentation. Available at SSRN 3446798.

Okoli, C. (2015). A guide to conducting a standalone systematic literature review. Communications of the Association for Information Systems, 37 Article 43.

Otto, B. (2011). Data Governance. WIRTSCHAFTSINFORMATIK, 53 235-238.

Patel, D. S., Asamoah, D. A. and Wamwara, W. (2024). Data management for customer relationship management: a web-based approach. International Journal of Business Information Systems, 45 343-374.

Pipino, L. L., Lee, Y. W. and Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45 211-218.

Revilla, E., Saenz, M. J., Seifert, M. and Ma, Y. (2023). Human–Artificial Intelligence Collaboration in Prediction: A Field Experiment in the Retail Industry. Journal of Management Information Systems, 40 1071-1098.

Rohani, N. and Yusof, M. M. (2023). Unintended consequences of pharmacy information systems: A case study. International Journal of Medical Informatics, 170 Article 104958.

Rowe, F. (2014). What literature review is not: diversity, boundaries and recommendations. European Journal of Information Systems, 23 241-255.

Roztocki, N., Strzelczyk, W., & Weistroffer, H. R. (2021). Driving Forces in Enterprise Systems Implementation in the Public Sector: A Conceptual Framework. In Proceedings of 13th Annual AIS SIG GlobDev Pre-ICIS Workshop, AIS, Austin, USA.

Sæbø, Ø., Federici, T. and Braccini, A. M. (2020). Combining social media affordances for organising collective action. Information Systems Journal, 30 699-732.

Schäffer, T. and Leyh, C. (2017). Master data quality in the era of digitization - toward inter-organizational master data quality in value networks: A problem identification. In Proceedings of International conference on enterprise resource Planning Systems, ISER, Hagenberg, Austria.

Schäffer, T. and Stelzer, D. (2017). Assessing Tools for Coordinating Quality of Master Data in Inter-organizational Product Information Sharing. In Proceedings of 13th International Conference on Wirtschaftsinformatik, AIS, St. Gallen, Switzerland.

Shang, S. and Seddon, P. B. (2002). Assessing and managing the benefits of enterprise systems: the business manager's perspective. Information Systems Journal, 12 271-299.

Silvola, R., Harkonen, J., Vilppola, O., Kropsu-Vehkapera, H. and Haapasalo, H. (2016). Data quality assessment and improvement. International Journal of Business Information Systems, 22 62-81.

Silvola, R., Tolonen, A., Harkonen, J., Haapasalo, H. and Mannisto, T. (2019). Defining one product data for a product. International Journal of Business Information Systems, 30 489-520.

Smith, H. A. and McKeen, J. D. (2008). Developments in practice XXX: master data management: salvation or snake oil? Communications of the Association for Information Systems, 23 Article 4.

Spruit, M. and Pietzka, K. (2015). MD3M: The master data management maturity model. Computers in Human Behavior, 51 1068-1076.

Strong, D., Lee, Y. W. and Wang, R. Y. (1997). Data quality in context. Communications of the ACM, 40 103-111.

Templier, M. and Paré, G. (2015). A Framework for Guiding and Evaluating Literature Reviews. Communications of the Association for Information Systems, 37 112-137.

Valencia-Parra, Á., Parody, L., Varela-Vaca, Á. J., Caballero, I. and Gómez-López, M. T. (2021). DMN4DQ: When data quality meets DMN. Decision Support Systems, 141 Article 113450.

Vilminko-Heikkinen, R. and Pekkola, S. (2017). Master data management and its organizational implementation: An ethnographical study within the public sector. Journal of Enterprise Information Management, 30 454-475.

Vilminko-Heikkinen, R. and Pekkola, S. (2019). Changes in roles, responsibilities and ownership in organizing master data management. International Journal of Information Management, 47 76-87.

Walsh, M. J., McAvoy, J. and Sammon, D. (2022). Grounding data governance motivations: a review of the literature. Journal of Decision Systems, 31 282-298.

Wand, Y. and Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39 86-96.

Webster, J. and Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly, 26 xiii-xxiii.

Wende, K. (2007). A Model for Data Governance - Organising Accountabilities for Data Quality Management. In Proceedings of 18th Australasian Conference on Information Systems (ACIS), ACIS, Toowoomba, Australia.

Whang, S. E., Roh, Y., Song, H. and Lee, J.-G. (2023). Data collection and quality challenges in deep learning: a data-centric AI perspective. The VLDB Journal, 32 791-813.

Wibisono, A., Sammon, D. and Heavin, C. (2023). Approaches to Identifying Data Quality Issues: The Role of the Data Broker. Information Systems Management, 41 226-237.

Xu, H. (2015). What are the most important factors for accounting information quality and their impact on AIS data quality outcomes? Journal of Data and Information Quality, 5 14-22.

A Scoping Literature Review of the IS Research on Equality, Equity, Diversity, and Inclusion

Liucen Pan
Newcastle University
Mo Moeini
University of Warwick
João Baptista
Lancaster University

Completed Research

Abstract

This literature review aims to enhance the current understanding of the knowledge and gaps concerning Equality, Equity, Diversity, and Inclusion (EEDI) in relation to technology within the field of Information Systems (IS). Utilising a scoping review methodology, this study identifies six emerging themes and eighteen corresponding sub-themes that illustrate how EEDI and technology have been explored in the top eleven IS journals. Additionally, three key relationships between EEDI and technology are identified, demonstrating how the two concepts influence one another within the context of IS literature. This study provides critical insights and implications for future research on EEDI and technology by identifying the current knowledge gaps within the IS field.

Keywords: Equity Equality Diversity Inclusion, Scoping Review, Information Systems

Introduction

Equality, Equity, Diversity, and Inclusion (EEDI) are crucial concepts in both social and organizational contexts. Initially, EEDI initiatives were introduced to create equal opportunities for marginalized groups, focusing on "ethnicity, race, age, religious beliefs, sexual orientations, and physical or mental disabilities" (Hellerstedt and Wennberg, 2024, p. 24). The goal is to achieve an inclusive society that treats everyone fairly (Thomas, 1992) and "unites individuals" (Hellerstedt and Wennberg, 2024, p. 25).

This belief has also been adopted by managers in organizational settings, where equal employment opportunities have begun to emerge, leading to increased representation of minority groups in the workplace (Grosby et al., 2006). Recent management studies have found that EEDI contributes significantly to organizational success, particularly in the literature on diversity management, as evidenced by the works of Bassett-Jones (2005) and Gupta (2013). Studies suggest that workplace diversity fosters innovation, creativity, and problem-solving through the diverse information and capabilities of individuals (Miller and del Carmen Traina, 2009). Inclusion, on the other hand, focuses on actively involving diverse individuals in the workplace, especially minority groups (Crosby et al., 2006). The goal of inclusion is to mitigate discrimination and prejudice, creating equal opportunities for every employee to access organizational resources regardless of their diverse backgrounds (Fleurbaey et al., 2017). Implementing EEDI practices in the workplace, therefore helps organizations reap the benefits of diversity and gain competitive advantages (Hellerstedt and Wennberg, 2024).

Fostering EEDI is thus essential in both social and organizational contexts, especially with the support of technology. For example, the well-known "Black Lives Matter" movement, represented by a hashtag found on social media platforms and protest posters, highlights the discrimination, racism, and inequality experienced by black individuals (BBC News, 2021). In the field of Information Systems (IS), studies show that the adoption of technology is crucial for both social and organizational EEDI. For instance, research has discussed using information and communication technologies (ICT) to enhance social inclusion for newly resettled refugees in host societies (Andrade and Doolin, 2016). Additionally, studies have examined the use of ICTs to support group decision-making processes among diverse employees to drive organisational success (Shachaf, 2008).

However, there is limited understanding of how EEDI is studied specifically in the IS field. Despite we have seen some studies mentioned the importance of having EEDI practices in IS contexts, for example, Marabelli and Newell (2023) discussed strategical implications and EDI considerations of metaverse, suggesting that "leveraging the metaverse strategically will require ethical and DEI considerations" (Marabelli and Newell, 2013, p. 1). Others, like Fedorowicz et al., (2023) explore the EDI in the IS discipline, focusing on identifying the underrepresented opportunities for minority groups in the IS academic community and suggesting further improvements on EDI practises. Our knowledge on how EEDI has been studied, especially with its interaction with technology, such as the relationship with IT and how IT contributes to EEDI and in what ways, is limited.

Therefore, this study aims to scope the topics of EEDI that have been researched and what they have been studied within the IS discipline. Our research question therefore is: what do we know about EEDI in the IS field? To answer this question, we employ a scoping literature review (Munn et al., 2018). We consider the scoping literature review to be the most suitable method for this study. Unlike other literature review methods, such as systematic literature review, which focuses on exploring international evidence on a given topic and identifying and addressing areas of research for future directions (Munn et al., 2018), the scoping literature review serves as "a precursor to a systematic review" and aims to uncover the types of available evidence in a field, and to examine how research is conducted on a certain topic or field, further "to identify and analyse knowledge gaps" (Munn et al., 2018, p. 2).

By conducting this scoping literature review, we contribute to the field of IS regarding EEDI in several ways. First, we examine 6 emerging themes of EEDI ad IT, with further 18 sub-themes in the IS field. Second, we identify three relationships between EEDI and technology that have been studied in the existing research articles. Lastly, we identify the knowledge gaps concerning EEDI and technologies and propose future research directions in this regard.

Background

In this section, we provide a conceptual background of EEDI based on existing literature. Understanding the definitions of each key concept will help us to sharpen our focus, particularly in clarifying our selection criteria for the articles included and excluded in the literature pool.

Diversity, in general, "is concerned with how individuals are classified across a range of attributes, where different groups are defined based on their similarity in terms of these attributes" (Hellerstedt and Wennberg, 2024, p. 27). These attributes can be observable or unobservable (Kilduff et al., 2000). Observable attributes include age, gender, ethnicity, nationality, and organizational tenure, while unobservable attributes are associated with individuals' beliefs, attitudes, and values (Kilduff et al., 2000). Harrison and Klein (2007) further note that diversity can be defined as separation, variety, or disparity among groups of individuals within organizations. Diversity as separation refers to differences in positions and opinions among group members; diversity as variety relates to differences in kind, source, or categories such as functional background, network ties, and industry experience; and diversity as disparity is associated with differences in "propositions of socially valued assets or resources held among unit members" (Harrison and Klein, 2007, p.1203). Diversity as separation and disparity also involves polarization and power, which may cause interpersonal conflicts and inequalities, hindering group members' involvement and inclusion (Hellerstedt and Wennberg, 2024, p. 28).

Inclusion, another important concept in this study, is related to but distinct from diversity. Roberson (2006) noted that the definition of diversity mainly focuses on individuals' similarities and differences, and their composition as a group within the organization. In contrast, inclusion focuses on organizational objectives that increase the participation and involvement of all employees, as well as leveraging the benefits of diversity management. Transforming from a monocultural, exclusive organization to an inclusive one requires organizations to seek and value all differences and develop systems and work practices that support all employees to contribute fully and equally within the group (Holvino et al., 2004). Inclusion also concerns individuals' psychological experiences (Ferdman, 2014). The experience of inclusion means "feeling safe, trusted, respected, supported, valued, fulfilled, engaged, and authentic in one's working environment, both as individuals and as members of particular identity groups" (Ferdman et al., 2009, p.6). Socially, inclusion involves having policies, norms, values, practices, and ideologies that support all individuals being fairly treated, allowing them to fully and equally belong to and participate in the larger society (Ferdman, 2014).

Equality, briefly touched upon in the previous section of inclusion, is crucial to consider in EEDI. Equality indicates that all individuals are offered the same opportunities, regardless of their personal characteristics (Berube et al., 2024). In this study, we also include another concept, equity. Equity, which is closely related to equality but has distinct meanings. Berube et al. (2024) define equity as justice, where individuals, regardless of their identity and background, are treated fairly, ensuring that resource allocation and decision-making mechanisms do not discriminate based on individual characteristics. This concept implies that “people stand in equal relationship to each other, rather than being treated better or worse” (Fourie, 2012, p. 112), even though some individuals may have a claim to additional support and resources to enable them to function as equals in society (Anderson, 2010). We will use these concepts to guide our paper search in our scoping review methodology. In the next section, we will explicitly outline the steps of the scoping review conducted in this study.

Method – Scoping Review

We employed a scoping literature review approach, which is an ideal method for determining “the scope or coverage of a body of literature on a given topic and providing a clear indication of the volume of literature and studies available, as well as an overview (broad or detailed) of its focus” (Munn et al., 2018, p. 2). This was the most appropriate review method for our study to address our research question. The extent to which the study seeks to provide in-depth coverage of existing literature primarily depends on the study's objectives (Arksey & O’Malley, 2005). By conducting a scoping review, we were able to effectively address our research question. Specifically, this approach allowed us to examine the "extent, range, and nature" (Arksey & O’Malley, 2005, p. 21) of how EEDI and technology have been studied and identify the emerging themes within the field of IS. Furthermore, it enabled us to identify research gaps in the existing literature and inform future research directions on the topics of EEDI and technology in IS.

Stages	Outcomes
Stage 1：Identifying the research question	• Identified the primary research question: *What do we know about EEDI in the IS field?* • Delineated the specific facets of the research question that we aimed to explore.
Stage 2: Identifying relevant studies	• Conducted a search of articles in electronic databases, using keywords. • Established the timeframe for the inclusion of published articles.

Stage 3: Study selection	• Specified the article selection and exclusion criteria • Read all searched articles' abstract. • Selected articles based on defined selection and exclusion criteria. • Reviewed the full paper when the abstract was insufficiently clear.
Stage 4: Charting the data	• Collected relevant information from the selected articles, including author, year, journal and abstract. • Summarised each article in one or two sentences, highlighting its key message.
Stage 5: Colleting, summarising and reporting the results	• Performed a numerical analysis of the article search results. • Identified emerging themes related to EEDI and technology. • Examined how these two topics have been studied within the IS field. • Wrote the findings.

Table 1. Scoping review stages

Our review process followed Arksey and O'Malley's (2005) scoping review methodological framework, which comprises five stages (Table 1). In the first stage, we identified our research question: *What do we know about EEDI in the IS field?* At this stage, we also carefully delineated the 'facets' (CRD, 2001) of the research question that we were particularly interested in, which included the themes of the topics studied, the relationship between EEDI and technology, and the identification of research gaps (i.e., future research directions). Defining these facets allowed us to focus our review on specific areas and scope the articles appropriately, ensuring that we found a sufficient number of relevant articles without missing key studies or becoming overwhelmed by an unmanageable volume of literature.

In the second stage, we identified relevant articles that aligned with the purpose of our study in the top 11 IS journals (listed on AIS website). At this stage, we initially searched electronic databases to access relevant studies, utilizing the Business Source Ultimate database from the University of Warwick e-library. We also employed "hand-searching of key journals" (Arksey & O'Malley, 2005) to retrieve articles from journals that may have been omitted from the university database, such as Information & Organisations, Journal of Strategic Information Systems, and Journal of Information Technology. We then used keywords, including Diversity, Equality, Equity, and Inclusion, to search for relevant articles. We ensured that these keywords appeared in the abstracts of the top 11 IS journals. Our initial search encompassed all references that appeared in the search results, spanning articles from the 1960s to 2024, this included all the available and published articles in the electronic databases.

In the third stage, we began selecting articles that were relevant to the study's purpose, linking them to the existing literature that we used to define our key concepts. Table 2 summarizes the selection criteria for relevant articles based on the existing EEDI literature. Two reviewers participated in this inclusion and exclusion process, ensuring consensus on the selected articles. If the relevance of a study was unclear during the selection and review process, the full articles were examined. As noted by Arksey and O’Malley (2005, p. 26), "abstracts cannot be assumed to be representative of the full article that follows, or to capture the full scope of an article" (Badger et al., 2000). Out of our initial search of 409 articles, 97 articles were selected for the review. The detailed number of selected articles per journal will be presented in the findings section.

Concepts	Inclusion Criteria	Exclusion Criteria
Diversity	Article focuses on individuals’ observable and unobservable diversity characteristics associated with technologies (i.e., in IT workplace and how IT foster diversity).	Diversity of things other than people, such as diversity in research topics (e.g., Baskerville and Wood-Harper, 1998, and diversity in products (e.g., Kui, 2018).
Inclusion	The act and practise of involve and accommodating individuals in an organisational or social setting.	Inclusion other than the given definition, such as focusing on adding of different digital features to a product (e.g., Gleasure et al., 2017).
Equity	Individuals are being treated fairly by giving same resources or opportunities associated with technologies (i.e., fair opportunities by IT or in IT workplace and industry).	Equity other than the given definition, such as financial equity (e.g., Kuang et al., 2019.)
Equality	Individuals given the exact resources based on their circumstances to reach a fair and equal outcome, associated with technologies (i.e., enabled by IT or in IT workplace and industry).	Equality other than the given definition.

Table 2. Selection criteria of articles.

In the fourth stage, we employed the "charting the data" method (Arksey & O'Malley, 2005). Charting involves synthesizing and interpreting qualitative data by organizing, sifting, and categorizing the material based on key issues and themes (Ritchie & Spencer, 1994). In this study, our charting approach focused on identifying emerging themes and topics related to EEDI and technologies in the selected articles. We concentrated on the "narrative" of the articles, complemented by the "descriptive-analytic" method within traditional narrative reviews (Pawson, 2002). We first collected standard information for each article, including title, abstract, authors, year of publication, and journal. We then summarized each article in one or two sentences, highlighting the study context, methodology, and key findings (Appendix A).

In the final stage, we collated, summarized, and reported our findings. Based on the charted information from the previous stage, we were able to present our narrative findings in two ways. First, we provided numerical analysis regarding the extent and distribution of the selected articles, this shows the number of relevant studies per journal under each key topic. Second, we identified nature of selected articles, this include the emerging themes and topics of EEDI and technologies and how they have been studied in the IS field. Our charted information and descriptive-analytic methods further enabled us to pinpoint research gaps in the existing literature and offer guidance for future research directions on EEDI and technologies in the IS field.

Findings

In this section, we present the findings from our scoping review. First, we provide a numerical analysis of the extent and distribution of the selected articles. Our initial search yielded a total of 409 articles (excluding duplicates) from the top 11 IS journals, all of which contained at least one keyword of EEDI in their abstracts. Based on our selection criteria, we included 97 articles that were deemed relevant and aligned with the research objectives of this scoping review, aimed at addressing our research question. Specifically, we identified 47 articles related to diversity, 33 related to inclusion, 7 focused on equality, and 10 on equity within the top 11 journals. It is important to note that some articles contained more than one of the keywords in their abstracts (e.g., Marabelli et al., 2023); however, we deliberately categorized each article according to one keyword to avoid confusion and duplication. Table 3 provides a detailed summary of the numerical distribution of the selected articles by journals and keywords.

Journals	Papers Identified with Keyword Search (Excluding Duplicates)	Papers Included in Review Pool	Diversity Papers Included	Inclusion Papers Included	Equality Papers Included	Equity Papers Included
DSS	65	6	3	0	1	2
EJIS	34	10	4	5	1	0
I&M	48	4	3	0	0	1
I&O	13	4	2	2	0	0
ISJ	40	22	8	13	0	0
ISR	45	5	4	0	1	0
JAIS	33	12	2	8	1	1
JIT	5	1	1	0	0	0
JMIS	30	9	5	0	2	2
JSIS	26	10	8	2	0	0
MISQ	72	15	7	3	1	4
Total	409	97	47	33	7	10

Table 3. Descriptive Summary of Selected Articles.

Using the descriptive-analytic method within traditional narrative reviews (Pawson, 2002), we identified emerging themes in the topics of EEDI and technologies being studied in the IS field. Specifically, we identified 6 emerging themes, along with an additional 18 sub-themes. Table 4a-Table 4f provide a detailed specification of these themes and sub-themes. Below, we present each emerging theme and its corresponding sub-themes.

Theme 1: EDI Management Studies

The first emerging theme identified in the selected articles is EEDI management. Relevant studies have explored how EEDI as a whole is managed across various contexts, including recruitment and employment, decision-making, strategy and innovation, and virtual spaces. These contexts have been further categorized as sub-themes (Table 4a). For example, Pessach et al. (2020) propose a framework for HR recruiters to use as a decision support tool to improve diversity and recruitment success rates. Similarly, Robert et al. (2018) emphasise the importance of tailoring communication technologies to meet the specific needs of diverse groups to optimize decision-making processes in a business decision context. Paul et al. (2004) argue that managing cultural diversity and conflict styles is crucial in virtual teams, as these factors significantly influence team performance.

Furthermore, several articles focus on the management of diversity characteristics solely, particularly age, race, and knowledge in IT professions or through the use of IT. For instance, Rhue and Clark (2022) investigate the role of race in crowdfunding campaigns on Kickstarter, finding that projects with racial identifiers, particularly African American or Asian cues, often have lower success rates compared to racially anonymous projects. Lastly, we identified a sub-theme related to diversity and IT security, with one study exploring the negative association between diversity and the likelihood of experiencing data breaches (Wang and Ngai, 2022).

Sub-themes & Definitions	Articles	Sample article summary
Recruitment and employment: How is EDI managed in the context of recruitment and employment in IT professions?	Pessach et al., 2020; Jia et al., 2022.	Pessach et al., 2020. This study proposed a framework as a decision support tool for HR recruiters to improve the diversity and recruitment success rate.
Decision making, strategy and innovation: How is EDI managed in the context of decision-making, strategy, and innovation?	Robert et al., 2018; Miranda et al., 2022; Carlo et al., 2012; Chau, 2002; Zhang et al., 2007; Fernandez and Olmedo, 2005.	Carlo et al., 2012. This study emphasizes the importance of knowledge diversity within software firms, showing that a diverse knowledge base is crucial for fostering various types of radical innovations.

Virtual places: How is EDI managed in the context of virtual spaces and communities?	Soltani Delgosha et al., 2024; Pinjani and Palvia, 2013; Ye and Jensen, 2022; Kankanhalli et al., 2006; Paul et al., 2004; Yin et al., 2023.	Ye and Jensen, 2022. This study investigates the impact of introducing an online community on contestant performance in crowdsourcing platforms, finding that having online community significantly enhances performance, particularly benefiting contestants with less and more diverse prior experience.
Diversity management: How are different diversity characteristics, particularly age, race, and knowledge, managed in the IT workplace or through IT?	Soh et al., 2011; Bergvall-Kåreborn and Howcroft, 2014; Dissanayake et al., 2021; Lee and Xia, 2010; Arazy et al.,2011; Daniel et al., 2018; He and King, 2007; Rhue and Clark, 2022; Igbaria and Wormley, 1992; Tams et al., 2014	Daniel et al., 2018. This study examines diversity of participant (i.e., language, role, and contribution) and project differences (development environment and connectedness), and examines how they interact to influence open-source project success.
IT security: How is EDI related to IT security?	Wang and Ngai, 2022.	Wang and Ngai, 2022. This study examines the relationship between firm diversity and the likelihood of experiencing data breaches, finding that the negative association between firm diversity and data breach risk, with certain conditions enhancing this protective effect.

Table 4a. Theme 1 of EDI Management and Studies.

Theme 2: Gender Studies

Gender studies represent a prominent theme within the IS field. We identified a group of studies focusing on sub-themes such as gender balance, women-focused topics, and pay gaps (Table 4b). For example, Gorbacheva et al. (2019) highlight significant research gaps in understanding and addressing gender imbalances in the IT profession, calling for future research to explore the consequences, explanations, and effective interventions for improving gender diversity in this field. Other studies on gender balance assess computing interventions aimed at improving gender parity (Craig, 2016), as well as compare men and women in terms of job satisfaction, workplace support, and experiences of sexual harassment (Gupta et al., 2019).

Studies focusing on women-related topics examine the career glass ceiling (Kirton and Robertson, 2018; Igbaria and Baroudi, 1995) and career support (Panteli, 2012; Armstrong et al., 2007). For instance, Kirton and Robertson (2018) assess how organizational culture, processes, and practices in a UK-based IT company contribute to the persistent under-representation of women in senior IT roles, applying Acker's 'inequality regimes' to highlight mechanisms that maintain gender inequality. Armstrong et al. (2007) explore how work-family conflict impacts women's advancement and turnover in IT, identifying key factors such as managing family responsibilities, work stress, work schedule flexibility, and job quality that influence women's career development in IT professions. Additionally, Levina and Xin (2007) analyse how national economic factors influence IT workers' compensation in the U.S., revealing that institutional differences, such as firm size, directly affect wages, and noting that female IT workers and those without college degrees face greater wage disparities during periods of labour market contraction.

Sub-themes & Definitions	Articles	Sample article summary
Pay gaps: How is gender related to pay gaps in IT professions?	Levina and Xin, 2007	Levina and Xin, 2007. This study analyses how national economic factors influence IT workers' compensation in the U.S., revealing that institutional differences, such as firm size, directly affect wages and they also found that female IT workers and those without college degrees face greater wage disparities during times of labour market contraction.

Gender balances: How fair access to the same resources for different genders is studied in IT professions or through IT?	Gorbacheva et al., 2019; Craig, 2016; Adam, 2001; Trauth, 2013; Gupta et al., 2019; Igbaria and Baroudi, 1995; Gorbacheva et al., 2019	Gorbacheva et al., 2019. This study highlights the significant research gaps in understanding and addressing the gender imbalance in the IT profession, calling for future research to explore the consequences, explanations, and effective interventions for improving gender diversity in this field.
Women-focused topics: How women-focused topics, such as the glass ceiling and access to IT resources, are studied in IT professions or related to IT?	Trauth and Connolly, 2021; Armstrong et al., 2007; Panteli, 2012; Annabi and Lebovitz, 2018; Quesenberry and Trauth, 2012; Kirton and Robertson, 2018; McGee, 2018	Trauth and Connolly, 2021. This study investigates how societal, organizational, and individual factors affecting gender equity in the IT field in Ireland have evolved over time, revealing seven themes that characterize changes impacting women IT professionals across five decades against the backdrop of socio-economic fluctuations.

Table 4b. Theme 2 of Gender Studies

Theme 3: IS Design Studies

IS planning and development represent another emerging theme within the selected articles. When IS design intersects with EEDI, studies frequently focus on the planning or development of IS artifacts that address EEDI issues or provide inclusive designs for underrepresented groups (Table 4c). For instance, Marabelli and Newell (2023) examine the strategic implications and DEI opportunities and challenges within the metaverse, emphasizing the necessity of ethical considerations. They propose a research agenda to advance both theoretical and practical knowledge on how this immersive sociotechnical system may impact organizations and industries. Additionally, Córdoba, (2009) proposes a methodological framework for IS planning in organizations, rooted in critical systems thinking, which emphasizes stakeholder engagement, continuous identification of concerns, and critical reflection on issues of inclusion, exclusion, and marginalization in the IS planning processes.

Sub-themes & Definitions	Articles	Sample article summary
EDI considerations: How is IS designed with EDI considerations?	Pee et al., 2021; Braa et al., 2023; Córdoba, 2009; Chipidza and Leidner, 2019; Wass and Purao, 2023; Hossain et al., 2012; Marabelli and Newell, 2023; Kankanhalli et al., 2004; Katzy et al., 2016; Bailey and Michaels, 2019	Pee et al., 2021. This article proposes guidelines for integrating future-oriented perspectives into IS design research, emphasizing the importance of considering potential futures to ensure that technological solutions are not only innovative but also equitable and inclusive, addressing potential future challenges and mitigating undesirable outcome during covid-19.

Table 4c. Theme 3 of IS Designs.

Theme 4: Digital Inclusion Studies

Digital inclusion also emerges as a significant theme in the selected articles. Within this theme, studies often concentrate on sub-themes such as financial inclusion, workplace inclusion, social inclusion, inclusion versus exclusion, and the digital divide (Table 4d). For example, Senyo et al. (2022) investigate how interactions between new entrants and incumbents in FinTech ecosystems in Ghana shape financial inclusion through innovative, collaborative, protectionist, equitable, legitimizing, and sustaining practices. They provide a theoretical model and propositions for scaling financial inclusion in developing countries. Heath and Babu (2017) in their study examine the relationship between IT use and workplace inclusion for people with disabilities. Similarly, Newman et al. (2017) explore the challenges of digital inclusion faced by young people with disabilities in accessing the Internet, using Bourdieu's critical theory to analyze how unequal resource distribution affects their online participation and access needs from a social inclusion perspective. Regarding inclusion versus exclusion, studies have investigated how social media can enhance social inclusion for first-generation college students at a Hispanic-serving institution, identifying various affordances and outcomes that improve social engagement and mitigate exclusion. Additionally, Cordoba and Midgley (2008) critically examine the need to expand IS planning processes beyond organizational concerns to address broader societal needs, discussing topics such as the "digital divide" to promote inclusion in IS design within a university setting.

Sub-themes & Definitions	Articles	Sample article summary
Financial inclusion: How do individuals and businesses have fair access to financial products and services?	McBride and Liyala, 2023; Mishra and Srivastava, 2022; Senyo et al., 2022; Tan et al., 2020; Lagna and Ravishankar, 2022; Siqueira et al., 2023	Mishra and Srivastava, 2022. This study explores how mobile payment adoption among unorganised retailers in India contributes to digital and financial inclusion. It underscores the pivotal role of these retailers in connecting marginalized citizens to state services and economic opportunities.
Workplace inclusion: How do employees feel they are being valued and treated fairly in the IT workplace or through IT?	Heath and Babu, 2017; Tarafdar et al., 2023; Mayer et al., 2024	Heath and Babu, 2017. This study examines the relationship between using IT and workplace inclusion for people with disabilities.
Social inclusion: How do individuals feel they are being valued and treated fairly in society through IT?	Qureshi et al., 2022; Zhou and Kordzadeh, 2023; Newman et al., 2017; Andrade and Doolin, 2016; Leong et al., 2022; Diaz Andrade et al., 2021; Pandey and Zheng, 2023; Qureshi et al., 2021; Ahuja et al., 2023	Newman et al., 2017. This study explores the digital inclusion challenges faced by young people with disabilities in accessing the Internet, using Bourdieu's critical theory to analyse how unequal distribution of resources impacts their online participation and access needs.

Inclusion vs exclusion: How are individuals included and excluded in the IT workplace or in society through IT?	Gonzalez and Deng, 2023; Iivari et al., 2018; Curto-Millet and Cañibano, 2023; Petter and Giddens, 2023	Pandey and Zheng, 2023.This study explores how social networking technologies like social media can enhance social inclusion for first-generation college students at a Hispanic-serving institution, identifying various affordances and outcomes that improve their social engagement and mitigate exclusion.
Digital divide: How are individuals separated by their access to IT?	Diaz Andrade, A., & Techatassanasoontorn, A. A, 2021; Fox and Connolly, 2018; Cordoba and Midgley, 2008; Chou et al, 2013; Narayanaswamy and Henry, 2013	Cordoba and Midgley, 2008. This study critically considers and expand the boundaries of IS planning processes beyond organizational concerns to encompass broader societal needs, consider topics such as "Digital divides', to promote inclusivity in IS designs in a university setting.

Table 4d. Theme 4: Digital Inclusion Studies

Theme 5: Equality and Equity Studies

While equality and equity as outcomes of diversity and inclusion management are mentioned in several studies (e.g., Kirton and Robertson, 2018; Diaz Andrade & Techatassanasoontorn, 2021), a few studies have thoroughly explored sub-themes centred on resource allocation, fair participation and treatment to address equality and equity issues (Table 4e). For example, Tong et al., (2022) explore the impact of health information technology on rural-urban healthcare access inequality through qualitative analysis, grounded in social transformation and affordance actualization theories, highlighting its potential role in addressing societal challenges and contributing to equity, diversity, and inclusion in healthcare. In terms of fair participation and treatment, Chou et al. (2013) highlight the importance of perceived equity in IS project teams, emphasizing how fairness in reward distribution and treatment can foster job commitment and supportive behaviours. This, in turn, promotes a more inclusive and equitable workplace environment for IS personnel.

Sub-themes & Definitions	Articles	Sample article summary

Resources allocation: How are individuals separated by accessing different resources enabled by IT?	Joshi, 1989; Goh et al., 2016; Tong et al., 2022;	Joshi, 1989. This study develops an instrument to measure perceptions of fairness in the allocation of information systems resources, drawing from equity and social justice literature, with implications for enhancing equity, diversity, and inclusion within centralized MIS functions.
Fair participation and treatment: How do employees have fair participation and treatment the IT workplace or through IT?	Fjermestad, 2004; McLeod and Liker, 1992; Tan et al., 2003; Mejias et al., 1996; Ho and Raman, 1991; Ellis et al., 1989	Chou et al, 2013. This study suggests the importance of perceived equity in IS project teams, highlighting how fairness in reward distribution and treatment can foster job commitment and supportive behaviors, thereby promoting a more inclusive and equitable workplace environment for IS personnel.

Table 4e. Fifth Theme of Equality and Equity Studies

Theme 6: IS Community Studies

The final emerging theme identified in the selected articles is the IS community. Within this theme, studies often focus on the significant role of IS in addressing EEDI-related issues and setting a future research agenda, as well as on existing EEDI challenges within the current IS community (Table 4f). For instance, several editorial papers have urged the IS community to consider the future direction of EEDI and how technology can advance these concepts (Marabelli and Chan, 2024; Windeler et al., 2023; Burton-Jones and Sarker, 2021). Additionally, studies by Payton et al. (2005; 2022) have addressed the severe underrepresentation of minority faculty in the global IS field, proposing mentoring as a key strategy to enhance diversity and inclusion in the IS community.

Sub-themes & Definitions	Articles	Sample article summary

Research agenda: How do IS play an important role in addressing EDI-related issues, and what are the future research suggestions?	Winter and Saunders,2019; Marabelli and Chan, 2024; Windeler et al., 2023; Burton-Jones and Sarker, 2021	Windeler et al., 2023. This study explores how access to information systems impacts societal inequalities and divisions, emphasizing the ongoing challenges and complexities in achieving equity and inclusion through technology.
EDI issues: How do EDI-related issues exist within the current IS community?	Marabelli et al., 2023; Davison, 2021; Fedorowicz et al., 2023; Payton et al., 2022; Payton et al., 2005	Payton et al., 2022. This study amplifies the voices of Black IS professors, highlighting their experiences of underrepresentation, tokenization, isolation, marginalization, and exclusion from positions of power within the IS field.

Table 4f. Sixth Theme of IS Community Studies.

Exploring the Relationships between EEDI and Technology

During the analysis, we identified three key relationships between the concepts of EEDI and technology. These relationships provide deeper insights into how EEDI and technology influence one another. Table 5 summarizes our findings regarding these relationships. In the following section, we detail each relationship and specifically explain how the interplay between these two concepts functions.

Type 1: IT has a causal role on EEDI issues

The first relationship is that IT has a causal impact on EEDI issues. In this dimension, IT can generate positive, negative, and even ambivalent affordances regarding EEDI concerns. For example, Qureshi et al., (2022) explore how videos can be used to curate and disseminate indigenous knowledge among farmers, context-specific knowledge, creating a knowledge common that leverages local resources, and overcoming socio-cultural and technological barriers to ensure equitable knowledge dissemination and utilization compared to top-down expert-driven approaches. In contrast, Diaz and Techatassanasoontorn (2021) examined the ethical issue of "digital enforcement," where an overemphasis on digital inclusion as a solution to the digital divide may have negative consequences, such as limiting the choices of individuals who prefer to live offline, thereby creating a new form of inequality. Additionally, technology can have an ambivalent effect on EEDI issues. For instance, Pandey and Zheng (2023) explored the dual nature of digital inclusion, demonstrating how technology can both empower marginalized individuals and reinforce existing power structures, offering an ambivalent understanding of the role of digital tools in inclusion and equity.

Type 2: EEDI Has a Casual Role on Technology

The second relationship dimension we identified is that EEDI also has a causal impact on technology, especially on its designs and planning. For example, the study by Fox and Connolly (2018) highlights how mistrust, risk perceptions, and privacy concerns related to the age-based digital divide can lead to recommendations for more inclusive design practices, tailored to different ages. Moreover, we discovered an intertwined relationship between EEDI issues and technology, where each can influence the other in a feedback loop. For example, Wass and Purao (2023) examined an IS design process involving individuals with intellectual and developmental disabilities. Their findings indicate that the design process not only led to a more inclusive IS solution to meet the special needs from the disadvantage group but also achieved social inclusion for disadvantaged groups by involving these individuals in the design process.

Type 3: IT Is the Context of EEDI Management

The third relationship dimension is that technology serves as a context in EEDI management (e.g., as a platform, in the IT industry workforce, or in IS research). Few articles also examined their studies in the IT industry and professions, for example, Gorbacheva et al., (2019) highlight the significant research gaps in understanding and addressing the gender imbalance in the IT profession, calling for future research to explore the consequences, explanations, and effective interventions for improving gender diversity in this field. Additionally, we found that IT is a context as a digital platform such as social media and crowdsourcing planforms. For example, Ye and Jensen (2022) find that having online community significantly enhances contestant performance in crowdsourcing, particularly benefiting contestants with less and more diverse prior experience. Lastly, few studies (e.g., Winter and Saunders, 2019; Payton et al., 2005) have addressed the EEDI issues in the IS academic community, as we previously discussed in theme 6.

Relationship Dimensions	**Effects**	**Definition**	**Sample Studies**	**Sample Article Summary**
IT has a causal role on EEDI issues	Positive influence	Technology helps to foster EEDI and make positive impact in either organisational or social context.	Pessach et al., 2020; Fjermestad, 2004; Zhang et al., 2007; Katzy et al., 2016; Andrade and Doolin, 2016; Tong et al., 2022	Pessach et al., 2020. This study proposed a framework as a decision support tool for HR recruiters to improve the diversity and recruitment success rate.
	Negative influence	Technology constrains EEDI and has negative impact in either organisational or social context.	Diaz and Techatassanasoontorn, 202; Siqueira et al., 2023	Diaz Andrade, A., & Techatassanasoontorn, A. A, 2021. This article argues that the overemphasis on digital inclusion as a solution to the digital divide forces individuals into increased Internet reliance, creating a new form of inequality called "digital enforcement," which limits personal choice and raises ethical concerns.
	Ambivalent (both)	Technology has contradicting findings on EEDI management.	Pandey and Zheng, 2023; Iivari et al., 2018; Curto-Millet and Cañibano, 2023	Pandey and Zheng, 2023. The study explores the dual nature of digital inclusion, showing how technology can both empower marginalized individuals and reinforce existing power structures, therefore, offering a nuanced understanding of how digital tools impact equity and inclusion in marginalized communities.

EEDI has a casual role on technology	Directly impact	EEDI influence directly how technology is designed	Pee et al., 2021; Braa et al., 2023; Hossain et al., 2012; Fox and Connolly, 2018	Pee et al., 2021. This article proposes guidelines for integrating future-oriented perspectives into IS design research, emphasizing the importance of considering potential futures to ensure that technological solutions are not only innovative but also equitable and inclusive, addressing potential future challenges and mitigating undesirable outcome during covid-19.
	Mutually impact	EEDI and technology design are mutually influenced, that each can influence the other in a feedback loop.	Córdoba, 2009; Wass and Purao, 2023; Heath and Babu, 2017	Wass and Purao, 2023. This study develops five principles for socially inclusive design-oriented research with marginalized groups, drawing from an empirical investigation of IT-based solutions for people with intellectual and developmental disabilities. Their findings indicate that the design process not only led to a more inclusive IS solution to meet the special needs from the disadvantage group but also achieved social inclusion for disadvantaged groups by involving these individuals in the design process.

IT is the context in EEDI management (e.g., a platform, IT industry workforce, IS research)	Research case setting	Technology serves as study/research setting when considering EEDI.	Tan et al., 2003; Dissanayake et al., 2021; Rhue and Clark, 2022; Daniel et al., 2013; Payton et al., 2022	Dissanayake et al., 2021. The study examines how board structure and changes in board composition affect organizational performance in the IT industry, finding that board size, gender diversity, and age have curvilinear effects, while board independence positively impacts performance and frequent changes negatively impact performance, especially in firms with fewer independent directors.

Table 5. Three Types of Casual Relationship Between EEDI and Technology Industrial Sectors Examined in Existing Articles.

Discussion

This scoping literature review aims to explore the current understanding of the concepts of EEDI and their integrations with technology in the field of IS. The study examined 97 relevant articles out of 409 published across 11 IS journals, covering a period from the 1960s to 2024. By employing a "descriptive-analytic" method within the framework of traditional narrative reviews (Pawson, 2002), we identified 6 emerging themes and an additional 18 sub-themes. These themes and sub-themes highlight the predominant topics of EEDI and technologies being studied in the IS field. For example, we include EEDI management in IT workplaces, gender differences among IT professionals, issues that women facing in their IT career path, inclusion versus exclusion, digital inclusion for disadvantaged groups, and considerations of equality and equity in participation within the IS community. Additionally, our review revealed several socially and organizationally significant themes that have received relatively less attention in the literature, such as diversity in pay gaps, diversity and IT security, and workplace inclusion.

This scoping review also uncovered three interrelated relationships between EEDI and technologies. First, we found that technologies can impact EEDI positively, negatively, or ambivalently. Second, EEDI considerations can influence technology, particularly in design processes. Incorporating EEDI considerations into IS design not only directly affects how IS artifacts are created to address EEDI challenges, resulting in more inclusive and equitable outcomes for end-users (directly impact); but also fosters an EEDI-friendly environment by involving diverse individuals in the design process through a mutual impact. Finally, we identified that technology often serves as a context for studying EEDI management in various studies, such as digital platforms (e.g., Ye and Jensen, 2022), specific industries (e.g., Panteli, 2012), and research communities (e.g., Marabelli et al., 2023).

This study contributes to the existing body of knowledge by identifying 6 emerging themes and 18 additional sub-themes related to EEDI and technology, thereby shedding light on both what is known and unknown in this area. Furthermore, it deepens our understanding of how EEDI and technology can be studied by examining their intertwined relationships. The study also highlights several themes that are currently underexplored, based on the selected articles, and suggests that future research could potentially address these areas, these research areas include as following:

- Gender pay gaps in IT professions: Future studies could potentially examine and explore how pay affects different genders in IT professions, whether this influences gender inequality in the IT workplace, and what factors contribute to such inequality.
- Workplace inclusion and diversity management enabled by IT: Future studies can explore how IT is being used or designed to manage inclusion and diversity in the workplace. What is the role of IT in fostering inclusion and diversity, and how does it achieve this?
- How does IT help marginalized groups to raise their voices both within organizations and society: Who are considered in a marginalized group across different contexts? How can IT address their needs across these contexts?

This scoping review has some limitations. First, the choice of search keywords was limited. We only used the keywords related to EEDI, which may have excluded other relevant terms with similar meanings, potentially resulting in the omission of pertinent articles. Additionally, our inclusion of the articles only from the top 11 IS journals, this will limit some insightful and valuable articles that published in other IS jouranls outside the 11 basket. Future research could extend the search keywords to include terms such as implicit bias, underrepresentation, tokenism, organizational silence, exclusion, elitism, ageism, sexism, implicit stereotypes, diversity metrics, and LGBTQ, and extend the journal articles selection broadly beyond the 11 IS journal basket. Second, due to the nature of a scoping review, our analysis primarily focused on the abstracts of the articles. Full articles were only examined when the abstract was unclear about the study's themes, context, or key findings. As a result, some articles with more nuanced findings regarding the relationship between technology and EEDI may not have been included. Nevertheless, our findings have deepened our understanding of what is known and unknown about EEDI and technology, as well as their interrelated relationships within the IS field.

Conclusion

This study seeks to answer the question of what is known and what remains unknown about EEDI and technology. Through a scoping review, we identified 6 emerging themes and 18 additional sub-themes, providing insights into the trending topics in EEDI and technology within the field of IS. By identifying three key relationships between EEDI and technology, this study explores how these concepts can be interdependently examined. Based on this scoping review, we suggest directions for future research on these topics.

References

Adam, A. (2001). Computer ethics in a different voice. *Information and organization*, *11*(4), 235-261.

Ahuja, S., Chan, Y. E., & Krishnamurthy, R. (2023). Responsible innovation with digital platforms: Cases in India and Canada. *Information Systems Journal*, *33*(1), 76-129.

Anderson, E. (2010). Defending the Capabilities Approach to Justice. *Measuring Justice: Primary Goods and Capabilities*, 81-100.

Andrade, A. D., & Doolin, B. (2016). Information and communication technology and the social inclusion of refugees. *Mis Quarterly*, *40*(2), 405-416.

Annabi, H., & Lebovitz, S. (2018). Improving the retention of women in the IT workforce: An investigation of gender diversity interventions in the USA. *Information Systems Journal*, *28*(6), 1049-1081.

Arazy, O., Nov, O., Patterson, R., & Yeo, L. (2011). Information quality in Wikipedia: The effects of group composition and task conflict. *Journal of management information systems*, *27*(4), 71-98.

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International journal of social research methodology*, *8*(1), 19-32.

Armstrong, D. J., Riemenschneider, C. K., Allen, M. W., & Reid, M. F. (2007). Advancement, voluntary turnover and women in IT: A cognitive study of work–family conflict. *Information & Management*, *44*(2), 142-153.

Badger, R., & White, G. (2000). A process genre approach to teaching writing. *ELT journal*, *54*(2), 153-160.

Bailey, M.D. and Michaels, D. (2019). An optimization-based DSS for student-to-teacher assignment: Classroom heterogeneity and teacher performance measures. *Decision Support Systems,* 119, pp.60-71.

Baskerville, R., & Wood-Harper, A. T. (1998). Diversity in information systems action research methods. *European Journal of information systems*, *7*(2), 90-107.

Bassett-Jones, N. (2005). The paradox of diversity management, creativity and innovation. *Creativity and innovation management*, *14*(2), 169-175.

Bérubé, J., Doris, J., & Pouliot, A. (2024). The role of cultural organisations in matters of equity, diversity, and inclusion. *Cultural Trends*, 1-16.

Bergvall-Kåreborn, B., & Howcroft, D. (2014). Persistent problems and practices in information systems development: a study of mobile applications development and distribution. *Information Systems Journal*, *24*(5), 425-444.

Braa, J., Sahay, S., & Monteiro, E. (2023). Design Theory for Societal Digital Transformation: The Case of Digital Global Health. *arXiv preprint arXiv:2311.09173*. p.1-57

Burton-Jones, A., & Sarker, S. (2021). Creating Our Editorial Board Position Statement on Diversity, Equity, and Inclusion (DEI). *MIS Quarterly*, *45*(4). iii-iv.

Campbell, A. (2021). What is Black Lives Matter and what are the aims? BBC News. https://www.bbc.co.uk/news/explainers-53337780

Carlo, J. L., Lyytinen, K., & Rose, G. M. (2012). A knowledge-based model of radical innovation in small software firms. *MIS quarterly*, 865-895.

Chau, P.Y., 2002. For the Special Issue on 'Personal Aspects of E-Business'. *European Journal of Information Systems*, *11*(3), pp.179-180.

Chipidza, W., & Leidner, D. (2019). A review of the ICT-enabled development literature: Towards a power parity theory of ICT4D. *The Journal of Strategic Information Systems*, *28*(2), 145-174.

Chou, T. Y., Seng-cho, T. C., Jiang, J. J., & Klein, G. (2013). The organizational citizenship behavior of IS personnel: Does organizational justice matter?. *Information & Management*, *50*(2-3), 105-111.

Crosby, F. J., Iyer, A., & Sincharoen, S. (2006). Understanding affirmative action. *Annal Review of Psychology*, *57*(1), 585-611.

Cordoba, J.R. and Midgley, G., 2008. Beyond organisational agendas: using boundary critique to facilitate the inclusion of societal concerns in information systems planning. *European Journal of Information Systems*, *17*(2), pp.125-142.

Córdoba, J. R. (2009). Critical reflection in planning information systems: A contribution from critical systems thinking. *Information Systems Journal*, *19*(2), 123-147.

Curto-Millet, D., & Cañibano, A. (2023). The design of social inclusion interventions: A paradox approach. *Journal of the Association for Information Systems*, *24*(5), 1271-1291.

Daniel, S., Midha, V., Bhattacherhjee, A., & Singh, S. P. (2018). Sourcing knowledge in open source software projects: The impacts of internal and external social capital on project success. *The Journal of Strategic Information Systems*, *27*(3), 237-256.

Daniel, S., Agarwal, R., & Stewart, K. J. (2013). The effects of diversity in global, distributed collectives: A study of open-source project success. *Information Systems Research*, *24*(2), 312-333.

Davison, R.M., 2021. Diversity and inclusion at the ISJ. *Information Systems Journal*, *31*(3). 347-355

Diaz Andrade, A., Techatassanasoontorn, A. A., Singh, H., & Staniland, N. (2021). Indigenous cultural re-presentation and re-affirmation: The case of Māori IT professionals. *Information Systems Journal*, *31*(6), 803-837.

Dissanayake, I., Jeyaraj, A., & Nerur, S. P. (2021). The impact of structure and flux of corporate boards on organizational performance: A perspective from the information technology industry. *The Journal of Strategic Information Systems*, *30*(2), 101667. 1-16.

Ellis, C. A., Rein, G. L., & Jarvenpaa, S. L. (1989). Nick experimentation: Selected results concerning effectiveness of meeting support technology. *Journal of Management Information Systems*, *6*(3), 7-24.

Fedorowicz, J., Payton, F. C., Chan, Y. E., Kim, Y. J., & Te'eni, D. (2023). DEI in the IS discipline: What can we do better?. *The Journal of Strategic Information Systems*, *32*(2), 101775.

Fernandez, E. and Olmedo, R., 2005. An agent model based on ideas of concordance and discordance for group ranking problems. *Decision Support Systems*, *39*(3), pp.429-443.

Ferdman, B. M., & Deane, B. (2014). *Diversity at work: The practice of inclusion*. John Wiley & Sons.

Ferdman, B. M., Barrera, V., Allen, A., & Vuong, V. (2009, August). Inclusive behavior and the experience of inclusion. In *BG Chung (Chair), Inclusion in organizations: Measures, HR practices, and climate. Symposium presented at the 69th Annual Meeting of the Academy of Management, Chicago*.

Fjermestad, J., 2004. An analysis of communication mode in group support systems research. *Decision Support Systems*, 37(2), 239-263.

Fleurbaey, M., and François, M. (2019): Well-Being Measurement with Non-Classical Goods. *Economic theory* 68. (3). 765–786.

Fourie, C. (2012). What is social equality? An analysis of status equality as a strongly egalitarian ideal. *Res Publica*, *18*(2), 107-126.

Fox, G., & Connolly, R. (2018). Mobile health technology adoption across generations: Narrowing the digital divide. *Information Systems Journal, 28*(6), 995-1019.

Craig, A. (2016). Theorising about gender and computing interventions through an evaluation framework. *Information Systems Journal*, *26*(6), 585-611.

Gupta, R. (2013). Workforce diversity and organisational performance. *International Journal of Business and Management invention*, 2(6), pp, 36-41.

Gupta, B., Loiacono, E. T., Dutchak, I. G., & Thatcher, J. B. (2019). A field-based view on gender in the information systems discipline: Preliminary evidence and an agenda for change. *Journal of the Association for Information Systems*, *20*(12), 2.

Goh, J. M., Gao, G., & Agarwal, R. (2016). The creation of social value. *MIS quarterly*, *40*(1), 247-264.

Gorbacheva, E., Beekhuyzen, J., vom Brocke, J. and Becker, J., 2019. Directions for research on gender imbalance in the IT profession. *European Journal of Information Systems*, *28*(1), pp.43-67.

Gonzalez, E., & Deng, X. N. (2023). Social Inclusion: The Use of Social Media and the Impact on First-Generation College Students. *Journal of the Association for Information Systems*, *24*(5), 1313-1333.

Harrison, D. A., & Klein, K. J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of management review*, *32*(4), 1199-1228.

He, J., Butler, B. S., & King, W. R. (2007). Team cognition: Development and evolution in software project teams. *Journal of Management Information Systems*, *24*(2), 261-292.

Heath, D., & Babu, R. (2017). Theorizing managerial perceptions, enabling IT, and the social inclusion of workers with disabilities. *Information and Organization*, *27*(4), 211-225.

Hellerstedt, K., Uman, T., & Wennberg, K. (2024). Fooled by diversity? When diversity initiatives exacerbate rather than mitigate bias and inequality. *Academy of Management Perspectives*, *38*(1), 23-42.

Ho, T. H., & Raman, K. S. (1991). The effect of GDSS and elected leadership on small group meetings. *Journal of Management Information Systems*, *8*(2), 109-133.

Holvino, E., Ferdman, B. M., & Merrill-Sands, D. (2004). Creating and sustaining diversity and inclusion in organizations: Strategies and approaches. In M. S. Stockdale & F. J. Crosby (Eds.), *The psychology and management of workplace diversity* (pp. 245–276). Blackwell Publishing.

Hossain, M.D., Moon, J., Yun, J.W. and Choe, Y.C., 2012. Impact of psychological traits on user performance in information systems delivering customer service: IS management perspective. *Decision Support Systems*, 54(1), pp.270-281.

Igbaria, M., & Wormley, W. M. (1992). Organizational experiences and career success of MIS professionals and managers: An examination of race differences. *MIS Quarterly*, 507-529

Igbaria, M., & Baroudi, J. J. (1995). The impact of job performance evaluations on career advancement prospects: An examination of gender differences in the IS workplace. *Mis Quarterly*, 107-123.

Iivari, N., Kinnula, M., Molin‐Juustila, T., & Kuure, L. (2018). Exclusions in social inclusion projects: Struggles in involving children in digital technology development. *Information Systems Journal*, *28*(6), 1020-1048.

Jia, R., Steelman, Z. R., & Jia, H. H. (2022). What makes one intrinsically interested in it? An explorative study of influence of autistic tendency and gender in the US and India. *MIS Quarterly*, *46*(3).

Jepson, R., Blasi, Z. D., Wright, K., &Riet, G.T. (2001) *Scoping review of the effectiveness of mental health services*, CRD Report 21. York: NHS Centre for Reviews and Dissemination, University of York.

Joshi, K. (1989). The measurement of fairness or equity perceptions of management information systems users. *MIS quarterly*, 343-358.

Kankanhalli, A., Tan, B. C., & Wei, K. K. (2006). Conflict and performance in global virtual teams. *Journal of management information systems*, *23*(3), 237-274.

Kankanhalli, A., Tan, B.C., Wei, K.K. and Holmes, M.C., 2004. Cross-cultural differences and information systems developer values. *Decision Support Systems*, 38(2), pp.183-195.

Katzy, B.R., Sung, G. and Crowston, K., 2016. Alignment in an inter-organisational network: the case of ARC transistance. *European Journal of Information Systems*, *25*(6), pp.553-568.

Kilduff, M., Angelmar, R., & Mehra, A. (2000). Top management-team diversity and firm performance: Examining the role of cognitions. *Organization science*, *11*(1), 21-34.

Kirton, G., & Robertson, M. (2018). Sustaining and advancing IT careers: Women's experiences in a UK-based IT company. *The Journal of Strategic Information Systems*, *27*(2), 157-169

Kuang, L., Huang, N., Hong, Y., & Yan, Z. (2019). Spillover effects of financial incentives on non-incentivized user engagement: Evidence from an online knowledge exchange platform. *Journal of Management Information Systems*, *36*(1), 289-320.

Lagna, A., & Ravishankar, M. N. (2022). Making the world a better place with fintech research. *Information Systems Journal*, *32*(1), 61-102.

Lee, G., & Xia, W. (2010). Toward agile: an integrated analysis of quantitative and qualitative field data on software development agility. *MIS quarterly*, *34*(1), 87-114.

Levina, N., & Xin, M. (2007). Research note—Comparing IT workers' compensation across country contexts: Demographic, human capital, and institutional factors. *Information Systems Research*, *18*(2), 193-210.

Leong, C., Tan, F. T. C., Tan, B., & Faisal, F. (2022). The emancipatory potential of digital entrepreneurship: A study of financial technology-driven inclusive growth. *Information & Management*, *59*(3), 103384.

Lu, T., & Zhang, Y. (2023). Profit vs equality? The case of financial risk assessment and a new perspective on alternative data. *MIS Quarterly*, *47*(4). 1517-1556.

Marabelli, M., & Newell, S. (2023). Responsibly strategizing with the metaverse: Business implications and DEI opportunities and challenges. *The Journal of Strategic Information Systems*, *32*(2), 101774.1-16.

Marabelli, M., Zaza, S., Masiero, S., Li, J. and Chudoba, K., (2023). Diversity, Equity and Inclusion in the AIS: Challenges and opportunities of remote conferences. *Information Systems Journal*, 33(6), 1370-1395.

Marabelli, M., & Chan, Y. E. (2024). The strategic value of DEI in the information systems discipline. *The Journal of Strategic Information Systems*, *33*(1), 101823, 1-8.

Mayer, A. S., Ihl, A., Grabl, S., Strunk, K., & Fiedler, M. (2024). A silver lining for the excluded: Exploring experiences that micro-task crowdsourcing affords workers with impaired work access. *Information Systems Journal*. 1-32.

McBride, N. and Liyala, S., 2023. Memoirs from Bukhalalire: a poetic inquiry into the lived experience of M-PESA mobile money usage in rural Kenya. *European Journal of Information Systems*, *32*(2), 173-194.

McGee, K. (2018). The influence of gender, and race/ethnicity on advancement in information technology (IT). *Information and Organization*, *28*(1), 1-36.

McLeod, P. L., & Liker, J. K. (1992). Electronic meeting systems: Evidence from a low structure environment. *Information Systems Research*, *3*(3), 195-223.

Miller, T., & del Carmen Triana, M. (2009). Demographic diversity in the boardroom: Mediators of the board diversity–firm performance relationship. *Journal of Management studies*, *46*(5), 755-786.

Miranda, S. M., Wang, D. D., & Tian, C. A. (2022). Discursive fields and the diversity -coherence paradox: An ecological perspective on the blackchin community discourse. *MIS Quarterly*, *46*(3).1421-1452.

Mishra, V., Walsh, I. and Srivastava, A., 2022. Merchants' adoption of mobile payment in emerging economies: the case of unorganised retailers in India. *European Journal of Information Systems*, *31*(1), pp.74-90.

Mejias, R. J., Shepherd, M. M., Vogel, D. R., & Lazaneo, L. (1996). Consensus and perceived satisfaction levels: A cross-cultural comparison of GSS and non-GSS outcomes within and between the United States and Mexico. *Journal of Management Information Systems*, *13*(3), 137-161.

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology*, *18 (1)*, 1-7.

Narayanaswamy, R., Grover, V., & Henry, R. M. (2013). The impact of influence tactics in information system development projects: A control-loss perspective. *Journal of Management Information Systems*, *30*(1), 191-226.

Newman, L., Browne-Yung, K., Raghavendra, P., Wood, D., & Grace, E. (2017). Applying a critical approach to investigate barriers to digital inclusion and online social networking among young people with disabilities. *Information Systems Journal*, *27*(5), 559-588.

Payton, F. C., Yarger, L., & Mbarika, V. (2022). Black Lives Matter: A perspective from three Black information systems scholars. *Information Systems Journal*, *32*(1), 222-232.

Payton, F. C., White, S. D., & Mbarika, V. W. (2005). A re-examination of racioethnic imbalance of IS doctorates: Changing the face of the IS classroom. *Journal of the Association for Information Systems*, *6*(1), 37-51.

Panteli, N. (2012). A community of practice view of intervention programmes: the case of women returning to IT. *Information Systems Journal*, *22*(5), 391-405.

Paul, S., Samarah, I. M., Seetharaman, P., & Mykytyn Jr, P. P. (2004). An empirical investigation of collaborative conflict management style in group support system-based global virtual teams. *Journal of management information systems*, *21*(3), 185-222.

Pawson, R. (2002). Evidence-based policy: in search of a method. *Evaluation*, *8*(2), 157-181.

Pessach, Dana et al. "Employees Recruitment: A Prescriptive Analytics Approach via Machine Learning and Mathematical Programming." *Decision Support Systems.* 134 (2020): 113290–113290.

Pessach, D., Singer, G., Avrahami, D., Ben-Gal, H. C., Shmueli, E., & Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, *134 (2020)*, 113290.

Pee, L.G., Pan, S.L., Wang, J. and Wu, J., 2021. Designing for the future in the age of pandemics: A future-ready design research (FRDR) process. *European Journal of Information Systems*, *30*(2), pp.157-175.

Pandey, P., & Zheng, Y. (2023). Technologies of Power in Digital Inclusion. *Journal of the Association of Information Systems*, *24*(5), 1334-1357.

Pinjani, P., & Palvia, P. (2013). Trust and knowledge sharing in diverse global virtual teams. *Information & management*, *50*(4), 144-153.

Petter, S., & Giddens, L. (2023). Is it Your Fault? Framing Social Media Inclusion and Exclusion Using Just World Theory. *Journal of the Association for Information Systems*, *24*(5), 1248-1270.

Rhue, L., & Clark, J. (2022). Who are you and what are you selling? Creator based and product based racial cues in crowdfunding. *MIS Quarterly*, *46*(4). 2229-2259

Quesenberry, J. L., & Trauth, E. M. (2012). The (dis) placement of women in the IT workforce: an investigation of individual career values and organisational interventions. *Information Systems Journal*, *22*(6), 457-473.

Qureshi, I., Bhatt, B., Parthiban, R., Sun, R., Shukla, D. M., Hota, P. K., & Xu, Z. (2022). Knowledge commoning: Scaffolding and technoficing to overcome challenges of knowledge curation. *Information and Organization*, *32*(2), 100410.

Qureshi, I., Pan, S. L., & Zheng, Y. (2021). Digital social innovation: An overview and research framework. *Information Systems Journal*, *31*(5). 647-671.

Roberson, Q. M. (2006). Disentangling the meanings of diversity and inclusion in organizations. *Group & organization management*, *31*(2), 212-236.

Robert Jr, L. P., Dennis, A. R., & Ahuja, M. K. (2018). Differences are different: Examining the effects of communication media on the impacts of racial and gender diversity in decision-making teams. Information Systems Research, 29(3), 525-545.

Senyo, P.K., Karanasios, S., Gozman, D. and Baba, M., 2022. FinTech ecosystem practices shaping financial inclusion: The case of mobile money in Ghana. *European Journal of Information Systems*, *31*(1),112-127.

Shachaf, P. (2008). Cultural diversity and information and communication technology impacts on global virtual teams: An exploratory study. *Information & Management*, *45*(2), 131-142.

Siqueira, É. S., Diniz, E. H., & Pozzebon, M. (2023). Surveilled inclusion and the pitfalls of social fintech platforms. *Journal of the Association for Information Systems*, *24*(5), 1292-1312.

Soh, C., Chua, C. E. H., & Singh, H. (2011). Managing diverse stakeholders in enterprise systems projects: a control portfolio approach. *Journal of Information Technology*, *26*(1), 16-31.

Soltani Delgosha, M., Hajiheydari, N., & Olya, H. (2024). A person-centred view of citizen participation in civic crowdfunding platforms: A mixed-methods study of civic backers. *Information Systems Journal*.

Tams, S., Grover, V., & Thatcher, J. (2014). Modern information technology in an old workforce: Toward a strategic research agenda. *The journal of strategic information systems*, *23*(4), 284-304.

Tan, C.W. and Pan, S.L., (2003). Managing e-transformation in the public sector: an e-government study of the Inland Revenue Authority of Singapore (IRAS). *European Journal of Information Systems*, *12*, pp. 269-281

Tan, T., Zhang, Y., Heng, C. S., & Ge, C. (2020). Empowerment of grassroots consumers: a revelatory case of a Chinese fintech innovation. *Journal of the Association for Information Systems, Forthcoming*.

Tarafdar, M., Rets, I., & Hu, Y. (2023). Can ICT enhance workplace inclusion? ICT-enabled workplace inclusion practices and a new agenda for inclusion research in Information Systems. *The Journal of Strategic Information Systems*, *32*(2), 101773.

Thomas, K. W. (1992). Conflict and conflict management: Reflections and update. *Journal of organizational behavior*, 13(1), 265-274.

Tong, Y., Tan, C. H., Sia, C. L., Shi, Y., & Teo, H. H. (2022). Rural-Urban Healthcare Access Inequality Challenge: Transformative Roles of Information Technology. *MIS Quarterly, 46*(4). 1937–1982.

Trauth, E. M. (2013). The role of theory in gender and information systems research. *Information and Organization*, *23*(4), 277-293.

Trauth, E., & Connolly, R. (2021). Investigating the nature of change in factors affecting gender equity in the IT sector: a longitudinal study of women in Ireland. *MIS Quarterly, 45*(4). 2055–2100.

Ritchie, J., & Spencer, L. (1994). *Qualitative data analysis for applied policy research*. In A. Bryman & R. G. Burgess (Eds.), Analysing qualitative data. London: Routledge.

Ye, J., & Jensen, M. (2022). Effects of introducing an online community in a crowdsourcing contest platform. *Information Systems Journal, 32*(6), 1203-1230.

Yin, K., Fang, X., Chen, B., & Sheng, O. R. L. (2023). Diversity preference-aware link recommendation for online social networks. *Information Systems Research, 34*(4), 1398-1414.

Wang, Q., & Ngai, E. W. (2022). Firm diversity and data breach risk: a longitudinal study. *The Journal of Strategic Information Systems*, *31*(4), 101743.

Wass, S., Thygesen, E., & Purao, S. (2023). Principles to Facilitate Social Inclusion for Design-Oriented Research. *Journal of the Association for Information Systems*, *24*(5), 1204-1247.

Windeler, J. B., Urquhart, C., Thatcher, J. B., Carter, M., & Bailey, A. (2023). Special Issue Introduction: JAIS Special Issue on Technology and Social Inclusion. *Journal of the Association for Information Systems, 24*(5), 1199-1203.

Winter, S. J., & Saunders, C. (2019). The personal in the policy cascade. *Journal of the Association for Information Systems, 20*(11), 1692–1699.

Zhang, D., Lowry, P. B., Zhou, L., & Fu, X. (2007). The impact of individualism—collectivism, social presence, and group diversity on group decision making under majority influence. *Journal of Management Information Systems*, *23*(4), 53-80.

Zhou, S., Loiacono, E.T. and Kordzadeh, N., 2023. Smart cities for people with disabilities: a systematic literature review and future research directions. *European Journal of Information Systems*, pp.1-18.

Interoperable Financial Technologies Breaking Down Barriers for Financial Inclusion of Smallholder Farmers: a case study of e-Credit

Muhammad Mushaf Khan

Maria Kutar

Salford Business School

Completed Research

Abstract

Interoperable financial technologies (i-FinTech) present transformative solutions to supply-side barriers in rural financial inclusion. This study examines how i-FinTech adoption reduces digital exclusion and influences formal financial institutions by reshaping lending behaviour, operational strategies, and policy frameworks to improve credit access for smallholder farmers). Using the Punjab Government's e-Credit initiative as a case study, the research explores the role of mobile wallets, APIs, and digital collateralization in fostering seamless connections between FIs, public institutions, technology and mobile network providers. Employing a qualitative methodology, including semi-structured interviews and thematic analysis of policy documents, the paper provides insights into supply-side barriers and practical evidence of institutional change. Findings highlight i-FinTech's potential to reduce digital divides, and address Financial Institution hesitations to engage with smallholder farmers. This paper offers actionable insights for policymakers and practitioners, demonstrating how i-FinTech can drive rural financial inclusion by aligning institutional strategies with digital innovation.

Keywords: Interoperable Financial Technologies (i-FinTech), Interoperability, FinTech, Financial Inclusion, Digital Divide, Pakistan.

1. Introduction

Formal financial institutions (FIs) face enduring challenges in extending credit to rural populations, particularly smallholder farmers (SHFs), who constitute a significant yet underserved segment of the global agricultural economy. These challenges include high operational costs, limited rural infrastructure, and the absence of scalable mechanisms to assess and mitigate risk. For SHFs in Punjab, Pakistan, a region that contributes significantly to national food security and agricultural exports, these barriers result in a reliance on informal lenders, perpetuating cycles of financial exclusion and economic vulnerability. The advent of interoperable financial technologies (i-FinTech) offers new possibilities for overcoming these institutional constraints. Tools such as mobile wallets, APIs, and digital collateralization platforms enable FIs to streamline operations, reduce costs, and improve transparency in credit delivery. However, making these tools available to SHFs is a complex challenge in a region characterised by digital divides, with many SHFs starting from a point of digital exclusion. This paper focuses on the use of interoperable technologies to address supply-side barriers to financial inclusion, analysing how i-FinTech influences the lending behaviour and operational strategies of FIs. Taking a case study approach, we examine the Punjab Government's e-Credit initiative to understand how i-FinTech adoption can influence institutional practices and credit policies, creating a more inclusive digital financial ecosystem. By examining the collaborative roles of financial and non-financial stakeholders, the work evaluates the potential of i-FinTech to address financial and digital exclusion.

Most existing studies have either concentrated on the technical aspects of interoperable technologies, such as the development of APIs, mobile wallets, and blockchain-based solutions, or examined demand-side barriers like limited financial literacy, risk aversion, and lack of trust faced by smallholder farmers (Bökle et al., 2022; Borgogno & Colangelo, 2019; Bourreau & Valletti, 2015; Grzybowski et al., 2023). There is limited research on how interoperable technologies can address these challenges in digitally excluded rural and agricultural contexts. We address this gap by examining the use of supply-side approaches that leverage interoperable financial technologies to reach excluded SHFs, reducing digital divides and providing access to credit.

For the Information Systems community, the digital divide represents a compelling area of study. The interplay between i-FinTech, digital literacy, and agricultural practices is a rich field for understanding how digital platforms can be leveraged to overcome traditional barriers in rural and underserved areas. The challenge lies not only in developing technology but in understanding the sociocultural, economic, and infrastructural factors that affect its adoption and use. By examining how i-FinTech platforms can help to bridge the digital divide, we contribute to a broader understanding of digital transformation in agriculture, exploring the ways technology can drive equitable growth and sustainability in marginalised communities. In this regard, the case of smallholder farmers in Pakistan provides an invaluable opportunity to study the nuanced impacts of digital access and financial inclusion, offering insights that can be applied to similar contexts globally.

The paper is structured as follows: in Section 2 we provide an overview of the literature on the digital divide, and outline the context of digital and financial exclusion in Punjab. Section 3 provides detail of the methodology employed, and Section 4 details the e-Credit case study and key findings. We provide discussion and conclusions in section 5.

2. Background

2.1 The Digital Divide

The digital divide has been recognised as an important topic both by scholars and through its inclusion by organisations such as the United Nations (Kim et al., 2022) and Organisation for Economic Co-operation and Development (OEDC) (Lavrikova et al., 2024) and by other academics (Lythreatis et al., 2022). Whilst Rogers (1962) Diffusion of Innovation Theory sought to explain how new ideas and technology spread (Rogers et al., 2014), it has been recognised that the adoption of technology is in practice uneven, influenced by various factors which leads to divides between groups. The OECD identifies ongoing divides in connectivity and in Internet use which impact the even growth of the digital technology ecosystem. Early research into the digital divide focused on the disparities in individuals' access to technology resources, including computers and the Internet (Van Dijk, 2006). However, as the technology ecosystems have developed, the conceptualisation of the divide has become more nuanced, examining access to technology / connectivity (access), the capability to use it effectively (capability), and the capacity to gain potential outcomes from it due to capability constraints and other contextual factors (outcomes). Wei et al. (2011) cited in Carter et al. (2020) explain it as follows:

> *"The digital access divide (the first-level digital divide) is the inequality of access to information technology (IT) in homes and schools. The digital capability divide (the second-level digital divide) is the inequality of the capability to exploit IT arising from the first-level digital divide and other contextual factors. The digital outcome divide (the third-level digital divide) is the inequality of outcomes (e.g., learning and productivity) of exploiting IT arising from the second-level digital divide and other contextual factors."*

The digital divide is recognised not as a single divide but multiple divides (Heeks, 2022), and these are apparent from different perspectives as technology evolves, such as growing recognition of AI divides. In this paper we are focused on the digital divides experienced by smallholder farmers in the Punjab; these include access to technology and connectivity, which in turn influence the ability to access financial institutions and credit, this is explored in more detail in subsequent sections below.

2.2 Context – Region and Farmers

The fertile plains of Punjab, Pakistan, have long been the foundation of the nation's agricultural economy, nourishing both its people and its industries. With its vast expanses of rich soil, favourable climate, and abundant water resources from the mighty Indus River, Punjab plays an integral role in feeding the nation and sustaining its economic growth. This region, known as the breadbasket of Pakistan, contributes a sizeable portion of the country's agricultural output and approximately 60% of the nation's exports (Agriculture Department, 2025). Punjab's vast agricultural landscape also fuels its position as a driver of national GDP, accounting for nearly 21% of the economy (Azam & Shafique, 2017). However, despite the region's undeniable importance, the agricultural sector, especially smallholder farmers, faces persistent challenges that threaten their livelihoods and the broader food security of the country.

Since 1947, the population of Punjab, Pakistan, has seen a dramatic increase. From an estimated 20 million people at the time of independence in 1947, the population has surged to approximately 127.4 million in 2024. This exponential growth has had profound implications for agricultural land ownership. Historically, large agricultural estates were common, but as the population grew, these lands were subdivided through inheritance, resulting in a significant rise in the number of smallholder farmers. Today, around 91% of farmers in Punjab own less than 12.5 acres of land. This fragmentation has been driven by the need to distribute land among multiple heirs, leading to smaller and smaller plots over generations. The trend of increasing smallholder farmers is expected to continue. By 2050, Punjab's population is projected to reach 176 million. This population pressure will likely further subdivide agricultural lands, increasing the number of smallholder farmers. Beneath the statistics of GDP contributions and workforce employment lies the harsh reality of smallholder farmers in Punjab, who manage over 65% of the agricultural land but remain marginalized in many ways.

"Census of agriculture 2016-17 shows that there were 5,249,800 agriculture farms in Punjab, these farms consist majorly of very small farms. 42% of the farms are even less than one hectare. Farms ranging from one hectare to 10 hectares make up half of the total number

of farms and they occupy 68.9% of the total area." (Agriculture Department of Government of Punjab, 2025)

2.3 Barriers which lead to financial exclusion

The smallholder farmers, managing the bulk of Punjab's farmland, continue to grapple with access to formal credit that meets their unique needs. The geographical dispersion of these farmers, scattered across remote and often inaccessible locations, creates significant cost burdens for traditional financial institutions, fuelling their reluctance to extend their services. With high operational costs, financial institutions tend to focus on larger clients, leaving small-scale borrowers underserved. The costs associated with building branch infrastructure, deploying loan officers, and maintaining administrative processes often outweigh the perceived benefits of serving these rural communities. Research suggests that it costs between five to ten times more to serve a smallholder farmer compared to a larger client, creating a disincentive for financial institutions to address their needs (Hazell & Varangis, 2020; Poulton et al., 2010; Stringer et al., 2020). These farmers are left in a bind, unable to access credit from formal institutions but unwilling to completely rely on the exploitative practices of informal moneylenders (Adams, 2019; Nannozi, 2019).

Financial inclusion in Pakistan, including Punjab, has historically been low. According to Honohan (2008), Pakistan had a financial inclusion index score of just 12% in the early 2000s, which was significantly lower than neighbouring countries like India and Bangladesh. By 2017, the World Bank's Global Findex database reported that only 21% of adults in Pakistan had an account at a formal financial institution (Demirguc-Kunt et al., 2018). Recent data indicates a steady improvement in financial inclusion. According to the Karandaaz Financial Inclusion Survey (K-FIS) 2022, financial inclusion in Punjab has exceeded 30% (Khan, 2023). This increase in financial inclusion, specifically in the rural areas where the small landholding farming community rests, is primarily attributed to the proliferation of digital financial services and mobile banking platforms, which have made financial services more affordable and accessible to rural and underserved populations. Additionally, the integration of digital advisory services through mobile applications has further contributed to this trend, providing personalized guidance and support to individuals.

The need for financial inclusion for these smallholders is critical. Access to financial services such as credit, savings, and insurance can significantly enhance their economic stability and productivity. Financial inclusion enables smallholder farmers to invest in better seeds, fertilizers, and irrigation systems, which can improve crop yields and income (Ahmed et al., 2017). Despite extensive government support and the rise of microfinance initiatives, smallholder farmers in Punjab face numerous barriers in accessing formal financial services. Formal account ownership, a key indicator of financial inclusion, has seen modest growth, rising from 10% in 2011 to 21% in 2021 across Pakistan. Over 115 million adults remain unbanked, with more than four out of five adults without bank accounts indicating they require assistance to use formal financial services. Many still resort to informal sources, including friends, family, and local moneylenders, for credit. According to the World Bank's Global Findex Database 2021 (Demirgüç-Kunt et al., 2022), nearly two-thirds of the population relies on such informal sources, a phenomenon that perpetuates cycles of poverty and financial exclusion.

The challenges do not end with operational costs. Traditional credit assessment methods, often reliant on formal documentation and collateral, present further barriers to smallholder farmers, who often lack the necessary paperwork or access to formalized assets. A World Bank study revealed that approximately 70% of loan applications from smallholder farmers in Pakistan are rejected due to insufficient documentation. Without proper credit histories, land titles, or official identification, these farmers are viewed as high-risk borrowers, further compounding their exclusion from formal financial services. This gap forces most smallholder farmers into informal credit markets, where exorbitant interest rates and exploitative lending practices trap them in an endless cycle of debt.

2.4 The Digital Divide and Smallholder Farmers

In the realm of agriculture, the digital divide poses a significant challenge to smallholder farmers. While digital agriculture offers the promise of revolutionizing farming practices through precision information and data-driven decision-making, its benefits remain largely inaccessible to those who lack digital literacy and resources. The absence of reliable internet connectivity and supportive infrastructure in rural areas exacerbates this divide. Without access to digital platforms, smallholder farmers are isolated from critical information on soil health, crop conditions, and climate patterns. This isolation limits their ability to adapt to the changing agricultural landscape, including the pressing threats of climate change and environmental degradation.

The lack of access to information services also hinders smallholder farmers' ability to engage in sustainable agricultural practices (Krell et al., 2021; Rodriguez et al., 2009). Digital platforms and extension services can disseminate valuable knowledge on crop rotation, organic farming, and integrated pest management, enabling farmers to adopt more environmentally friendly approaches (Fabregas et al., 2019; Munthali et al., 2018; Naika et al., 2021; Rajkhowa & Qaim, 2021). The philosophical implications of this digital divide are profound. It perpetuates a cycle of inequality, where those with access to information and technology have a distinct advantage over those who do not. This inequality can lead to further marginalization and poverty, particularly in rural areas.

Another major barrier, that stems from access to finance, is the limited access to updated and modern information services, which includes digital agricultural extension and advisory services. These services are essential for providing farmers with up-to-date knowledge on best practices, market trends, weather forecasts, and pest management strategies. However, many smallholder farmers in developing regions, especially in Pakistan, are unable to access these services (Deichmann et al., 2016)ue to inadequate financial resources and digital infrastructure leading to low digital literacy, and insufficient investments in digital human capital development (Gumbi et al., 2023).

The promise of interoperable financial technologies (i-FinTech) lies in their potential to bridge this divide, specifically for smallholder farmers who have been historically excluded from traditional financial systems. i-FinTech platforms can provide a digital lifeline, connecting farmers to credit, insurance, and financial services that are crucial for investing in modern agricultural practices. With the integration of i-FinTech, digital platforms can become more accessible, leveraging mobile technology and user-friendly interfaces to accommodate those with limited digital literacy. Furthermore, i-FinTech has the potential to act as a catalyst in fostering digital inclusion by incentivizing financial institutions and stakeholders to invest in digital infrastructure and capacity-building initiatives. By making financial services accessible, i-FinTech can drive demand for digital literacy training and technological adaptation, gradually narrowing the digital divide. This effort aligns with broader goals of digital agriculture, positioning i-FinTech as a key interest for the IS community seeking to understand how digital platforms can support inclusion and equitable access in underserved regions. In the next section we detail the methodology employed to examine these matters in the research.

3. Methodology

The research presented here aimed to identify changes in agriculture credit policy, behaviour of financial institutions, and the resulting impact for digitally and financially excluded SMFs following the implementation of interoperable financial technologies. This aim was to understand the lived experiences of those who navigate these changes, both the smallholder farmers seeking inclusion and the financial institutions adjusting their strategies and practices. By embedding the human element, this research seeks to understand how policy changes ripple through the layers of financial and agricultural ecosystems, reshaping the landscape for those who are often marginalized. The methodological approach is thus both philosophical and pragmatic. It acknowledges that the evolution of credit policy and institutional behaviour is not merely a mechanical process but a complex human journey, influenced by culture, regulation, and the socio-economic realities of smallholder farmers and the technology tools. The study aims to capture the human narrative around interoperable financial technology. This human-centric view is essential for a deeper understanding of i-FinTech's role within Information Systems, to capture the essence of how digital solutions can foster inclusivity and empowerment in a context where access has long been a barrier, and a key contributor to digital divides.

The research takes a qualitative case study approach, examining the e-Credit system as case study. e-Credit is selected as a suitable case due to its comprehensive nature. There are other examples of interoperable i-Fin-Tech solutions which aimed to address financial inclusion, but these are more limited in scope. For example M-Pesa and UPI do not integrate lending and agriculture credit (Jack & Suri, 2011; Kumar et al., 2022; Mbiti & Weil, 2015; Mukherjee & Banerjee, 2023), whilst Braizils' Pix mobile banking, and China's AliPay are primarily urban focused (Chong, 2019; Duarte et al., 2022; Li et al., 2019). The e-Credit initiative addresses matters of both financial and digital exclusion due to the characteristics of the target population outlined in section 2 above, and therefore provides a suitable case to explore these matters.

The research was designed to capture diverse perspectives from both financial and non-financial institutions involved in the e-Credit project. The journey began with identifying and selecting key stakeholders, representatives from financial institutions, and regulatory bodies like the State Bank of Pakistan, Punjab Information & Technology Board, and Punjab Land Record Authority. These institutions, each playing a unique role, provided a window into the operational dynamics and challenges faced in implementing interoperable technologies. A series of 26 interviews were conducted with these stakeholders during September 2023 to April 2024. In parallel, key documents such as the 'Growth Strategy Punjab 2017', the e-Credit Detail Document, the 'Punjab Agriculture Policy 2018', the 'Punjab Agriculture Sector Plan 2015', the 'National Financial Inclusion Strategy 2023', and the 'Licensing and Regulatory

Framework for Digital Banks 2022' were meticulously analysed. These documents served as historical and policy-oriented reference points, grounding the participants' testimonials in the broader strategic context of the e-Credit system, and enabling triangulation of findings. Interview questions were structured around key themes. Table 1 (below) outlines the key themes explored during interviews, sample questions posed to stakeholders, and the rationale for each theme, aligning with existing literature on fintech adoption, operational outcomes, technical challenges, and policy implications.

Key Theme	Sample Interview Question	Rationale
Adoption Factors	What factors influenced your institution's decision to adopt or not adopt interoperable technologies?	Helps identify technological, regulatory, and institutional drivers of adoption, aligning with research on fintech adoption in financial institutions (Arner et al., 2020).
Operational Outcomes	How has the implementation of interoperable technologies affected your institution's efficiency, outreach, and transparency?	Explores whether interoperable technologies adoption leads to cost savings, improved service delivery, and reduced fraud, in line with digital financial inclusion literature (World Bank, 2022).
Technical & Regulatory Challenges	What technical or regulatory challenges has your institution faced in integrating interoperable technologies?	Identifies barriers such as API standardization, cybersecurity risks, and compliance issues, which impact interoperability (Arner et al., 2020; Zetzsche et al., 2017).
Policy Implications	How has the adoption of interoperable technologies influenced your institution's lending practices and credit policies?	Examines how financial institutions adapt their policies to accommodate digital credit systems, supporting research on fintech-driven policy shifts (Philippon, 2016).

Table 1: Key Themes, Sample Interview Questions, and Rationale for Exploring Interoperable Technologies in Financial Institutions

A thematic analysis of the interview data and case study documentation was conducted to identify key themes. We present these alongside the e-Credit case study in section 4.

4. Case Study: The e-Credit Ecosystem

In this section we detail the e-Credit case study, the details of which are underpinned by the document analysis and interviews. We first provide an overview of the goals, timeline and stakeholders, which is followed by examination of the way in which it utilises interoperable technologies. Finally, we explore the ways in which accessibility and affordability are addressed within e-Credit, together with the leadership approaches required to coordinate its implementation.

4.1 e-Credit Overview

In 2016, the Government of Punjab (GoPb), Pakistan, launched the 'Empowerment of Kissan (Farmer) through Digital & Financial Inclusion' (e-Credit) scheme, which ran from 2016 to 2022. The initiative, led by the Department of Agriculture (DoA), aimed to enhance agricultural development by integrating financial and digital services for smallholder farmers. To support this goal, the Agriculture Delivery Unit was established, bringing together experts in finance, strategy, policy, and information technology for accomplishing the objective as shared in figure 1. The unit worked with various financial institutions, including traditional banks and microfinance organizations, to increase financial access for farmers.

A key partnership with the Punjab Land Record Authority streamlined land verification and collateralization processes through digital mutations, addressing collateral barriers. Additionally, the Punjab Information Technology Board (PITB) provided the necessary digital infrastructure and expertise. The unit's Planning and Evaluation Cell oversaw the initiative's alignment with broader agricultural goals. Through these collaborations, the GoPb disbursed interest-free loans to smallholder farmers via commercial banks and microfinance institutions, aiming to bring them into the formal financial sector. The initiative provided recurring loans, with a maximum of five subsidized loans per farmer, before transitioning them to self-sufficiency within the formal financial industry.

The key goals and objectives of the 'e-Credit' initiative, as presented in the 'e-Credit Details Document of Government of Punjab, Agriculture Department, were to:

1. Extend credit to underbanked and unbanked smallholder farmers to help them improve on farm productivity and subsequently livelihood.
2. Disburse loans using distributed legers in instalments to ensure funds are used for agricultural purposes only and enable credit history for smallholder farmers.
3. Provide timely agriculture advice and support through ICTs.
4. Provide digital marketplaces and platforms for farmers to buy quality inputs and sell their produce and to collect and analyse data on agricultural practices, enabling informed decision making.

Figure 1 below provides an overview of the e-Credit initiatives ecosystem, illustrating the diverse stakeholders and services involved.

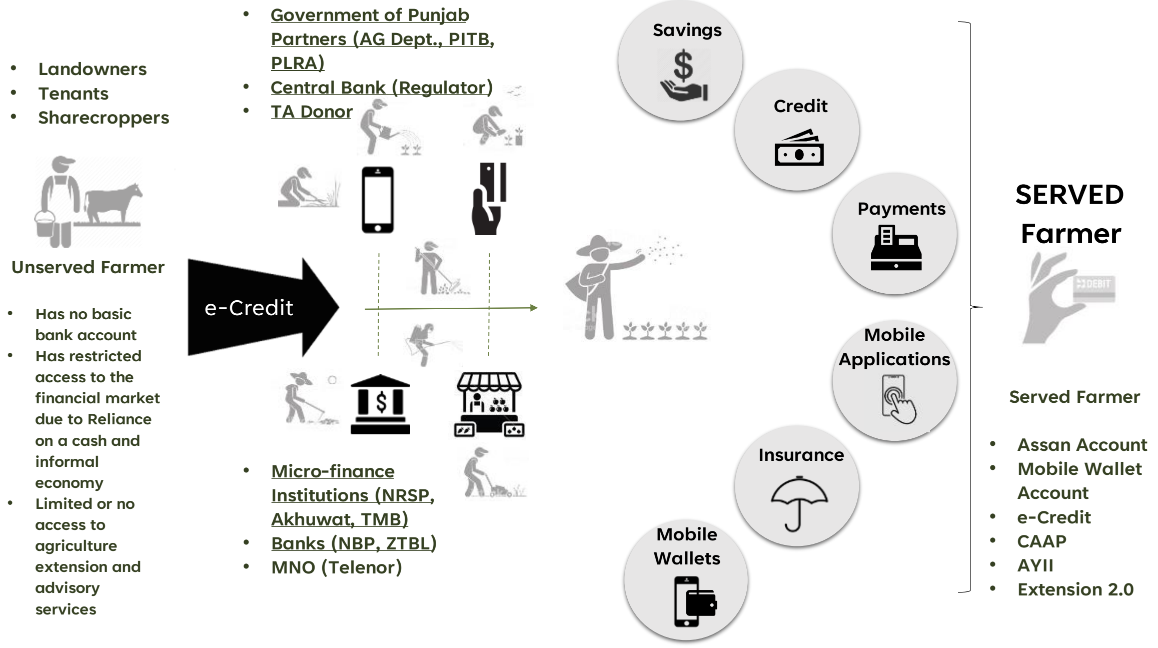

• Landowners
• Tenants
• Sharecroppers
Unserved Farmer
• Has no basic bank account
• Has restricted access to the financial market due to Reliance on a cash and informal economy
• Limited or no access to agriculture extension and advisory services
• Government of Punjab Partners (AG Dept., PITB, PLRA)
• Central Bank (Regulator)
• TA Donor
e-Credit
• Micro-finance Institutions (NRSP, Akhuwat, TMB)
• Banks (NBP, ZTBL)
• MNO (Telenor)
Savings
Credit
Payments
Mobile Applications
Insurance
Mobile Wallets
SERVED Farmer
DEBIT
Served Farmer
• Assan Account
• Mobile Wallet Account
• e-Credit
• CAAP
• AYII
• Extension 2.0

Figure 1: e-Credit partners and outcomes (Created by the author)

Figure 2 (below) provides a quantitative overview of the impact of the e-Credit initiative on agricultural financing for smallholder farmers in Punjab. It shows that a total amount of 72.4 billion PKR was disbursed through this initiative, highlighting the substantial financial support extended to the agricultural sector. The "Digital Mutations" metric, with 164,015 entries, represents the number of digital land record updates, indicating the adoption of 'e-Passbook', a step towards transparent and accessible land records, which can facilitate credit eligibility assessments. The total number of loans issued is 1,068,942, reflecting the initiative's outreach in delivering credit to farmers. Among the loan recipients, 513,447 are unique farmers, showcasing a broad base of beneficiaries. Significantly, 286,383 of these recipients are "Landless Farmers," demonstrating the program's inclusivity in reaching farmers who typically lack collateral for traditional loans. Another noteworthy statistic is that 73% of the recipients are "New to Bank" customers, indicating the program's success in onboarding previously unbanked individuals, thereby expanding financial inclusion. Moreover, the initiative granted 368,154 loans to female farmers, supporting gender inclusion within agricultural finance. The "Farmers Applied" metric, with 136,980 applicants, shows the level of demand and interest in the program. Finally, there are 11,828 "Active CAPP Users," suggesting a growing adoption of digital platforms by farmers for credit access and financial management, reflecting a shift towards digital financial ecosystems in the rural sector.

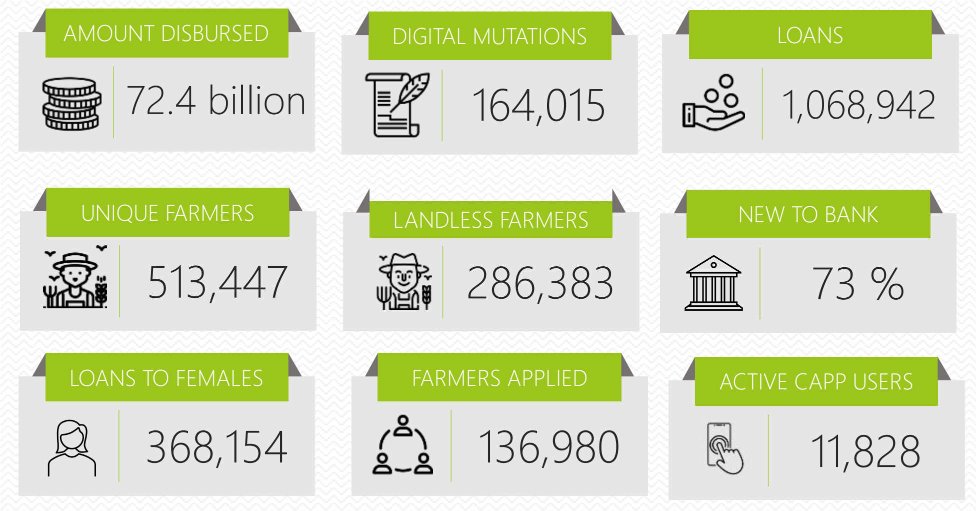
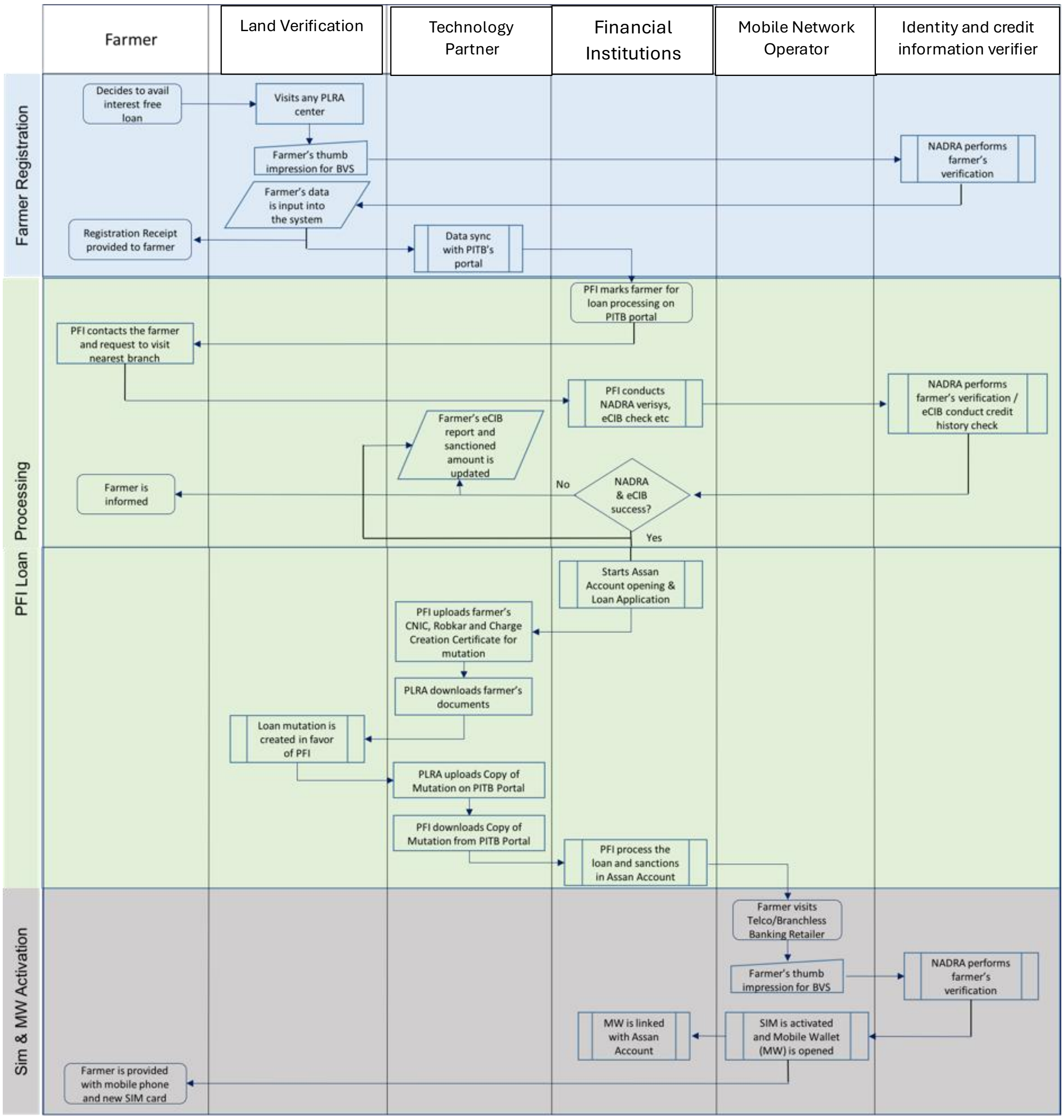


Figure 2: Progress of e-Credit (ADU, 2021)

4.1.1 Scheme Operation and Process Flow

e-Credit farmers were provided with 3G/4G-enabled smartphones with pre-installed farming applications, along with free SIM cards and data bundles. To enable farmers to utilize the full potential of this platform, hundreds of facilitation centres and booths were established across the province to ensure training of smallholder farmers at each Village level. e-Credit aimed to promote financial inclusion by providing subsidised loans to smallholder farmers Initially, some loans were disbursed, without opening of formal bank accounts of smallholder farmers, via cheques. These cheques were cashed at the respective financial institutions, which required the smallholders' farmers to visit the branch in person. However, during the second cropping season, loan disbursement in formal bank accounts was made mandatory to ensure timely disbursement to ensure avoidance of late repayments and defaults; this also embedded use of the technology by farmers. e-Credit was delivered in partnership with formal financial institutions, two commercial banks and three microfinance institutions, to ensure that the registered smallholder farmers had access to formal bank accounts. These formal bank accounts enabled farmers to receive and repay loans. Figure 3 below illustrates the process flow.



Farmer
Land Verification
Technology Partner
Financial Institutions
Mobile Network Operator
Identity and credit information verifier
Farmer Registration
Decides to avail interest free loan
Visits any PLRA center
Farmer's thumb impression for BVS
NADRA performs farmer's verification
Farmer's data is input into the system
Registration Receipt provided to farmer
Data sync with PITB's portal
PFI Loan Processing
PFI marks farmer for loan processing on PITB portal
PFI contacts the farmer and request to visit nearest branch
PFI conducts NADRA verisys, eCIB check etc
NADRA performs farmer's verification / eCIB conduct credit history check
Farmer's eCIB report and sanctioned amount is updated
No
NADRA & eCIB success?
Yes
Farmer is informed
Starts Assan Account opening & Loan Application
PFI uploads farmer's CNIC, Robkar and Charge Creation Certificate for mutation
PLRA downloads farmer's documents
Loan mutation is created in favor of PFI
PLRA uploads Copy of Mutation on PITB Portal
PFI downloads Copy of Mutation from PITB Portal
PFI process the loan and sanctions in Assan Account
Sim & MW Activation
Farmer visits Telco/Branchless Banking Retailer
Farmer's thumb impression for BVS
NADRA performs farmer's verification
SIM is activated and Mobile Wallet (MW) is opened
MW is linked with Assan Account
Farmer is provided with mobile phone and new SIM card

Figure 3: e-Credit & 'CAPP' process flow (ADU, 2016)

4.2 Interoperability and e-Credit

Interoperability was ensured through technology led farmer-centric service delivery. The technologies underpinning this are summarised in Table 2 below. It is important to note that in addition to the technical and semantic interoperability offered by i-Fintech, organisational interoperability to align coordinate service delivery was also an important aspect. To ensure convenient and transparent disbursement of micro-credit, farmers were registered through the 'Farmer Registration Portal'. Financial institutions could disburse loans directly to farmers' Assan accounts or mobile wallets. The intention was to make it easier and faster for farmers to receive their loans. It was intended to allow farmers to repay their loans from their Assan accounts or mobile wallets using the interoperable agent which had a more comprehensive footprint in the rural areas compared to the limited branch network. Farmers could use their Assan accounts or mobile wallets to make payments for agricultural inputs, such as seeds and fertilizers. By addressing the logistical and cost barriers, access to the agricultural inputs was made convenient.

Name of Interoperable Technology in 'e-Credit'	**Interoperable Technologies**	**Interoperability Type**
Farmer Registration Portal	Farmer Know Your Customer (KYC) registration portal, Application programming Interfaces (APIs)	**Technical Interoperability**: APIs enable the portal to connect with other systems (e.g., NADRA for identity verification). **Semantic Interoperability**: The KYC portal ensures that farmer data (e.g., name, CNIC, location) is standardized and understood across different systems.
Agri-loan Portal	Digital Information Recording, Monitoring and Evaluation Platform, APIs	**Technical Interoperability**: APIs allow the Agri-loan portal to share data with banks, mobile wallets, and government systems. **Organizational Interoperability**: The platform requires coordination between financial institutions, government agencies, and farmers to record and monitor loan data.
Assan account & Mobile Wallets	Distributed ledgers, EMV Chip Technology, Unstructured Supplementary Service Data (USSD), Real-Time Gross Settlement (RTGS), QR code, 1 Link and PakPay.	**Technical Interoperability:** Distributed ledgers, EMV chips, USSD, RTGS, QR codes, and payment systems like 1Link and PayPak enable seamless transactions across different platforms. **Semantic Interoperability:** QR codes and standardized payment protocols ensure that data (e.g., transaction amounts, recipient details) is understood across systems.
Connected Agriculture Punjab Platform (CAPP)	Mobile Applications, APIs	**Technical Interoperability:** APIs enable CAPP to integrate with other systems (e.g., Agri-loan portal, mobile wallets). **Organizational Interoperability:** CAPP requires collaboration between farmers, financial institutions, and government agencies to provide advisory services and financial inclusion.
e-Passbook	APIs	**Technical Interoperability:** APIs enable the e-Passbook to sync with Agri-loan portal to provide real-time updates on collateralisation.

Table 2 : Interoperable Financial Technologies of e-Credit

Interoperable financial technologies like the ones clubbed under the e-Credit initiative, emerge as a beacon of hope for smallholder farmers. The initiative sought to address the entrenched supply-side barriers that have long hindered the financial inclusion of Punjab's farming community. Interoperable technologies, including Assan Banking Accounts, Mobile Wallets, Application Programming Interfaces (APIs), Mobile Applications, and digital Portals provided innovative solutions that can help bridge the gap between smallholder farmers and formal financial services. These digital platforms allowed for the seamless integration of information and payment systems, enabling farmers to access financial products and services without the traditional hurdles posed by physical infrastructure and documentation requirements.

Interoperability was at the heart of the e-Credit system and incorporates administrative as well as the more specialised fin-tech elements.

"*The e-Credit 'Farmer Registration Portal' was designed with a robust framework to standardize data collection processes. This ensured that all farmer data is uniformly captured and maintained, facilitating seamless access and assessment by participating financial institutions. By adhering to the prudential regulations set forth by the Central Bank of Pakistan, we ensured that the data standardisation integrity and security was upheld, thereby fostering trust and efficiency in the credit application process."* (Director IT, Punjab Land Record Authority, 2024)

This portal enabled the verification and registration of farmers wanting access to e-Credit facility and sought to reduce farm to bank distance. The existing infrastructure and workforce of revenue offices/offices was trained to record interest of agricultural credit. After the registration process was completed, the farmer's data was integrated through Application Programming Interface (API), an interoperable technology, with the 'Agri-loan Portal'.

"*The portal enabled participating financial institutions to select registered farmers, avoid duplication, track credit approvals, facilitate the opening of bank accounts and lastly ensure digital collateralisation of land for credit. We streamlined the entire process to further reduce both the time and cost ensuring greater efficiency and transparency in the system*." (Joint Director, Punjab Information Technology Board, 2024)

This digital portal enabled the selection of registered farmers while avoiding duplication, tracking of credit approvals, and opening of bank accounts and 'Digital Mutation' of land (which sought to reduce both time and cost of collateralisation and ensured efficiency). Furthermore, the 'Agri-loan Portal', another interoperable platform enabled the tracking of loan disbursement and viewing of seasonal reports mentioning useful data of project implementation and record 'e-Passbook' Another purpose of the 'Agri-loan Portal' was to reduce or if possible, eliminate credit rationing.

"*Asaan Bank Accounts and Mobile Wallet Accounts have been game-changers in promoting financial inclusion. They provided a simple, accessible way for us the financial institutions and smallholder farmers, especially in rural and underserved areas, to engage with the formal financial system. By allowing users to perform transactions securely from their basic mobile phones, these accounts have significantly reduced barriers to banking, enabling more people to save, transfer money, and access financial services with ease."* (NKSP, 2024)

Assan accounts and mobile wallets, highly interoperable payment accounts, played a crucial role in the success of the e-Credit scheme in Punjab, Pakistan. Assan account is a digital financial services platform that provides users with access to a range of financial services, including savings, loans, and payments. The Assan account was linked to the user's mobile phone number, making it easy for users to access financial services on the go. Mobile wallets are another interoperable transaction account that allowed users to store, send, and receive information and money using their mobile phones. Mobile wallets are convenient and affordable, making them a popular way for people to make payments in developing countries. The use of Assan accounts and mobile wallets played a significant role in the success of the e-Credit scheme in Punjab, Pakistan as they required fewer trips to a formal banking institution for opening, operating, and maintaining.

Lastly, a 'Connected Agriculture Punjab Platform' (CAPP) the interoperable advisory platform was established to cater to best agriculture practices by connecting the smallholder farmer with all stakeholders in the agriculture value chain, including but not limited to agriculture input providers, research institutions, commodity buyers, supply chain services providers, and agriculture extension workers. CAPP comprised of several mobile phone-based applications bringing the much needed (and previously ignored) advisory services from the financial institutions to the palms of smallholder farmers. The mobile phone base applications included Farmer Registration, Zarai Mashwara (Agricultural Advice), Subsidy Claim, Kissan (Farmer) TV, Kasht (Produce) Calculator, Muqami Musam (Local Weather), Kissan Dukan (Farmer Shop), Mandi (Market), Zarai Jantry (Agriculture Register). Annexure A provides visual screen shots of these mobile applications.

"*The Connected Agriculture Punjab Platform (CAPP) has been instrumental in transforming agricultural practices. The aim was to connect smallholder farmers with key stakeholders across the agriculture value chain to create a comprehensive advisory ecosystem that promotes best practices and enhances productivity*." (Strategy & Policy Advisor, Agriculture Department, government of Punjab, 2022)

To increase digital literacy among the farmer community and provide them with real-time weather updates, pesticide warnings, best crop practices, subject matter expert's advice, etc., subsidised smartphones (costing £2 to £5) were provided. These smartphones include preinstalled applications aimed at providing the necessary digital platform to the farming community. However, only a small number of smartphones were disbursed after which this digital intervention was halted due to political interference.

4.3 Digital Literacy

The adoption of digital financial services requires a certain level of digital literacy, which is often lacking among rural populations. The government established facilitation centres and training booths across Punjab to educate farmers on how to use digital platforms for loan applications, repayments, and accessing agricultural advice. These programs were tailored to the needs of smallholder farmers, many of whom had limited prior exposure to digital technologies. To bridge the digital divide, the initiative provided 3G/4G-enabled smartphones pre-installed with farming applications to registered farmers. These smartphones, along with free SIM cards and data bundles, enabled farmers to access digital services and participate in the e-Credit ecosystem. The Agri-loan Portal and Connected Agriculture Punjab Platform (CAPP) Were designed with user-friendly interfaces to ensure that even farmers with limited digital literacy could navigate the platforms. Features such as Unstructured Supplementary Service Data (USSD) codes, mobile applications, and QR Payments systems allowed farmers to access services using basic mobile phones, further reducing barriers to adoption. The programmes supporting digital literacy were essential to address issues around digital inclusion at the access and capability levels, which in turn were necessary foundations to address financial exclusion.

4.4 Accessibility and Affordability

One of the central promises of interoperable technologies is their potential to significantly reduce operational costs for financial institutions and bringing operations efficiency. By digitizing processes and utilizing data-driven approaches, these technologies streamline the credit assessment process, automate decision-making, and optimize resource allocation. Research suggests that such innovations can reduce service costs by 30-50%, making it more financially viable for institutions to serve smallholder farmers. Moreover, interoperable platforms enable secure data sharing from various sources, including mobile money transactions, utility payments, and other digital footprints, creating more accurate credit profiles for farmers. These enhanced profiles can increase loan approval rates by 20-30%, providing smallholder farmers with much-needed access to formal credit.

The e-Credit initiative leverages these technological advancements to provide timely and accessible financial assistance to smallholder farmers, improving their cash flow and financial stability. Through the initiative, farmers are empowered to access critical inputs, such as seeds, fertilizers and machinery, at the right time in the agricultural cycle, enabling them to enhance productivity and crop yields. This, in turn, leads to better profitability and the ability to invest in improved farming practices and technologies, such as precision farming tools. By tailoring financial services to the needs of smallholder farmers, with flexible repayment plans that align with the agricultural cycle, the e-Credit initiative also alleviates the financial burden of debt.

At the heart of this initiative is the creation of a digital, robust, and centralized system for efficient loan management, tracking, and monitoring. This system ensures transparency and accountability, reducing the risks of mismanagement or corruption. ‘Mobile wallets’ and ‘Assan banking accounts’, powered by ‘APIs’, ‘mobile applications’ and other ‘digital portals’, simplify financial transactions and improve accessibility for farmers in even the most remote areas. The system provides a user-friendly interface that allows smallholder farmers, many of whom are digitally illiterate, to interact with financial services providers in a way that is intuitive and easy to understand. By doing so, the e-Credit initiative not only fosters financial inclusion but also promotes digital literacy, further empowering the rural population.

The interoperability of financial technologies within the e-Credit initiative is a game changer for Punjab’s smallholder farmers, offering them a digital lifeline in a world where financial exclusion has long been the norm. The ability to streamline processes, reduce operational costs, and enhance credit assessments has the potential to transform the way financial institutions engage with small-scale borrowers. More importantly, by creating a system that accommodates the financial realities of smallholder farmers, the initiative paves the way for increased agricultural productivity, food security, and economic prosperity in Pakistan. The challenges that have long plagued smallholder farmers in Punjab restricting access to finance, physical isolation, and a lack of tailored financial products are not insurmountable. The e-Credit initiative demonstrates that by leveraging interoperable financial technologies, these barriers can be dismantled. In doing so, it not only promotes financial inclusion but also lays the foundation for a more resilient and sustainable agricultural sector. With greater access to formal financial services, smallholder farmers can break free from the cycle of poverty and take control of their economic futures, contributing to a more prosperous and secure Pakistan.

In recent years, the discourse surrounding financial inclusion has become increasingly focused on the potential of digital financial technologies to address long-standing barriers, particularly for marginalized populations such as smallholder farmers. The proliferation of mobile banking and interoperable financial technologies has opened new avenues for enhancing access to formal financial services, especially in rural and underserved areas.

4.5 Leadership fostering Partnerships and Interoperability

The success of the e-Credit initiative lies in the coordinated efforts of multiple stakeholders from both financial and non-financial sectors "*We understood from the beginning that a digital solution like e-Credit could not succeed in isolation. We needed the expertise of financial institutions, mobile network providers, technology providers, and regulatory bodies to come together, and that's where the true success of this project lies*." (CHIEF, Planning & Evaluation Cell, Agriculture Department, Government of Punjab). This statement reflects the department's role as a coordinator, ensuring that each stakeholder played a part in driving financial inclusion. The Punjab Information Technology Board (PITB) provided the technological infrastructure rather the technology backbone that made e-Credit operational. The Joint Director of Digital Financial Services at PITB shared, "*Our role was to ensure the interoperability of platforms, enabling seamless data sharing between financial institutions and government departments.*" PITB's efforts were crucial in overcoming the logistical and infrastructural challenges of reaching remote farming communities. The Assistant Director remarked, "*It was a challenge to convince traditional financial institutions to adopt digital solutions, but once they saw the cost savings and the efficiency, they became more open to integration. The e-Credit initiative allowed us to scale solutions using interoperability that could be customized to the needs of smallholder farmers.*" This insight underscores the transformative potential of technology in reducing operational costs and expanding financial outreach, key objectives of the e-Credit initiative.

The Punjab Land Record Authority (PLRA) played a pivotal role in the e-Credit initiative by providing two critical functions: the 'Farmer Registration Portal' and the 'e-Passbook' system. The 'e-Passbook' system was another transformative solution introduced by PLRA, completely digitizing the process of land collateralization for farmers seeking loans. Historically, land collateralization involved a time-consuming and costly manual process. Banks were required to hire lawyers to verify land ownership through revenue offices, after which the collateralization process itself was conducted manually. This could take several months and incurred substantial costs due to legal fees, administrative delays, and the potential for corruption, including speed money to expedite procedures. The introduction of the 'e-Passbook' digitized the entire process, significantly reducing both time and cost. The system allowed for real-time verification of land ownership, automatically linking land records with financial institutions and eliminating the need for lawyers. The process, which previously took months, was now completed within three days. "*With the 'e-Passbook', we reduced the turnaround time for land collateralization to just 72 hours. This digital transformation not only saved time but also eliminated the opportunities for corruption that were prevalent in the manual process*," said the official from PLRA.

The microfinance institutions involved in the e-Credit initiative, including NRSP, Akhuwat, and Telenor Microfinance Bank (TMB), played a crucial role in ensuring the accessibility of credit for smallholder farmers. The chief executive from NRSP explained, "*We've always been committed to serving the underbanked, and the e-Credit initiative gave us the tools to reach even more farmers. By integrating with digital wallets and utilizing data from PLRA, we were able to reduce the time and cost associated with loan processing*." Other stakeholders were initially more cautious in their approach to the e-Credit initiative. A representative from ZTBL admitted, "*We were sceptical at first because serving smallholder farmers comes with high operational costs. But after seeing the success and ease of use for the interoperable financial technologies the efficiencies brought by digital integration, we decided to participate fully. And train more staff.*" ZTBL's role in the initiative helped bridge the gap between large financial institutions and smallholder farmers, demonstrating how digital solutions could reduce costs and improve service delivery. Similarly, the NBP viewed the initiative as a strategic opportunity to expand its rural banking services. An NBP senior official shared, "*We saw the e-Credit initiative as a way to diversify our portfolio and reach a market that had been largely underserved. The interoperable technology allowed us to lower the barriers to entry and provide more tailored financial products to farmers*." (Agriculture Finance Head, 2024) This statement highlights how the initiative has encouraged even the more traditional financial institutions to innovate and embrace digital financial solutions. The Central Regulatory Banking Body of Pakistan (SBP) played a key role in providing the regulatory framework that supported the adoption of interoperable technologies. A senior official explained, "*We were involved from the outset in ensuring that the e-Credit initiative complied with national regulations while also promoting innovation. Our role was to create an enabling environment that balanced innovation with financial stability and consumer protection*." (NKSP, 2024) Another official from SBP also highlighted the importance of collaboration between regulatory bodies and financial institutions: "*The success of the e-Credit initiative is a testament to what can be achieved when regulators and industry stakeholders work together. We provided the guidelines, which allowed the financial institutions to innovate without compromising on security and rights.*" (UASB, 2024)

4.6 Discussion

The e-Credit system is a complex collaboration between a large number of stakeholders, which has had significant impact in the Punjab region, as illustrated above. It can be seen that the leadership of key stakeholders in policy development and alignment of the e-Credit system design to interoperable principles was central to its success. These supported the operation of the ecosystem and provided the foundation for farmers to receive the digital tools and skills to access credit. Addressing the digital divide was an essential precursor to addressing financial exclusion, whilst the goal of addressing financial exclusion provided the scheme and resources to address the digital divide. In the following section we present our conclusions and identify future directions.

5. Conclusion: Interoperability and the Digital Divide

The e-Credit initiative has demonstrably expanded financial inclusion and empowered smallholder farmers in Punjab. Supported by interoperable technologies and efficient digital processes, the initiative has reduced both the cost and time of accessing formal financial services. Quantitative and qualitative data indicate that farmers have experienced improved financial well-being and greater control over their agricultural decisions. The program has also fostered a collaborative ecosystem between financial and non-financial institutions, positioning it as a scalable model for financial inclusion. While the initiative's success is clear, sustained effort is required to ensure its long-term impact. Continuous investment in digital infrastructure, flexible financial products, and ongoing stakeholder collaboration will be critical to maintaining the gains made and ensuring that the initiative remains a beacon of empowerment and sustainability for smallholder farmers in Punjab.

The insights from these key stakeholders highlight the multifaceted nature of the e-Credit initiative, with each institution contributing to its success in distinct yet interconnected ways. The collaboration between financial and non-financial institutions, combined with the interoperable technological infrastructure provided, created an ecosystem that facilitated the financial inclusion of smallholder farmers in Punjab. This case study not only illustrates the potential of interoperable financial technologies but also underscores the importance of cross-sector collaboration in driving meaningful social and economic change. An important element of the success of the scheme has been that it has been use as a lever to drive digitisation of processes – for example in e-Credit the move to digital land records was an essential step.

The e-Credit scheme was an interoperable digital initiative, intended to address financial exclusion, in a context where a significant digital divide existed. Bridging the digital divide was an essential first step to unlocking the financial inclusion aspects of the scheme and enabling access to other ICTs which support smallholder farmers. The findings provide evidence for the potential of interoperable technology to underpin schemes to address social and economic change as a mechanism to address the challenges of digital divides. This research has focused on a single case study, and the financial context. Further research is required to understand the extent to which this might be applicable in other contexts, but the results suggest that addressing social and economic issues can be an effective mechanism for addressing digital divides.

References

Adams, D. W. (2019). Taking a fresh look at informal finance. In *Informal finance in low-income countries* (pp. 5-23). Routledge.

ADU. (2021). *Steering Committee Presentation 2021*. Agriculture Delivery Unit

Agriculture Department, G. o. P., Pakistan. (2025). *Overview of Punjab Agriculture*. Retrieved 11/02/2025 from https://www.agripunjab.gov.pk/overview

Ahmed, U. I., Ying, L., Bashir, M. K., Abid, M., & Zulfiqar, F. (2017). Status and determinants of small farming households' food security and role of market access in enhancing food security in rural Pakistan. *Plos one*, *12*(10), e0185466.

Arner, D. W., Buckley, R. P., Zetzsche, D. A., & Veidt, R. (2020). Sustainability, FinTech and Financial Inclusion [Article]. *European Business Organization Law Review*, *21*(1), 7-35. https://doi.org/10.1007/s40804-020-00183-y

Azam, A., & Shafique, M. (2017). Agriculture in Pakistan and its Impact on Economy. *A Review. Inter. J. Adv. Sci. Technol*, *103*, 47-60.

Bökle, S., Paraforos, D. S., Reiser, D., & Griepentrog, H. W. (2022). Conceptual framework of a decentral digital farming system for resilient and safe data management. *Smart Agricultural Technology*, *2*, 100039.

Borgogno, O., & Colangelo, G. (2019). Data sharing and interoperability: Fostering innovation and competition through APIs. *Computer Law & Security Review*, *35*(5), 105314.

Bourreau, M., & Valletti, T. (2015). Enabling digital financial inclusion through improvements in competition and interoperability: What works and what doesn't. *CGD Policy Paper*, *65*, 1-30.

Carter, L., Liu, D., & Cantrell, C. (2020). Exploring the intersection of the digital divide and artificial intelligence: A hermeneutic literature review. *AIS Transactions on Human-Computer Interaction*, *12*(4), 253-275.

Chong, G. P. L. (2019). Cashless China: Securitization of everyday life through Alipay's social credit system—Sesame Credit. *Chinese Journal of Communication*, *12*(3), 290-307.

Deichmann, U., Goyal, A., & Mishra, D. (2016). Will digital technologies transform agriculture in developing countries? *Agricultural Economics*, *47*(S1), 21-33.

Demirguc-Kunt, A., Klapper, L., Singer, D., & Ansar, S. (2018). *The Global Findex Database 2017: Measuring financial inclusion and the fintech revolution*. World Bank Publications.

Demirgüç-Kunt, A., Klapper, L., Singer, D., & Ansar, S. (2022). *The Global Findex Database 2021: Financial inclusion, digital payments, and resilience in the age of COVID-19*. World Bank Publications.

Duarte, A., Frost, J., Gambacorta, L., Koo Wilkens, P., & Shin, H. S. (2022). Central banks, the monetary system and public payment infrastructures: lessons from Brazil's Pix. *Available at SSRN 4064528*.

Fabregas, R., Kremer, M., & Schilbach, F. (2019). Realizing the potential of digital development: The case of agricultural advice. *Science*, *366*(6471), eaay3038.

Grzybowski, L., Lindlacher, V., & Mothobi, O. (2023). Interoperability Between Mobile Money Agents and Choice of Network Operators: The Case of Tanzania. *Review of Network Economics*, *22*(1), 27-52. https://doi.org/doi:10.1515/rne-2023-0024

Gumbi, N., Gumbi, L., & Twinomurinzi, H. (2023). Towards sustainable digital agriculture for smallholder farmers: A systematic literature review. *Sustainability*, *15*(16), 12530.

Hazell, P., & Varangis, P. (2020). Best practices for subsidizing agricultural insurance. *Global food security*, *25*, 100326. https://doi.org/https://doi.org/10.1016/j.gfs.2019.100326

Heeks, R. (2022). Digital inequality beyond the digital divide: conceptualizing adverse digital incorporation in the global South. *Information Technology for Development*, *28*(4), 688-704.

Honohan, P. (2008). Cross-country variation in household access to financial services. *Journal of Banking & Finance*, *32*(11), 2493-2500.

Jack, W., & Suri, T. (2011). *Mobile money: The economics of M-PESA*.

Khan, I. (2023). *FINANCIAL INCLUSION INDEX FOR PAKISTAN*.

Kim, J. Y., Park, J., & Jun, S. (2022). Digital transformation landscape in Asia and the Pacific: Aggravated digital divide and widening growth gap.

Krell, N., Giroux, S., Guido, Z., Hannah, C., Lopus, S., Caylor, K., & Evans, T. (2021). Smallholder farmers' use of mobile phone services in central Kenya. *Climate and Development*, *13*(3), 215-227.

Kumar, A., Choudhary, R. K., Mishra, S. K., Kar, S. K., & Bansal, R. (2022). The growth trajectory of UPI-based mobile payments in India: Enablers and inhibitors. *Indian Journal of finance and banking*, *11*(1), 45-59.

Lavrikova, N. I., Magomaeva, L. R., Kochyan, G. A., Ponomarev, S. V., & Borshchevskaya, E. P. (2024). Assessing the digital divide in OECD and BRICS countries: implications for public policy. *International Journal of Technology, Policy and Management*, *24*(3), 285-302.

Li, J., Wang, J., Wangh, S., & Zhou, Y. (2019). Mobile payment with alipay: An application of extended technology acceptance model. *IEEE Access*, *7*, 50380-50387.

Lythreatis, S., Singh, S. K., & El-Kassar, A.-N. (2022). The digital divide: A review and future research agenda. *Technological forecasting and social change*, *175*, 121359.

Mbiti, I., & Weil, D. N. (2015). Mobile banking: The impact of M-Pesa in Kenya. In *African successes, Volume III: Modernization and development* (pp. 247-293). University of Chicago Press.

Mukherjee, S., & Banerjee, K. (2023). Importance of UPI in Socio-Economic Development: An Empirical Study. *AIMS International Journal of Management*, *17*(1).

Munthali, N., Leeuwis, C., van Paassen, A., Lie, R., Asare, R., van Lammeren, R., & Schut, M. (2018). Innovation intermediation in a digital age: Comparing public and private new-ICT platforms for agricultural extension in Ghana. *NJAS-Wageningen Journal of Life Sciences*, *86*, 64-76.

Naika, M. B., Kudari, M., Devi, M. S., Sadhu, D. S., & Sunagar, S. (2021). Digital extension service: quick way to deliver agricultural information to the farmers. In *Food technology disruptions* (pp. 285-323). Elsevier.

Nannozi, A. (2019). A case study: exploring the influence of the informal financial sector on food security among smallholder farmers in Uganda, Greater Luweero.

Philippon, T. (2016). *The fintech opportunity*. N. B. O. E. RESEARCH.

Poulton, C., Dorward, A., & Kydd, J. (2010). The Future of Small Farms: New Directions for Services, Institutions, and Intermediation. *World Development*, *38*(10), 1413-1428. https://doi.org/https://doi.org/10.1016/j.worlddev.2009.06.009

Rajkhowa, P., & Qaim, M. (2021). Personalized digital extension services and agricultural performance: Evidence from smallholder farmers in India. *Plos one*, *16*(10), e0259319.

Rodriguez, J. M., Molnar, J. J., Fazio, R. A., Sydnor, E., & Lowe, M. J. (2009). Barriers to adoption of sustainable agriculture practices: Change agent perspectives. *Renewable Agriculture and Food Systems*, *24*(1), 60-71.

Rogers, E. M., Singhal, A., & Quinlan, M. M. (2014). Diffusion of innovations. In *An integrated approach to communication theory and research* (pp. 432-448). Routledge.

Stringer, L. C., Fraser, E. D. G., Harris, D., Lyon, C., Pereira, L., Ward, C. F. M., & Simelton, E. (2020). Adaptation and development pathways for different types of farmers. *Environmental Science & Policy*, *104*, 174-189. https://doi.org/https://doi.org/10.1016/j.envsci.2019.10.007

Van Dijk, J. A. (2006). Digital divide research, achievements and shortcomings. *Poetics*, *34*(4-5), 221-235.

Wei, K.-K., Teo, H.-H., Chan, H. C., & Tan, B. C. (2011). Conceptualizing and testing a social cognitive model of the digital divide. *Information Systems Research*, *22*(1), 170-187.

World Bank. (2022). Financial Inclusion. *Financial inclusion is a key enabler to reducing poverty and boosting prosperity.* Retrieved 19/05/2022, from https://www.worldbank.org/en/topic/financialinclusion/overview#2

Zetzsche, D. A., Buckley, R. P., Barberis, J. N., & Arner, D. W. (2017). Regulating a revolution: From regulatory sandboxes to smart regulation. *Fordham J. Corp. & Fin. L.*, *23*, 31.

Annexure A: Glossary

GoP- Government of Pakistan

GoPb – Government of Punjab

AD – Agriculture Department

PITB – Punjab Information Technology Board

PLRA – Punjab Land Record Authority

TMB – Telenor Microfinance Bank

NRSP – National Rural Support Programme

NBP – National Bank of Pakistan

ZTBL – Zarai Tariqiyati Bank Limited

Comparative Perspectives to Inform Digital Leadership Qualities for SMEs in Developing Countries

Richard Adrian Tjorko
UCL Centre for Systems Engineering, University College London, UK
richard.tjokro.22@alumni.ucl.ac.uk

Chekfoung Tan
UCL Centre for Systems Engineering, University College London, UK
chekfoung.tan@ucl.ac.uk

Completed Research

Abstract

As technological advancements, including emerging technologies like artificial intelligence, reshape business landscapes, SMEs face unique challenges in adapting to volatility, uncertainty, complexity, and ambiguity (VUCA). Digital transformation (DT) has become a critical strategy for sustaining competitiveness, particularly for SMEs, which play a significant role in driving economic growth and innovation. While research has focused on digital-native and large organisations across both developed and developing regions, significant opportunities remain to further explore the leadership attributes essential for DT within SMEs This study aims to identify key leadership qualities necessary for navigating DT complexities through comparative analysis across different contexts. Using an inductive qualitative approach that combines an extensive literature review with survey-based empirical insights, this paper proposes a digital leadership framework that contributes to the existing knowledge base. In addition, this framework aligns with the unique needs of SMEs in developing contexts, supporting their resilience, adaptability, and potential for growth.

Keywords: Digital Leadership, Digital Transformation, Leadership Qualities, Small and Medium-sized Enterprises, Developing Countries, Emerging Technologies

1.0 Introduction

Technological advancements and their implications have been fundamentally reshaping businesses (Benitez et al., 2022). Organisations are increasingly required to enhance their creativity, flexibility, and resilience to address the growing volatility, uncertainty, complexity, and ambiguity (VUCA) in the business landscape (Santarsiero et al., 2019; Schiuma, 2012). To thrive, they must adapt their strategies and behaviours, transforming challenges into opportunities for growth (Schiuma et al., 2022).

In this context, digital transformation (DT) has emerged as a crucial factor for sustaining organisational competitiveness, particularly for small and medium-sized enterprises (SMEs) (Fachrunnisa et al., 2020; Kokot et al., 2023; Li et al., 2016; Scuotto et al., 2021). DT is defined as the strategic adoption of digital technologies, such as mobile, big data, cloud computing, Internet of Things (IoT), and emerging technologies like artificial intelligence (AI), to drive business model innovation and create new revenue-generating and value-creating opportunities (Malodia et al., 2023; Parida et al., 2019; Verhoef et al., 2021). DT involves profound changes across an organisation's operations, products, processes, and business models, necessitating a shift in organisational culture, leadership, mindsets, attitudes towards risk, and adaptability to continuous change (Kane et al., 2015; Kraus et al., 2022). AI, in particular, enhances leadership effectiveness by providing data-driven insights, optimising decision-making, and automating routine tasks (Al-Bayed et al., 2024).

SMEs represent a significant pillar of most economies, accounting for approximately 90% of businesses and over half of global employment (Faye & Goldbulm, 2022). Their contributions extend to socio-economic goals such as fostering economic growth, job creation, and innovation, particularly in developing countries. Given their importance, governments and stakeholders place substantial emphasis on supporting SME growth.

For SMEs, DT offers significant opportunities to develop high-value products and services, enhance existing offerings, and expand market reach while improving operational efficiency (Li et al., 2016). However, prior studies indicate that SMEs in developing countries have been slow to adopt DT, which hinders their survival and growth (Hai et al., 2021; Malodia et al., 2023; OECD, 2021). The success of DT is frequently attributed to strong leadership, especially within SMEs (AlNuaimi et al., 2022; Fachrunnisa et al., 2020; Li et al., 2016; Promsri, 2019). Leaders play a pivotal role in driving DT initiatives by recognising digitalisation as essential to business activities and cultivating a digital mindset that aligns IT with business strategy (Sia et al., 2016). As DT introduces substantial organisational change, leaders must develop new capabilities that enable adaptability, innovation, and resilience. However, leadership attributes may vary based on organisational size, economic environment, and leadership levels. Understanding how leadership qualities differ across these contexts is crucial to informing a digital leadership framework tailored to SMEs in developing countries.

Despite extensive research on DT in developing countries, further studies can still be undertaken with a particular focus on leadership within the SME context, given its significance in fostering economic growth. Previous research has explored the conceptualisation and analysis of digital leadership, as well as the role of digital leaders in the digital economy (Avolio et al., 2000; El Sawy et al., 2016; Li et al., 2016), often concentrating on digital-native enterprises and large organisations in developed economies (Belitski & Liversage, 2019; Malodia et al., 2023). However, as highlighted by Erhan et al. (2022), there remains an opportunity to explore the specific leadership qualities that contribute to the development of a digital leadership framework. Therefore, this paper seeks to address the research question: *What digital leadership qualities are essential for SME leaders in developing countries?*

This paper is structured as follows: Section 2 reviews the literature on leadership in the digital era, while Section 3 outlines the research methodology. Section 4 presents the results, including comparative insights across different contexts, followed by a discussion in Section 5 on how these findings inform the digital leadership framework. Finally, Section 6 concludes with research implications, limitations, and recommendations for future research.

2.0 Leadership in the Digital Era

2.1 Digital Leadership

The concept of digital leadership has emerged, blending conventional leadership with digital competencies (De Waal et al., 2016; Kane et al., 2019; Schiuma et al., 2022), while incorporating essential qualities such as agility, creativity, and stakeholder collaboration to address the VUCA business landscape (Kazim, 2019). Digital leaders manage transformation by employing diverse leadership styles, including transformational, servant, and transactional (Sow & Aborbie, 2018). They are instrumental in developing digital strategies aligned with organisational goals (Can, 2021), promoting cultural change, securing stakeholder buy-in, and fostering enthusiasm (Benitez et al., 2022; Hinings et al., 2018). Ko et al. (2022) suggest that digital leaders' commitment fosters a cohesive environment supportive of transformation, while their focus on talent development enhances employees' digital knowledge (Vial, 2019). Digital leaders also act as change facilitators, addressing resistance and managing transformation-related tensions, which helps organisations navigate turbulent environments (Benitez et al., 2022; Leso et al., 2023). Essential attributes for digital leaders include digital vision, the ability to envision and communicate a digital future, and digital knowledge, or an understanding of technology’s business impact (Cortellazzo et al., 2019; Imran et al., 2020).

Other key traits are agility, empowerment, and the capacity to “fail fast,” learning quickly from setbacks to redirect efforts productively. Empowerment involves creating a supportive environment for employee growth, and managing diverse teams is vital as leaders must integrate expertise across business and IT domains. Klus and Müller (2021) identify core digital leadership traits such as predicting the future, motivating others, and digital proficiency, while agility enables leaders to respond rapidly to new challenges (Erhan et al., 2022; Fachrunnisa et al., 2020). Thus, digital leadership blends traditional qualities with new digital capabilities, positioning leaders to effectively guide organisations through transformation. This approach is particularly suited to the current landscape, where anticipating and adapting to technological advancements is crucial for success.

2.2 Transformational Leadership

Transformational leadership is widely examined in the context of digital transformation (AlNuaimi et al., 2022; Karippur & Balaramachandran, 2022; Schiuma et al., 2022). This leadership style enhances followers' performance and personal growth by encouraging them to exceed expectations through four key dimensions: charisma, inspiration, intellectual stimulation, and individualised consideration (Northouse, 2021). Charisma establishes leaders as role models, while inspiration secures followers' commitment through clear communication of the organisational vision. Intellectual stimulation promotes innovative thinking by encouraging followers to approach problems from multiple perspectives. Individualised consideration involves empowering followers with opportunities for growth and socio-emotional support.

Transformational leaders foster a culture of innovation and open dialogue by reshaping followers' beliefs, attitudes, and behaviours (Karippur & Balaramachandran, 2022; Northouse, 2021). This approach cultivates an experimental mindset and collaborative environment, both essential for digital transformation.

In the context of digital transformation, Sow and Aborbie (2018) find that transformational leadership drives more favourable outcomes than other styles. Supporting studies indicate that this leadership approach strengthens an organisation's innovation capabilities (Ardi et al., 2020; Lei et al., 2020), fosters creativity (AlNuaimi et al., 2022), promotes e-business adoption (Alos-Simo et al., 2017), and enhances agility (Lin, 2011; Veiseh & Eghbali, 2014; Wanasida et al., 2020). In light of these findings, his study argues that transformational leadership has a positive impact on digital transformation.

2.3 Servant Leadership

While transformational leadership focuses on organisational goals, servant leadership prioritises followers' well-being, viewing it as an end in itself. Servant leadership is a moral approach that provides both tangible and emotional support, creating an environment where employees can reach their full potential, thereby helping the organisation achieve its goals (Jin et al., 2022; Liden et al., 2014). Servant leaders trust their followers to act in the organisation's best interests (Van Dierendonck, 2011), prioritising collective needs, showing empathy, and supporting personal and professional growth (Northouse, 2021). Van Dierendonck (2011) identified ten attributes of servant leadership, including listening, empathy, healing, awareness, and stewardship. Mittal and Dorfman (2012) added humility, authenticity, and interpersonal acceptance. Collins (2009) suggests that humility is a critical factor for long-term organisational success.

Research indicates that servant leadership positively impacts employees' engagement, with commitment and empowerment as mediating factors (Jin et al., 2022; Larjovuori et al., 2016). In the context of digital transformation, servant leadership enhances individual creativity and innovation by fostering a service-oriented culture, psychological empowerment, and job autonomy (Jin et al., 2022; Liden et al., 2014). Additionally, servant leadership reduces stress and enhances well-being during demanding processes like digital transformation (Jin et al., 2022). Thus, this study argues that organisations demonstrating higher levels of servant leadership are better positioned for successful digital transformation.

2.4 Inclusive Leadership

Literature indicates that inclusive leadership has become a popular approach among digital leaders to address the challenges of a diverse organisational landscape, a common scenario in digital transformation (Bourke, 2016). As Northouse (2021) argues, to remain competitive, firms must proactively foster inclusive environments that value and embrace differences. Such environments allow individuals to contribute based on their unique abilities, fostering motivation and leading to optimal performance (Cox & Blake, 1991).

Inclusive leadership is defined as the behaviour of leaders that promotes both belonging and individuality, encouraging active employee participation in group processes (Simmons & Yawson, 2022). This leadership style leverages diverse knowledge, perspectives, and skills to promote organisational learning and growth (Northouse, 2021). Inclusive leaders value diverse viewpoints, appreciate contributions, and are accessible and available to their teams (Ye et al., 2019).

Research consistently highlights the benefits of inclusive leadership. Inclusive environments allow full utilisation of talent, align focus on shared goals, and lead to stronger group performance (Dixon-Fyle et al., 2020). Northouse (2021) notes that inclusion enhances innovation, enabling individuals to share ideas freely. In digital transformation contexts, diverse teams have demonstrated success in creative problem-solving (Tidd & Bessant, 2020), improved decision-making, and meeting the needs of varied stakeholders (Mosher et al., 2017; Ye et al., 2019). A study from Deloitte (2013) similarly found that employees who perceive their organisations as inclusive are more likely to develop innovative solutions, meet customer needs, and collaborate effectively towards common goals.

3.0 Research Methodology

This study employed an inductive qualitative approach, as outlined by Saunders et al. (2019), to explore the emerging phenomenon of leadership in digital transformation and was conducted in two stages. First, an extensive literature review was undertaken to theoretically examine the qualities and attributes associated with digital leadership, inspired by the four key leadership styles outlined in Section 2 (see Table 1). This review served as the foundation for designing the survey used to collect primary data in the second stage.

The survey consisted of a mix of open- and closed-ended questions. Measurement employed a five-point Likert scale, ranging from “strongly disagree” (1) to “strongly agree” (5). A total of 32 questions comprised the questionnaire, with 30 being closed-ended and two open-ended, each accompanied by definitions or examples to ensure clarity and consistency in participants' understanding. The survey was developed as a self-administered, internet-mediated questionnaire, providing participants the flexibility to respond at their convenience and enhancing control over data collection quality. Departmental research ethics approval was obtained prior to data collection.

Qualities	Attributes	Sources
Agility	Agile culture	(Abbu et al., 2022; Belitski & Liversage, 2019; Eller et al., 2020; Erhan et al., 2022; Kane et al., 2019; Karippur & Balaramachandran, 2022; Larjovuori et al., 2018; Li et al., 2016; Wrede et al., 2020)
	Agile strategy	
	Proactiveness	
	Adaptive and flexible	
Digital literacy	Digital skills	(Abbu et al., 2022; Eller et al., 2020; González-Varona et al., 2020; Kane et al., 2019; Kokot et al., 2023; Malodia et al., 2023; Schiuma et al., 2022; Wrede et al., 2020)
	Digital knowledge	
	Digital attitude	
Digital visionary	Clear digital vision and strategy	(Abbu et al., 2022; Eller et al., 2020; Kane et al., 2019; Karippur & Balaramachandran, 2022; Larjovuori et al., 2018; Li et al., 2016; Schiuma et al., 2022; Weber et al., 2022; Wrede et al., 2020)
	Digitalisation as a strategic imperative	
	Data-driven	
Digital entrepreneurship	Multi competent	(Abbu et al., 2022; G. Kane et al., 2019; Karippur & Balaramachandran, 2022; Larjovuori et al., 2018; Li et al., 2016)
	Creative and disruptive	
	Growth mindset	
	Risks taking	
Foster innovation	Cultivate innovative culture	(Karippur & Balaramachandran, 2022; Larjovuori et al., 2018; Weber et al., 2022)
	Encourage innovative thinking	
	Provide support and resources	
	Appreciate achievements	
Empowerment	Encourage to do more	(Kane et al., 2019; Larjovuori et al., 2018; Schiuma et al., 2022; Weber et al., 2022; Wrede et al., 2020)
	Coaching	
	Empathy	
Inspire and motivate	Effective communication	(Larjovuori et al., 2018; Schiuma et al., 2022; Wrede et al., 2020)
	Enthusiast	
	Role modelling	
Collaboration and partnership	Strategic partnership	(Eller et al., 2020; Karippur & Balaramachandran, 2022; Larjovuori et al., 2018; Li et al., 2016; Schiuma et al., 2022; Wrede et al., 2020)
	Cultivate sharing culture	
Foster inclusivity	Value diversity	(Abbu et al., 2022; Larjovuori et al., 2018; Schiuma et al., 2022; Weber et al., 2022; Wrede et al., 2020)
	Promote participation	
Other intrapersonal qualities	Trustworthy	(Abbu et al., 2022)
	Commitment	
	Humility	

Table 1. Digital Leadership Qualities and Attributes

A pilot survey was conducted with five individuals to confirm the survey's effectiveness and the reliability of the data, following Hardy and Ford (2014). Two rounds of adjustments were made until the researcher was confident that respondents faced no difficulties in understanding or responding to the questions. For sample recruitment, a combined approach of snowball sampling and the maximum variation technique from Palinkas et al. (2015) was utilised. This non-probability sampling method aimed to reach rare or hard-to-find populations.

The study's sample included individuals from companies of various sizes, industries, and countries, with varying levels of experience in digital-related projects. This diverse inclusion aimed to provide a comprehensive, multifaceted, and unbiased perspective on leadership. First, incorporating respondents from different organisational backgrounds and experience levels allowed for a broader understanding of leadership dynamics. Employees at different levels (e.g., executives versus staff) were expected to offer distinct perspectives on the traits and qualities they consider essential in a leader.

Second, involving firms of varying sizes facilitated comparative analysis, enabling the identification of best leadership practices or attributes from larger, well-resourced organisations that could be adapted to benefit SMEs. As noted by Hyvönen (2018), examining a phenomenon from multiple perspectives enhances research robustness and validity.

For data analysis, all survey attributes were ranked by average scores to identify those with higher consensus and those considered less significant by respondents. Responses were also compared across different contexts, including job level and industry, to identify any supporting or conflicting perspectives. This comparative analysis ensures the applicability of digital leadership qualities and attributes for SMEs across diverse industries. Results were tabulated, and visual representations generated in Microsoft Excel supported the analysis. Content analysis, as outlined by Weber (1990), was employed to examine open-ended responses, which later informed the development of the digital leadership framework.

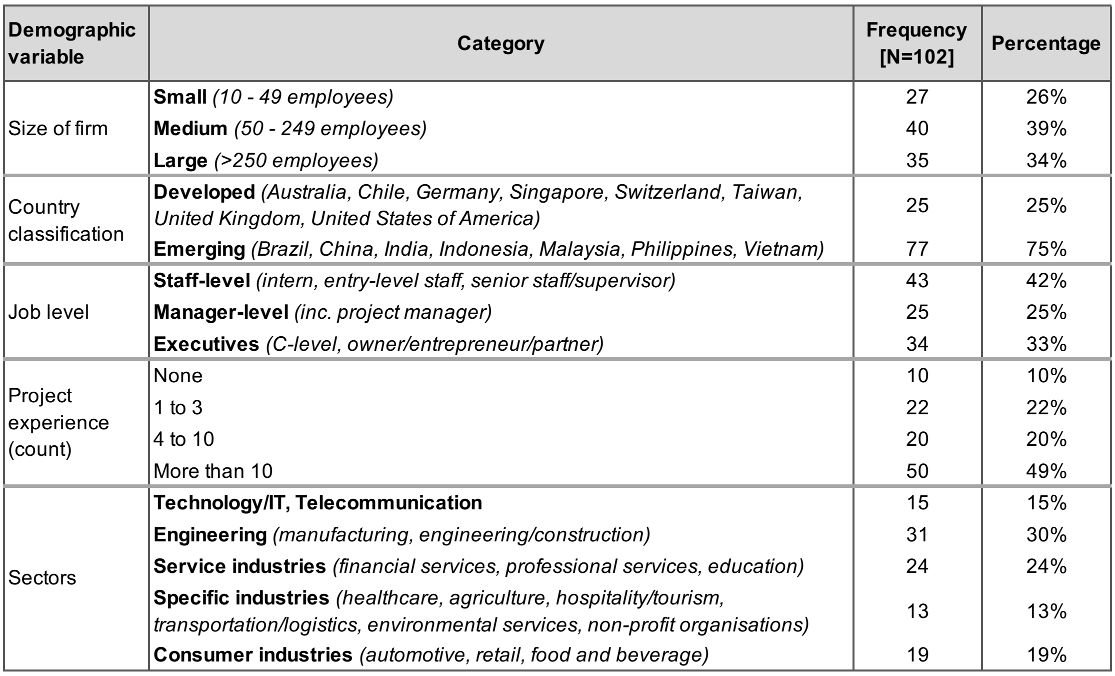
4.0 Results

4.1 Demographics

A total of 102 survey responses were collected and analysed. The demographic characteristics covered several key dimensions, including firm size, country classification, job level, project experience, and industry sectors, as shown in Figure 1.

4.2 Digital Leadership Qualities

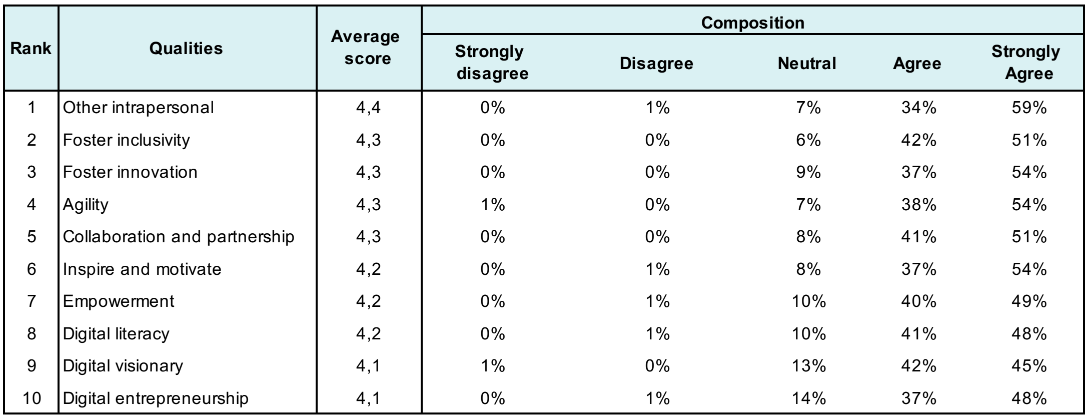
The survey results indicate a high level of agreement across all digital leadership qualities, with each average score exceeding 4 on a 5-point scale (see Figure 2). The top three leadership qualities considered essential by respondents are *other intrapersonal*, *foster inclusivity*, and *foster innovation*. These qualities received high average scores of approximately 4.3, indicating strong agreement from over 50% of respondents, with minimal disagreement. In contrast, attributes related to digital skills, such as *digital literacy*, *digital visionary*, and *digital entrepreneurship*, received relatively lower emphasis.



Demographic variable	Category	Frequency [N=102]	Percentage
Size of firm	**Small** *(10 - 49 employees)*	27	26%
	Medium *(50 - 249 employees)*	40	39%
	Large *(>250 employees)*	35	34%
Country classification	**Developed** *(Australia, Chile, Germany, Singapore, Switzerland, Taiwan, United Kingdom, United States of America)*	25	25%
	Emerging *(Brazil, China, India, Indonesia, Malaysia, Philippines, Vietnam)*	77	75%
Job level	**Staff-level** *(intern, entry-level staff, senior staff/supervisor)*	43	42%
	Manager-level *(inc. project manager)*	25	25%
	Executives *(C-level, owner/entrepreneur/partner)*	34	33%
Project experience (count)	None	10	10%
	1 to 3	22	22%
	4 to 10	20	20%
	More than 10	50	49%
Sectors	**Technology/IT, Telecommunication**	15	15%
	Engineering *(manufacturing, engineering/construction)*	31	30%
	Service industries *(financial services, professional services, education)*	24	24%
	Specific industries *(healthcare, agriculture, hospitality/tourism, transportation/logistics, environmental services, non-profit organisations)*	13	13%
	Consumer industries *(automotive, retail, food and beverage)*	19	19%

Figure 1. Demographics of Survey Respondents

Although these qualities still scored relatively high (around 4.1), fewer than half of the respondents expressed strong agreement. The results also show a notable proportion of respondents who were neutral, suggesting that personality and interpersonal traits of leaders may hold greater importance than specific digital competencies. Each attribute is ranked in descending order by average score in Figure 3.



Rank	Qualities	Average score	Composition				
			Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
1	Other intrapersonal	4,4	0%	1%	7%	34%	59%
2	Foster inclusivity	4,3	0%	0%	6%	42%	51%
3	Foster innovation	4,3	0%	0%	9%	37%	54%
4	Agility	4,3	1%	0%	7%	38%	54%
5	Collaboration and partnership	4,3	0%	0%	8%	41%	51%
6	Inspire and motivate	4,2	0%	1%	8%	37%	54%
7	Empowerment	4,2	0%	1%	10%	40%	49%
8	Digital literacy	4,2	0%	1%	10%	41%	48%
9	Digital visionary	4,1	1%	0%	13%	42%	45%
10	Digital entrepreneurship	4,1	0%	1%	14%	37%	48%

Figure 2. **Overall Rating of Digital Leadership Qualities**

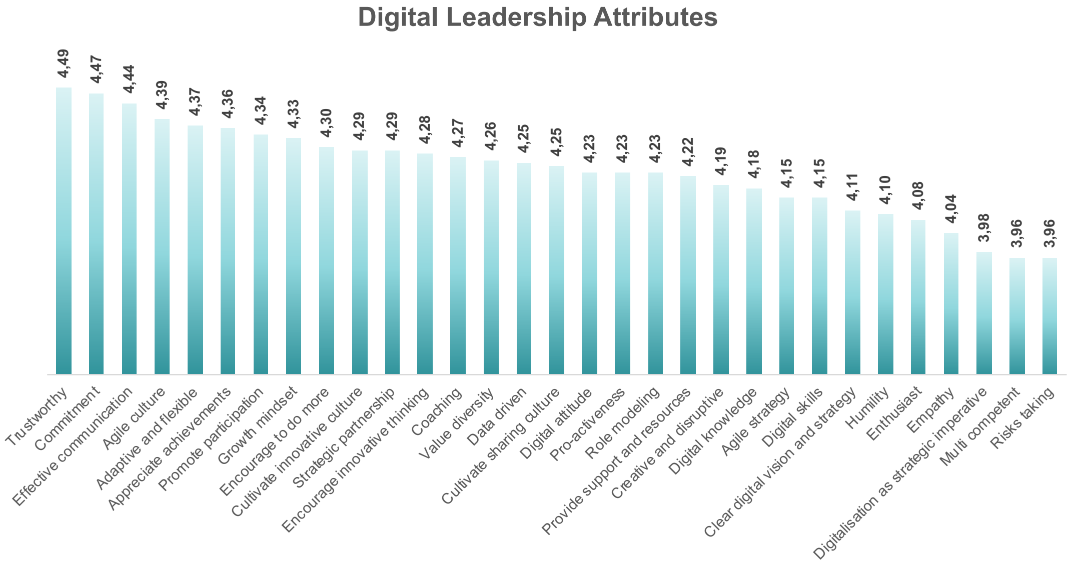


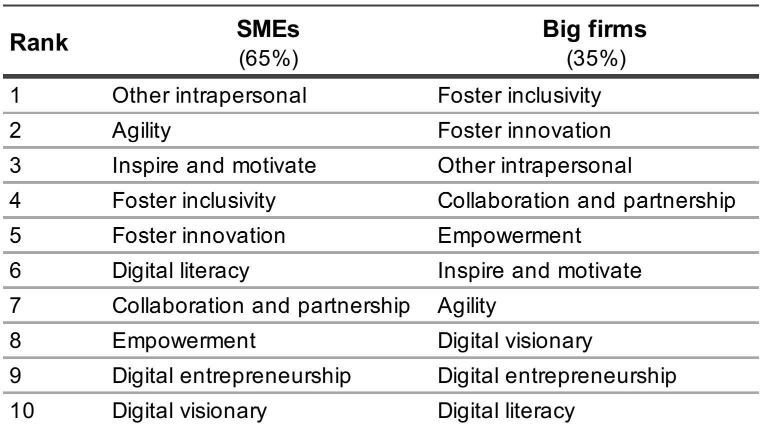
Figure 3. **Ranking of Digital Leadership Attributes**

Other intrapersonal is perceived as the most important quality of digital leadership, with an average score of 4.4. A significant majority (59%) strongly agreed on the importance of this quality, while only about 1% disagreed. Among the attributes contributing to this quality, *trustworthy* was particularly valued, with 62% of respondents strongly agreeing and 28% agreeing. It ranked first out of 31 attributes overall, underscoring the importance of leaders who establish and maintain trust with their teams. Ranked as the second most important attribute, *commitment* scored an average of 4.5, with 56% of respondents strongly agreeing on its significance. In contrast, *humility* emerged as the least important attribute in this group, ranking 6th from the bottom overall, as shown in Figure 3.

4.3 Comparisons of Leadership Qualities Across Different Contexts

4.3.1 Comparison Between SMEs and Big Firms

The ranking of leadership qualities as perceived by SMEs and large firms is presented in Figure 4, with percentages representing the distribution of responses. Overall, the results reveal a distinct pattern in the qualities valued by each group. Both groups emphasise the importance of leaders with a positive personality (i.e., *other intrapersonal* qualities). For SMEs, *agility* is among the most valued leadership qualities, while respondents from big firms view *innovation* as more essential. Additionally, the qualities of *inspiring and motivating* emerge as key priorities for SME leaders. The concept of diversity (*fostering inclusivity)* is highly valued by respondents in large firms. However, this quality is regarded as less important by SME leaders. Furthermore, the quality of *empowerment* is rated relatively low in importance for SMEs, as respondents expressed doubts about its relevance in supporting digital transformation.

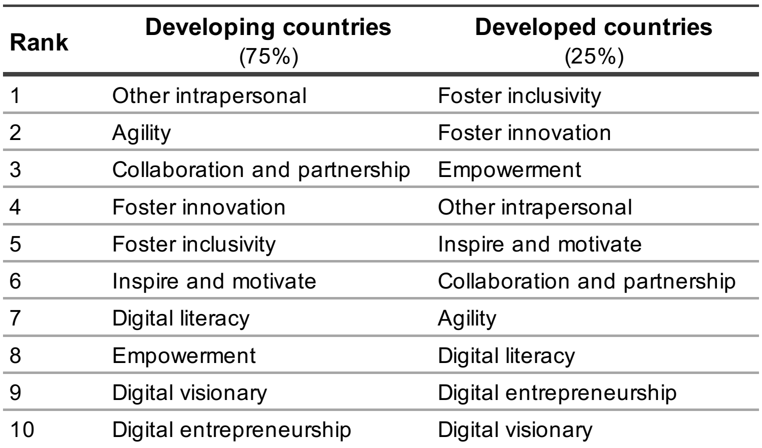


Rank	SMEs (65%)	Big firms (35%)
1	Other intrapersonal	Foster inclusivity
2	Agility	Foster innovation
3	Inspire and motivate	Other intrapersonal
4	Foster inclusivity	Collaboration and partnership
5	Foster innovation	Empowerment
6	Digital literacy	Inspire and motivate
7	Collaboration and partnership	Agility
8	Empowerment	Digital visionary
9	Digital entrepreneurship	Digital entrepreneurship
10	Digital visionary	Digital literacy

Figure 4. Comparison of Leadership Qualities Between SMEs and Big Firms

4.3.2 Comparison between Developing and Developed Countries

Regarding countries, notable variations in perceptions of leadership qualities exist between developed and developing nations. One prominent difference is the importance of *inclusivity*, which is highly valued by respondents in developed countries (ranked 1st) but ranked lower in developing countries (ranked 5th), as shown in Figure 5. In developing countries, *other intrapersonal* qualities are seen as the most important for digital leaders, followed by *agility* and *collaboration and partnership*. In contrast, respondents from developed countries place higher importance on *fostering innovation* and *empowerment*. Additionally, both groups share a less favourable perspective on digital-related attributes, such as *digital literacy, digital visionary,* and *digital entrepreneurship.*

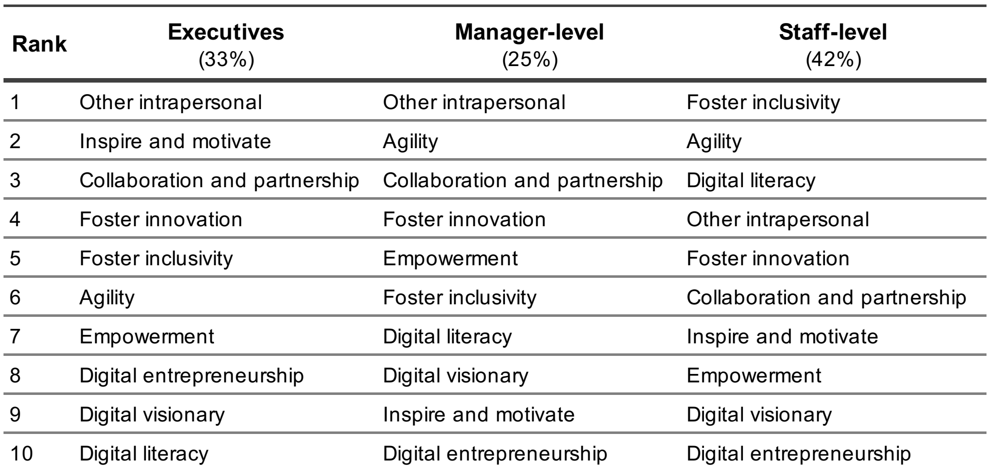


Rank	**Developing countries** (75%)	**Developed countries** (25%)
1	Other intrapersonal	Foster inclusivity
2	Agility	Foster innovation
3	Collaboration and partnership	Empowerment
4	Foster innovation	Other intrapersonal
5	Foster inclusivity	Inspire and motivate
6	Inspire and motivate	Collaboration and partnership
7	Digital literacy	Agility
8	Empowerment	Digital literacy
9	Digital visionary	Digital entrepreneurship
10	Digital entrepreneurship	Digital visionary

Figure 5. Comparison of Leadership Qualities Between Developing and Developed Countries

4.3.3 Comparison Among Different Seniority Levels

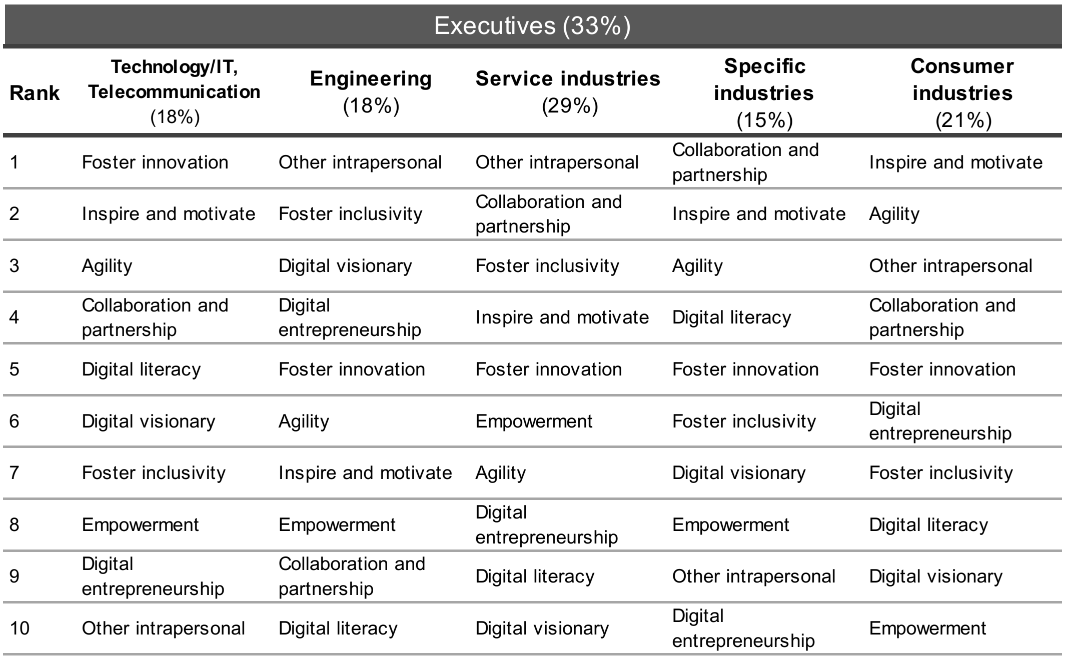
The Executives category, comprising C-level leaders and entrepreneurs/owners/partners, represents 33% of survey respondents. As shown in Figure 6, *other intrapersonal*, *inspire and motivate*, and *collaboration and partnership* emerge as the top three qualities deemed important for digital leaders in SMEs. Although *other intrapersonal* ranks highest overall, its perceived importance varies across industries. Specifically, this quality is highly valued by leaders in the engineering, service, and consumer industries but less so in the Technology/IT & Communication and Specific Industries sectors (see Figure 7). Both *inspire and motivate* and *collaboration and partnership* are considered important across four industries, with the exception of the Engineering sector. This suggests that engineering leaders may prioritise individual expertise and technological proficiency over interpersonal qualities.



Rank	**Executives** (33%)	**Manager-level** (25%)	**Staff-level** (42%)
1	Other intrapersonal	Other intrapersonal	Foster inclusivity
2	Inspire and motivate	Agility	Agility
3	Collaboration and partnership	Collaboration and partnership	Digital literacy
4	Foster innovation	Foster innovation	Other intrapersonal
5	Foster inclusivity	Empowerment	Foster innovation
6	Agility	Foster inclusivity	Collaboration and partnership
7	Empowerment	Digital literacy	Inspire and motivate
8	Digital entrepreneurship	Digital visionary	Empowerment
9	Digital visionary	Inspire and motivate	Digital visionary
10	Digital literacy	Digital entrepreneurship	Digital entrepreneurship

Figure 6. Comparison of Leadership Qualities Among Different Seniority Levels

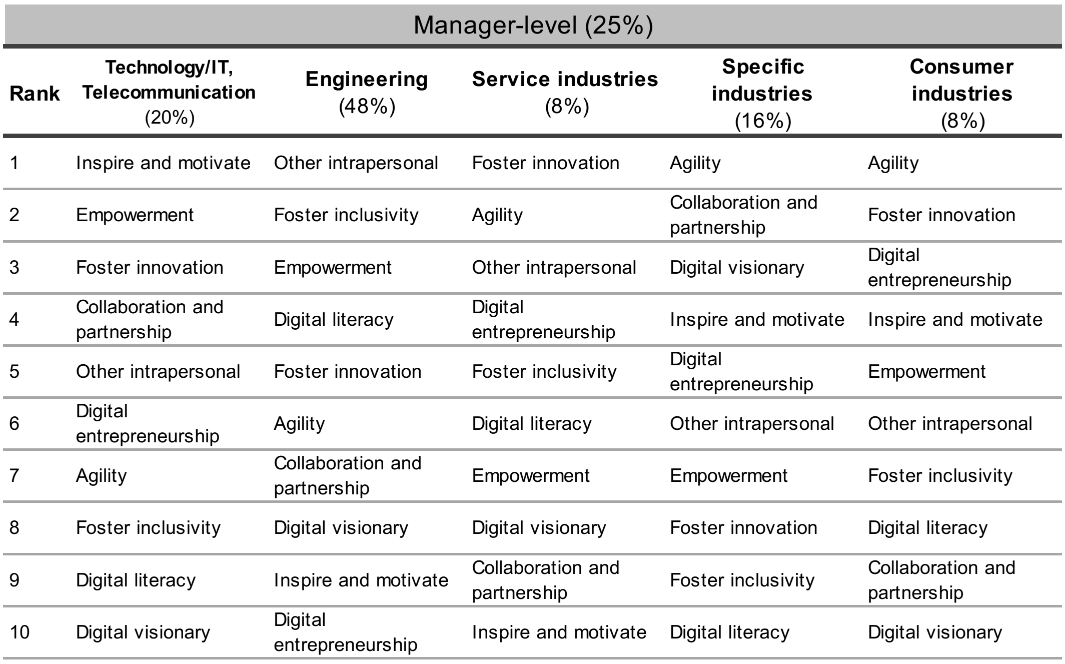
In the fast-paced Technology/IT & Telecommunication sector, the quality of *foster innovation* ranks as the most important. Additionally, the results reveal increased awareness of *foster inclusivity* among leaders, particularly in the Engineering and Service industries. A noteworthy observation is that executives across all sectors assign relatively lower importance to the quality of *empowerment* (see Figure 7).



Executives (33%)					
Rank	**Technology/IT, Telecommunication** (18%)	**Engineering** (18%)	**Service industries** (29%)	**Specific industries** (15%)	**Consumer industries** (21%)
1	Foster innovation	Other intrapersonal	Other intrapersonal	Collaboration and partnership	Inspire and motivate
2	Inspire and motivate	Foster inclusivity	Collaboration and partnership	Inspire and motivate	Agility
3	Agility	Digital visionary	Foster inclusivity	Agility	Other intrapersonal
4	Collaboration and partnership	Digital entrepreneurship	Inspire and motivate	Digital literacy	Collaboration and partnership
5	Digital literacy	Foster innovation	Foster innovation	Foster innovation	Foster innovation
6	Digital visionary	Agility	Empowerment	Foster inclusivity	Digital entrepreneurship
7	Foster inclusivity	Inspire and motivate	Agility	Digital visionary	Foster inclusivity
8	Empowerment	Empowerment	Digital entrepreneurship	Empowerment	Digital literacy
9	Digital entrepreneurship	Collaboration and partnership	Digital literacy	Other intrapersonal	Digital visionary
10	Other intrapersonal	Digital literacy	Digital visionary	Digital entrepreneurship	Empowerment

Figure 7. Comparison of Leadership Qualities Among Executives Across Different Industries

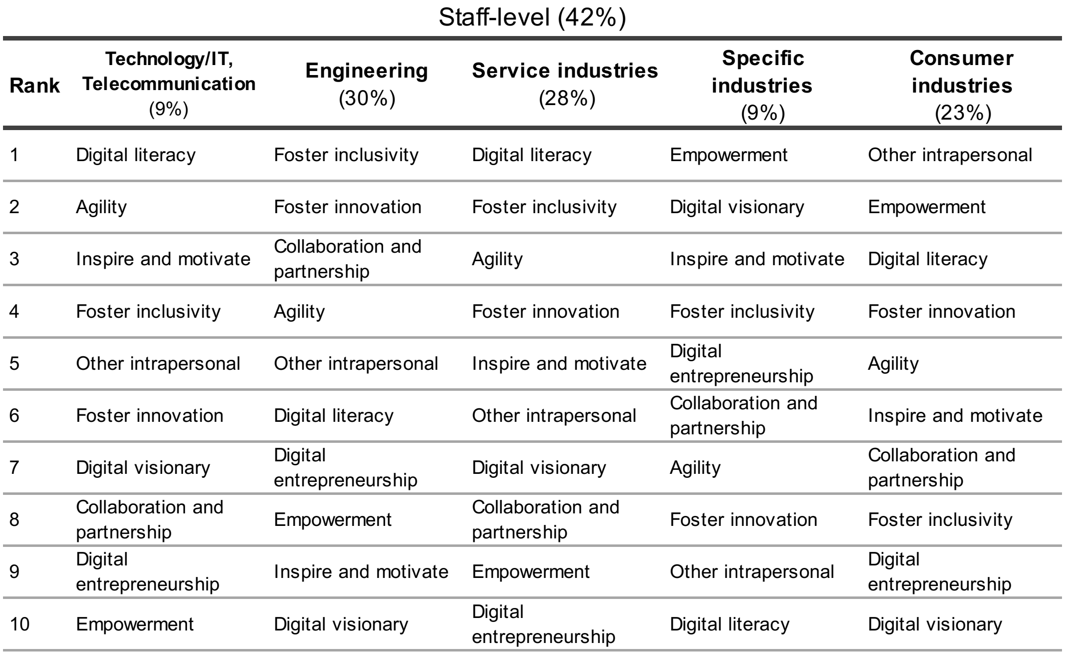
A notable similarity in perceptions is observed between managers and executives (see Figure 6). Specifically, *other intrapersonal*, *collaboration and partnership*, and *foster* *innovation* are ranked highly by both groups. However, managers, slightly differing from executives, consider *agility* as the second most important quality, with this attribute ranking first across three different industries (see Figure 8). Interestingly, managers in the Technology/IT & Communication sector consider agility to be of lesser importance, a finding that contrasts with responses from both staff and executives in the same sector, where it ranks within the top three qualities. The importance of *collaboration and partnership* is also notable. Although this is generally regarded as a key aspect of managerial responsibilities, managers in three industries (Engineering, Service, and Consumer industries) appear to view it as less significant. Additionally, unlike the perspectives of both executives and staff, *inclusivity* has not yet emerged as a prominent theme for managers across industries.



Manager-level (25%)					
Rank	**Technology/IT, Telecommunication** (20%)	**Engineering** (48%)	**Service industries** (8%)	**Specific industries** (16%)	**Consumer industries** (8%)
1	Inspire and motivate	Other intrapersonal	Foster innovation	Agility	Agility
2	Empowerment	Foster inclusivity	Agility	Collaboration and partnership	Foster innovation
3	Foster innovation	Empowerment	Other intrapersonal	Digital visionary	Digital entrepreneurship
4	Collaboration and partnership	Digital literacy	Digital entrepreneurship	Inspire and motivate	Inspire and motivate
5	Other intrapersonal	Foster innovation	Foster inclusivity	Digital entrepreneurship	Empowerment
6	Digital entrepreneurship	Agility	Digital literacy	Other intrapersonal	Other intrapersonal
7	Agility	Collaboration and partnership	Empowerment	Empowerment	Foster inclusivity
8	Foster inclusivity	Digital visionary	Digital visionary	Foster innovation	Digital literacy
9	Digital literacy	Inspire and motivate	Collaboration and partnership	Foster inclusivity	Collaboration and partnership
10	Digital visionary	Digital entrepreneurship	Inspire and motivate	Digital literacy	Digital visionary

Figure 8. Comparison of Leadership Qualities Among Managerial-Level Roles Across Different Industries

Staff-level respondents, including interns, entry-level, and senior staff, represent the majority in this study (42%) and therefore offer perspectives that warrant attention. According to this group, the top three qualities they value in leaders are *fostering inclusivity*, *agility*, and *digital literacy*, as shown in Figure 9. *Inclusivity* is recognised as important across all industries except the Consumer sector. *Agility* ranks second, receiving particular emphasis in the Technology/IT & Communication, Engineering, and Service industries. At the bottom of the rankings are *empowerment*, *digital* *visionary*, and *digital entrepreneurship*. *Empowerment*, in particular, shows notable variation: it ranks highly in the Specific (1st) and Consumer (2nd) industries but is less valued in others.



Staff-level (42%)

Rank	Technology/IT, Telecommunication (9%)	Engineering (30%)	Service industries (28%)	Specific industries (9%)	Consumer industries (23%)
1	Digital literacy	Foster inclusivity	Digital literacy	Empowerment	Other intrapersonal
2	Agility	Foster innovation	Foster inclusivity	Digital visionary	Empowerment
3	Inspire and motivate	Collaboration and partnership	Agility	Inspire and motivate	Digital literacy
4	Foster inclusivity	Agility	Foster innovation	Foster inclusivity	Foster innovation
5	Other intrapersonal	Other intrapersonal	Inspire and motivate	Digital entrepreneurship	Agility
6	Foster innovation	Digital literacy	Other intrapersonal	Collaboration and partnership	Inspire and motivate
7	Digital visionary	Digital entrepreneurship	Digital visionary	Agility	Collaboration and partnership
8	Collaboration and partnership	Empowerment	Collaboration and partnership	Foster innovation	Foster inclusivity
9	Digital entrepreneurship	Inspire and motivate	Empowerment	Other intrapersonal	Digital entrepreneurship
10	Empowerment	Digital visionary	Digital entrepreneurship	Digital literacy	Digital visionary

Figure 9. Comparison of Leadership Qualities Among Staff-Level Roles Across Different Industries

4.4 Additional attributes

The questionnaire included optional open-ended questions, inviting respondents to share their perspectives on leadership characteristics, skills, or behaviours not covered in the survey. Out of 102 participants, 35 provided responses. Content analysis identified ten new attributes: integrity (n=8), respect (n=4), teamwork (n=3), resilience (n=2), listening (n=2), confidence (n=1), self-efficacy (n=1), responsibility (n=1), gratitude (n=1), and charisma (n=1), as shown in Figure 10.

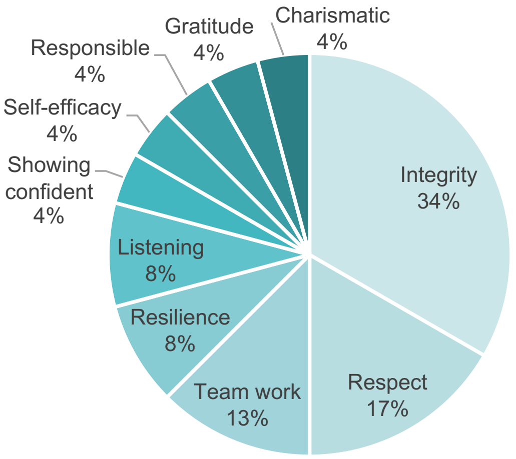


Figure 10. **New Attributes Identified from Open-Ended Responses**

4.5 Summary of Findings

In conclusion, the empirical findings strongly validate the significance of all ten digital leadership qualities. However, the priority of these qualities varies among different contexts. Within the context of this study, which focuses on SMEs in developing countries, Figure 11 presents the overall ranking of these qualities according to responses from participants in developing countries.

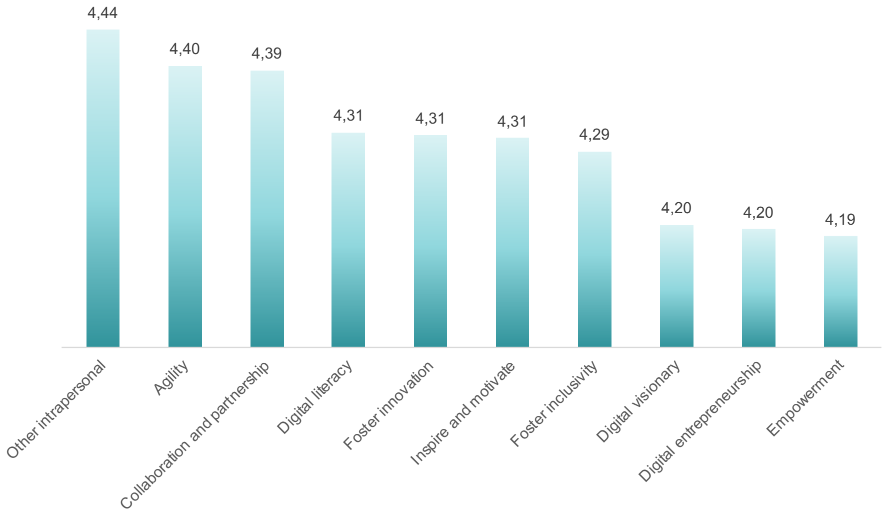


Figure 11. **Ranking of the Digital Leadership Qualities of SMEs in Developing Countries**

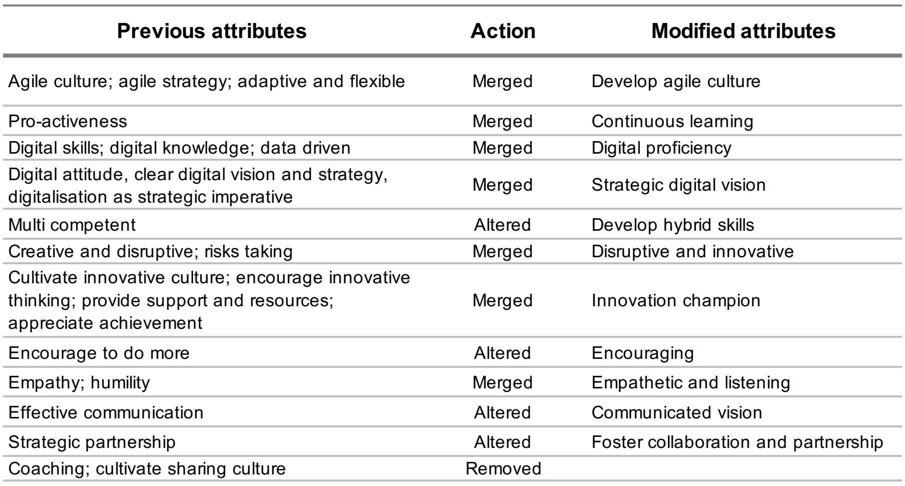
Among digital leadership qualities, *other intrapersonal*, which comprises leaders' personal traits such as *trustworthiness*, *commitment*, and *humility*, emerges as the most universally valued quality. This prominence remains consistent across different contexts, including SMEs, executives, managers, and respondents from developing countries. The quality *foster inclusivity* ranks second, though its prioritisation varies; it is rated highest among staff-level respondents but is considered less important by executives, managers, and SMEs in developing countries. *Foster innovation*, ranked third in importance, is particularly valued by respondents from large corporations and developed countries. In contrast, for SMEs in developing countries, *agility* and *collaboration and partnership* take precedence, ranking second and third, respectively. This highlights the pivotal role these qualities play for SMEs in developing countries as they navigate shifting market dynamics and resource limitations.

Placed sixth overall, *inspire and motivate* is notably significant for executive respondents but is seen as comparatively less important by managers and staff members. Additionally, the quality of *empowerment* is generally lower in priority across all respondent groups, especially within SMEs in developing countries, where it is rated as the least important quality. As a consistent trend, digital-specific qualities such as *digital literacy*, *digital visionary*, and *digital entrepreneurship*, rank lower across most groups. However, this pattern shifts in SMEs from developing countries, where *digital literacy* holds a notable fourth-place position, reflecting its growing importance on par with *foster innovation* and *inspire and motivate*.

5.0 Discussions

Given the new insights derived from the survey results, it became necessary to further refine the digital leadership qualities and attributes outlined in Table 1 to ensure that the digital leadership framework more accurately reflects the findings. Therefore, additional analysis was conducted. As a result, several attributes were added, merged, or reassigned to more suitable categories. For example, attributes such as *digital skills*, *digital knowledge*, and *data-driven* were consolidated into a single attribute, *digital proficiency*, as the literature indicated they are closely related concepts. Additionally, some attributes, such as *coaching* and *cultivating a sharing culture*, were removed due to a lack of perceived importance in the survey results and insufficient supporting evidence from the literature. A summary of this process is presented in Figure 12. Ultimately, this process led to the development of 21 attributes (reduced from 31) and

7 digital leadership qualities (reduced from 10), proposed as the digital leadership framework for SMEs in developing countries (see Figure 13).



Previous attributes	Action	Modified attributes
Agile culture; agile strategy; adaptive and flexible	Merged	Develop agile culture
Pro-activeness	Merged	Continuous learning
Digital skills; digital knowledge; data driven	Merged	Digital proficiency
Digital attitude, clear digital vision and strategy, digitalisation as strategic imperative	Merged	Strategic digital vision
Multi competent	Altered	Develop hybrid skills
Creative and disruptive; risks taking	Merged	Disruptive and innovative
Cultivate innovative culture; encourage innovative thinking; provide support and resources; appreciate achievement	Merged	Innovation champion
Encourage to do more	Altered	Encouraging
Empathy; humility	Merged	Empathetic and listening
Effective communication	Altered	Communicated vision
Strategic partnership	Altered	Foster collaboration and partnership
Coaching; cultivate sharing culture	Removed	

Figure 12. Summary of the Modified Attributes

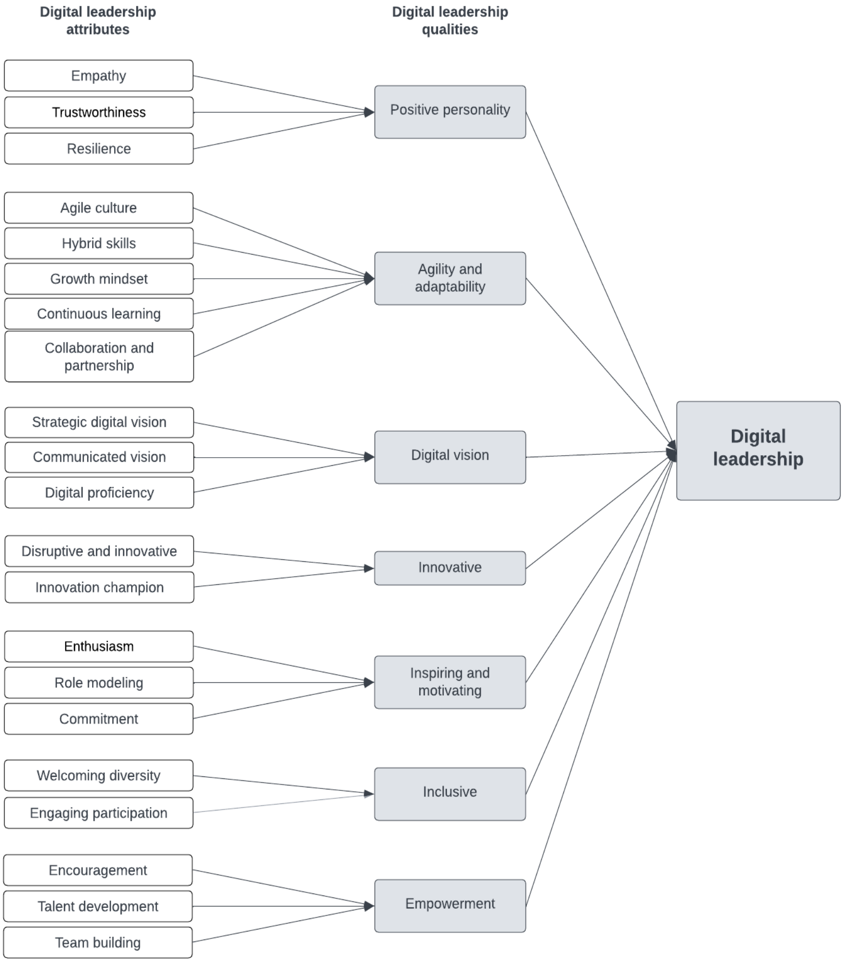


Figure 13. The Digital Leadership Framework

Survey results highlight a *positive personality* as the most valued quality in a digital leader, forming the foundation for effective leadership. Attributes such as *empathy*, *trustworthiness*, and *resilience* shape leaders' interactions, decision-making, and the cultivation of a positive organisational culture.

In SMEs, agility emerges as a crucial organisational quality, linked to a leader’s ability to respond swiftly to opportunities and threats (Li et al., 2016). Adaptability, or a leader’s capacity to adjust to changing circumstances complements agility (Trenerry et al., 2021). Together, these qualities enable leaders to navigate dynamic, ambiguous situations effectively. Attributes within *agility and adaptability* include fostering an *agile culture*, *hybrid skills*, a *growth mindset*, *continuous learning*, and *collaboration and partnership*.

The *digital vision* quality relates to a leader’s ability to envision and communicate a digital future for the organisation. Although survey responses were mixed, literature such as Chen and Chang (2013), Eberl and Drews (2021), and Kane et al. (2019) underscores its significance. This quality involves *strategic digital vision*, *communicated vision*, and *digital proficiency*.

The digital business environment presents challenges across strategies, processes, and operational models (Kane et al., 2018). Respondents, particularly from large firms, favour leaders who act as *innovative champions*, aligning with research that highlights innovation as crucial for organisational survival during disruption (Abbu et al., 2022; Ye et al., 2019). Therefore, SME digital leaders could adopt a similar approach, fostering a *disruptive and innovative* mindset that empowers teams to take risks and develop creative ideas, viewing digital disruption as an opportunity.

Executives identify *inspiring and motivating* as essential for leaders managing digital transformation. Given the substantial changes and uncertainties, employees may see transformation as a threat to their roles (Wrede et al., 2020). Digital leaders need to inspire commitment and instil enthusiasm, as supported by Zoppelletto et al. (2023), fostering employee support through *enthusiasm*, *role modelling*, and *commitment.*

For SMEs, digital transformation calls for an inclusive approach to leadership (Schwarzmüller et al., 2018). Survey findings affirm inclusivity as the top-quality staff desire in leaders. *Inclusive* leaders are open, accessible, and foster engagement by making employees feel valued and empowered to share unique perspectives (Nembhard & Edmondson, 2006). Attributes such as *welcoming diversity* and *engaging participation* help promote creativity and innovation, encouraging diverse viewpoints and experimentation. This leads to higher engagement in innovative activities, aligning with findings by Choi et al. (2017) and Ye et al. (2019).

Empowerment focuses on enabling individuals, nurturing a proactive and confident mindset, and instilling self-belief (Van Dierendonck, 2011). Empowering leaders recognise the unique skills, knowledge, and potential each employee brings to drive creativity and innovation (Laub, 2000). This quality is supported by studies on digital transformation from Frick et al. (2021) and Imran et al. (2020), showing that empowerment fosters readiness and confidence, enabling employees to take independent action and practice self-leadership. Key attributes of empowerment include *encouragement*, *talent development*, and *team building*.

6.0 Conclusion

6.1 Research Implications

This research makes a significant twofold contribution to the theoretical understanding of digital leadership. First, it extends the existing literature by exploring the role of leaders in the digital transformation of SMEs in developing countries. Second, it proposes a digital leadership framework tailored to the specific needs of SMEs in these regions. This framework is grounded in an extensive literature review and supported by empirical data, outlining essential leadership qualities and attributes needed to navigate the complexities of digital transformation effectively. The research outcomes are especially relevant in the current era of AI, where leadership qualities are crucial for effectively harnessing AI technologies to benefit the organisation.

This study also offers valuable practical insights and recommendations. The findings highlight the pivotal role of leaders in SME digital transformation, shaping the transformative vision, setting strategic direction, and establishing the context within which transformation unfolds. While digital transformation is a collective organisational effort, the commitment and actions of leaders are ultimately decisive in its success. Leaders are crucial drivers of structural and cultural changes, creating an environment that encourages active workforce engagement in the transformation process. Moreover, they play a key role in talent development, ensuring that employees are equipped and ready to contribute meaningfully. These insights provide SME leaders with a guidepost, clarifying their multifaceted roles in fostering successful digital transformations.

Building upon the digital leadership framework, it is essential to consider the three proposed principles as fundamental to effective leadership in the digital era:

- **The centrality of positive personality in digital leadership** – A leader's disposition significantly influences leadership effectiveness. Employees often model their behaviour based on their leaders; thus, leaders must exemplify a positive attitude and mindset, including a growth-oriented approach, enthusiasm, and resilience.
- **An employee-centred approach to digital strategy** – The success of digital transformation is contingent upon securing employee engagement and minimising resistance. Leaders must effectively communicate the vision, mission, and strategic intent of the transformation, ensuring that employees understand its significance. Moreover, fostering inclusivity and demonstrating empathy in addressing concerns enhances employee commitment, thereby facilitating a smoother transition.
- **Establishing a supportive organisational culture** – A conducive environment is integral to the transformation process. Leaders should foster a culture that promotes innovation, agility, and collaboration, ensuring that employees feel supported and empowered to adapt to change.

6.2 Limitations and Future Work

This study introduces several limitations. First, while the research focused on SMEs across various developing countries, a significant proportion of respondents (67%) were from Indonesia, with the remaining participants primarily concentrated in the Asian region. This concentration, along with the limited available literature linking leadership with digital transformation in developing countries, may limit the generalisability of the findings to all developing regions, given potential variations in work culture and regional dynamics. Second, the absence of a validation stage, such as follow-up interviews or statistical tests, may raise concerns regarding the reliability and validity of the findings. Lastly, the cross-sectional nature of this study means its context and outcomes are specific to a particular point in time and may not fully account for the continuously evolving digital technologies and business landscape, potentially impacting the framework’s ongoing relevance.

For addressing these limitations, future research could adopt a multi-method approach, combining qualitative and quantitative analysis to validate the framework through statistical measurements. Additionally, an extended analysis could incorporate case studies or follow-up interviews to assess the framework’s applicability and effectiveness across diverse contexts, particularly in exploring the reasons why leadership qualities are perceived differently, as discussed in Section 4.3. Broadening the scope to include data from a more varied range of developing countries or industries could further enhance the framework’s generalisability. Future research could also consider additional variables, such as the firm’s digital maturity level, leaders’ experience, and gender, to deepen understanding of how these factors interact with the framework. Furthermore, a correlation analysis between this framework and the transformation outcomes of SMEs could reveal valuable patterns, particularly in assessing variations across different contexts.

References

Abbu, H., Mugge, P., Gudergan, G., Hoeborn, G., & Kwiatkowski, A. (2022). Measuring the human dimensions of digital leadership for successful digital transformation. *Research-Technology Management*, *65*(3), 39-49.

Al-Bayed, M. H., Hilles, M., Haddad, I., Al-Masawabe, M. M., Alhabbash, M. I., Abu-Nasser, B. S., & Abu-Naser, S. S. (2024). AI in Leadership: Transforming Decision-Making and Strategic Vision.

AlNuaimi, B. K., Singh, S. K., Ren, S., Budhwar, P., & Vorobyev, D. (2022). Mastering digital transformation: The nexus between leadership, agility, and digital strategy. *Journal of Business Research*, *145*, 636-648.

Alos-Simo, L., Verdu-Jover, A. J., & Gomez-Gras, J.-M. (2017). How transformational leadership facilitates e-business adoption. *Industrial Management & Data Systems*, *117*(2), 382-397.

Ardi, A., Djati, S., Bernarto, I., Sudibjo, N., Yulianeu, A., Nanda, H., & Nanda, K. (2020). The relationship between digital transformational leadership styles and knowledge-based empowering interaction for increasing organisational innovativeness. *International Journal of Innovation, Creativity and Change*, *11*(3), 259-277.

Avolio, B. J., Kahai, S., & Dodge, G. E. (2000). E-leadership: Implications for theory, research, and practice. *The leadership quarterly*, *11*(4), 615-668.

Belitski, M., & Liversage, B. (2019). E-Leadership in small and medium-sized enterprises in the developing world. *Technology Innovation Management Review*, *9*(1), 64-74.

Benitez, J., Arenas, A., Castillo, A., & Esteves, J. (2022). Impact of digital leadership capability on innovation performance: The role of platform digitization capability. *Information & management*, *59*(2), 103590.

Bourke, J. (2016). *The six signature traits of inclusive leadership: Thriving in a diverse new world*.

Can, O. (2021). The Role of Leadership in Digital Transformation: A Review and Suggestions for Future Research. ECMLG 2021 17th European Conference on Management, Leadership and Governance,

Chen, Y.-S., & Chang, C.-H. (2013). The determinants of green product development performance: Green dynamic capabilities, green transformational leadership, and green creativity. *Journal of business ethics*, *116*, 107-119.

Choi, S. B., Tran, T. B. H., & Kang, S.-W. (2017). Inclusive leadership and employee well-being: The mediating role of person-job fit. *Journal of Happiness Studies*, *18*, 1877-1901.

Collins, J. (2009). Good to Great-(Why some companies make the leap and others don't). In: SAGE Publications Sage India: New Delhi, India.

Cortellazzo, L., Bruni, E., & Zampieri, R. (2019). The role of leadership in a digitalized world: A review. *Frontiers in psychology*, *10*, 1938.

Cox, T. H., & Blake, S. (1991). Managing cultural diversity: Implications for organizational competitiveness. *Academy of Management Perspectives*, *5*(3), 45-56.

De Waal, B., Van Outvorst, F., & Ravesteyn¹, P. (2016). Digital leadership: The objective-subjective dichotomy of technology revisited. 12 th European Conference on Management, Leadership and Governance ECMLG 2016,

Deloitte. (2013). Waiter, is that inclusion in my soup? A new recipe to improve business performance. In: Deloitte Sydney.

Dixon-Fyle, S., Dolan, K., Hunt, V., & Prince, S. (2020). Diversity wins: How inclusion matters. *McKinsey & Company*, *6*.

Eberl, J. K., & Drews, P. (2021). Digital Leadership–Mountain or molehill? A literature review. *Innovation through information systems: Volume III: A collection of latest research on management issues*, 223-237.

El Sawy, O. A., Kraemmergaard, P., Amsinck, H., & Vinther, A. L. (2016). How LEGO Built the Foundations and Enterprise Capabilities for Digital Leadership [Article]. *Mis Quarterly Executive*, *15*(2), 141-166. <Go to ISI>://WOS:000377109100004

Eller, R., Alford, P., Kallmünzer, A., & Peters, M. (2020). Antecedents, consequences, and challenges of small and medium-sized enterprise digitalization. *Journal of Business Research*, *112*, 119-127.

Erhan, T., Uzunbacak, H. H., & Aydin, E. (2022). From conventional to digital leadership: exploring digitalization of leadership and innovative work behavior [Article]. *Management Research Review*, *45*(11), 1524-1543. https://doi.org/10.1108/mrr-05-2021-0338

Fachrunnisa, O., Adhiatma, A., Lukman, N., & Ab Majid, M. N. (2020). Towards SMEs' digital transformation: The role of agile leadership and strategic flexibility. *Journal of Small Business Strategy*, *30*(3), 65-85.

Faye, I., & Goldbulm, D. (2022). *Quest to better understand the relationship between SME finance and job creation: Insights from new report*. World Bank Blogs. Retrieved 18 April 2023 from https://blogs.worldbank.org/en/psd/quest-better-understand-relationship-between-sme-finance-and-job-creation-insights-new-report

Frick, N. R., Mirbabaie, M., Stieglitz, S., & Salomon, J. (2021). Maneuvering through the stormy seas of digital transformation: the impact of empowering leadership on the AI readiness of enterprises. *Journal of Decision Systems*, *30*(2-3), 235-258.

González-Varona, J. M., Acebes, F., Poza, D., & López-Paredes, A. (2020). Fostering digital growth in SMEs: organizational competence for digital transformation. Boosting Collaborative Networks 4.0: 21st IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2020, Valencia, Spain, November 23–25, 2020, Proceedings 21,

Hai, T. N., Van, Q. N., & Thi Tuyet, M. N. (2021). Digital transformation: Opportunities and challenges for leaders in the emerging countries in response to COVID-19 pandemic. *Emerging Science Journal*, *5*(1), 21-36.

Hardy, B., & Ford, L. R. (2014). It's not me, it's you: Miscomprehension in surveys. *Organizational Research Methods*, *17*(2), 138-162.

Hinings, B., Gegenhuber, T., & Greenwood, R. (2018). Digital innovation and transformation: An institutional perspective. *Information and Organization*, *28*(1), 52-61.

Hyvönen, J. (2018). *Strategic leading of digital transformation in large established companies–a multiple case-study*

Imran, F., Shahzad, K., Butt, A., & Kantola, J. (2020). Leadership competencies for digital transformation: evidence from multiple cases. Advances in Human Factors, Business Management and Leadership: Proceedings of the AHFE 2020 Virtual Conferences on Human Factors, Business Management and Society, and Human Factors in Management and Leadership, July 16-20, 2020, USA,

Jin, S., Li, Y., & Xiao, S. (2022). What drives Employees' innovative behaviors in emerging-market multinationals? An integrated approach. *Frontiers in psychology*, *12*, 803681.

Kane, Phillips, A. N., Copulsky, J., & Andrus, G. (2019). How digital leadership is (n't) different. *MIT Sloan Management Review*, *60*(3), 34-39.

Kane, G., Palmer, D., Phillips, A. N., Kiron, D., & Buckley, N. (2015). Strategy, not technology, drives digital transformation. *MIT Sloan Management Review and Deloitte University Press*, *14*(1-25).

Kane, G., Palmer, D., Phillips, A. N., Kiron, D., & Buckley, N. (2018). *Coming of age digitally: Learning, leadership, and legacy*.

Kane, G., Phillips, A., Nguyen, Copulsky, J., & Andrus, G. (2019). How digital leadership is (n't) different. *MIT Sloan management review*, *60*(3), 34-39.

Karippur, N. K., & Balaramachandran, P. R. (2022). Antecedents of effective digital leadership of enterprises in Asia Pacific. *Australasian Journal of Information Systems*, *26*.

Kazim, F. A. (2019). Digital transformation and leadership style: a multiple case study. *The ISM journal of international business*, *3*(1), 24-33.

Klus, M. F., & Müller, J. (2021). The digital leader: what one needs to master today's organisational challenges. *Journal of Business Economics*, *91*(8), 1189-1223.

Ko, A., Fehér, P., Kovacs, T., Mitev, A., & Szabó, Z. (2022). Influencing factors of digital transformation: management or IT is the driving force? *International Journal of Innovation Science*, *14*(1), 1-20.

Kokot, K., Kokotec, I. Đ., & Čalopa, M. K. (2023). Digital Leadership and Maturity as a Key to Successful Digital Transformation: Country Case Study of Croatia. *TEM Journal*, *12*(1).

Kraus, S., Durst, S., Ferreira, J. J., Veiga, P., Kailer, N., & Weinmann, A. (2022). Digital transformation in business and management research: An overview of the current status quo. *International Journal of Information Management*, *63*, 102466.

Larjovuori, R.-L., Bordi, L., & Heikkilä-Tammi, K. (2018). Leadership in the digital business transformation. Proceedings of the 22nd international academic mindtrek conference,

Larjovuori, R. L., Bordi, L., Makiniemi, J. P., & Heikkila-Tammi, K. (2016, Sep 08-10). THE ROLE OF LEADERSHIP AND EMPLOYEE WELL-BEING IN ORGANIZATIONAL DIGITALIZATION. [What's ahead in service research?: New perspectives for business and society]. 26th Annual Conference of the European-Association-for-Research-on-Services (RESAR), Naples, ITALY.

Laub, J. A. (2000). Assessing the servant organization: Development of the organizational. *PhD diss., Florida Atlantic University Boca Raton, Florida*.

Lei, H., Leaungkhamma, L., & Le, P. B. (2020). How transformational leadership facilitates innovation capability: the mediating role of employees' psychological capital. *Leadership & Organization Development Journal*, *41*(4), 481-499.

Leso, B. H., Cortimiglia, M. N., & Ghezzi, A. (2023). The contribution of organizational culture, structure, and leadership factors in the digital transformation of SMEs: a mixed-methods approach. *Cognition, Technology & Work*, *25*(1), 151-179.

Li, W., Liu, K., Belitski, M., Ghobadian, A., & O'Regan, N. (2016). e-Leadership through strategic alignment: An empirical study of small- and medium-sized enterprises in the digital age. *Journal of Information Technology*. https://doi.org/10.1057/jit.2016.10

Liden, R. C., Wayne, S. J., Liao, C., & Meuser, J. D. (2014). Servant leadership and serving culture: Influence on individual and unit performance. *Academy of management journal*, *57*(5), 1434-1452.

Lin, W.-B. (2011). Factors affecting the effects of service recovery from an integrated point of view. *Total Quality Management*, *22*(4), 443-459.

Malodia, S., Mishra, M., Fait, M., Papa, A., & Dezi, L. (2023). To digit or to head? Designing digital transformation journey of SMEs among digital self-efficacy and professional leadership. *Journal of Business Research*, *157*, 113547.

Mittal, R., & Dorfman, P. W. (2012). Servant leadership across cultures. *Journal of World Business*, *47*(4), 555-570.

Mosher, D. K., Hook, J. N., Captari, L. E., Davis, D. E., DeBlaere, C., & Owen, J. (2017). Cultural humility: A therapeutic framework for engaging diverse clients. *Practice Innovations*, *2*(4), 221.

Nembhard, I. M., & Edmondson, A. C. (2006). Making it safe: The effects of leader inclusiveness and professional status on psychological safety and improvement efforts in health care teams. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *27*(7), 941-966.

Northouse, P. G. (2021). *Leadership : theory and practice / Peter G. Northouse* (Ninth edition. ed.). SAGE.

OECD. (2021). *The Digital Transformation of SMEs*.

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Administration and policy in mental health and mental health services research*, *42*(5), 533-544. https://doi.org/10.1007/s10488-013-0528-y

Parida, V., Sjödin, D., & Reim, W. (2019). Reviewing literature on digitalization, business model innovation, and sustainable industry: Past achievements and future promises. In (Vol. 11, pp. 391): MDPI.

Promsri, C. (2019). The developing model of digital leadership for a successful digital transformation. *GPH-International Journal of Business Management*, *2*(08), 01-08.

Santarsiero, F., Carlucci, D., & Schiuma, G. (2019). Understanding the phenomenon of innovation labs. Annual Gsom Emerging Markets Conference 2019,

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students*. Pearson Education, Limited.

Schiuma, G. (2012). Managing knowledge for business performance improvement. *Journal of Knowledge Management*, *16*(4), 515-522.

Schiuma, G., Schettini, E., Santarsiero, F., & Carlucci, D. (2022). The transformative leadership compass: six competencies for digital transformation entrepreneurship. *International Journal of Entrepreneurial Behavior & Research*, *28*(5), 1273-1291.

Schwarzmüller, T., Brosi, P., Duman, D., & Welpe, I. M. (2018). How does the digital transformation affect organizations? Key themes of change in work design and leadership. *Management Revue*, *29*(2), 114-138.

Scuotto, V., Nicotra, M., Del Giudice, M., Krueger, N., & Gregori, G. L. (2021). A microfoundational perspective on SMEs' growth in the digital transformation era. *Journal of Business Research*, *129*, 382-392.

Sia, S. K., Soh, C., & Weill, P. (2016). How DBS bank pursued a digital business strategy. *MIS Quarterly Executive*, *15*(2).

Simmons, S. V., & Yawson, R. M. (2022). Developing leaders for disruptive change: An inclusive leadership approach. *Advances in Developing Human Resources*, *24*(4), 242-262.

Sow, M., & Aborbie, S. (2018). Impact of leadership on digital transformation. *Business and Economic Research*, *8*(3), 139-148.

Tidd, J., & Bessant, J. R. (2020). *Managing innovation: integrating technological, market and organizational change*. John Wiley & Sons.

Trenerry, B., Chng, S., Wang, Y., Suhaila, Z. S., Lim, S. S., Lu, H. Y., & Oh, P. H. (2021). Preparing workplaces for digital transformation: An integrative review and framework of multi-level factors. *Frontiers in psychology*, *12*, 620766.

Van Dierendonck, D. (2011). Servant leadership: A review and synthesis. *Journal of management*, *37*(4), 1228-1261.

Veiseh, S., & Eghbali, N. (2014). A study on ranking the effects of transformational leadership style on organizational agility and mediating role of organizational creativity. *Management Science Letters*, *4*(9), 2121-2128.

Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J. Q., Fabian, N., & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*, *122*, 889-901.

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems*, *28*(2), 118-144. https://doi.org/https://doi.org/10.1016/j.jsis.2019.01.003

Wanasida, A. S., Bernarto, I., & Sudibjo, N. (2020). The effect of millennial transformational leadership on IT capability, organizational agility and organizational performance in the pandemic era: An empirical evidence of fishery startups in Indonesia. Conference Series,

Weber, E., Krehl, E. H., & Büttgen, M. (2022). The digital transformation leadership framework: Conceptual and empirical insights into leadership roles in technology‐driven business environments. *Journal of Leadership Studies*, *16*(1), 6-22.

Weber, R. P. (1990). *Basic content analysis* (Vol. 49). Sage.

Wrede, M., Velamuri, V. K., & Dauth, T. (2020). Top managers in the digital age: Exploring the role and practices of top managers in firms' digital transformation. *Managerial and Decision Economics*, *41*(8), 1549-1567.

Ye, Q., Wang, D., & Guo, W. (2019). Inclusive leadership and team innovation: The role of team voice and performance pressure. *European Management Journal*, *37*(4), 468-480.

Zoppelletto, A., Orlandi, L. B., Zardini, A., Rossignoli, C., & Kraus, S. (2023). Organizational roles in the context of digital transformation: A micro-level perspective. *Journal of Business Research*, *157*, 113563.

Impact of Generative AI on Students Teaching and Learning Success in Higher Education

Dr. Mohammed Albakri
University of Salford
Salford Business School
m.albakri@salford.ac.uk

Eduard Buzila
Otto-von-Guericke University Magdeburg
Eduard.Buzila@ovgu.de

Abstract

Large Language Models (LLMs) have become an integral part of higher education (HE), with millions of students and lecturers interacting with them weekly. Students utilise LLMs for exam preparation and personalised learning, while educators employ them as pedagogical tools to enhance teaching efficacy. This study conducts a systematic literature review (SLR) to evaluate the impact of LLMs on teaching effectiveness, student engagement, and overall learning success (LS). It examines both progressive and regressive factors influencing LLM integration in HE, emphasising the importance of structured and ethical implementation. Findings indicate that when responsibly integrated, LLMs support personalised learning, foster critical thinking, and enhance digital literacy, all of which are key skills for future professionals. However, challenges such as ethical concerns, biases, and over-reliance on AI-generated content pose potential risks to LS. This study contributes to the ongoing discourse on AI-driven education, offering insights into the progressive and regressive issues that influence students' LS.

Keywords: Generative Artificial Intelligence, Large Language Models, Learning Success, Higher Education

1.0 Introduction

The general public's accessibility of LLMs through the release of Chat Generative Pre-trained Transformer 3.5 (ChatGPT-3.5) on 30 November 2022 has heralded a new era in human history, since a new entity has been created, which has expanded the realm of relationships that humans can have. Throughout history, humans could only have a human-human, a human-animal, or a human-thing relationship (Cheng, 2023). As such, humans could only interact with other human beings (homo sociologicus), with wild, farm and/or domestic animals (homo naturalis), and/or with tools and machines (homo faber) (Edwards, 2018).

Through new advancements in Artificial Intelligence (AI), through the use of artificial neuronal networks in machine and deep learning, mimicking the functions of the human brain, humans can now also interact with LLMs. Even though almost two years have passed since the release of ChatGPT-3.5, we still cannot tell what it actually is with regard to the aforementioned trichotomy: human, animal, thing. What we know is that it is not a human being, it is not an animal but it is also not ‘just’ a thing. However, the capabilities of the newest version of LLMs such as ChatGPT (GPT-4.5, introduced by OpenAI in late February 2025, is the newest version) possesses more similarities to a human (Bubeck et al., 2023; Guzman & Lewis, 2020; J. Li et al., 2024) than to an animal or an ‘unalive’ thing, forming a new category of possible human interactions, namely a “human-AI interaction” (Amershi et al., 2019).

Since hundreds of millions of users engage regularly, on a weekly basis, with one or more LLMs, almost all scientific disciplines now focus on this “human-LLM interaction” to understand how LLMs might affect their own scientific community (use of LLMs by researchers) as well as the general public (use of LLMs by laypersons). Hence, one has to specify the group of users within the specific domain where they use LLMs when analysing the human-LLM interaction. For us educators, it is not surprising that students in HE have been using LLMs for different purposes one of which is studying and preparing for exams (Duong et al., 2023).

Therefore, in this paper, we focus on students and on lecturers (group of users) in HE (specific domain), presenting, analysing, and discussing the effects of LLMs on students’ learning success (LS). By LS, we are referring to the achievement of students' educational goals, measured by their academic performance, learning experiences, and development of skills such as critical thinking, creativity, and AI literacy to improve learning processes (Deckler et al, 2024). Hence, we focus on how students and teachers perceive, use, and interact with an LLM from the viewpoint of Information Systems (IS) since “IS is the only discipline with a primary focus to study the applications of technology by organisations and society” (Avison & Elliot 2006:5), focusing more on “interactions between people and organisations (the ‘soft’ issues) and technology than on the technologies (the ‘hard’ issues) themselves” (Avison & Elliot 2006:7). Hence, the key research question (RQ) of our paper is the following: *How might the use of an LLM such as ChatGPT impact the teaching and the learning success (LS) of students in HE?*

To answer this RQ, we have conducted a SLR that allowed us to look at both possible sides of the use of an LLM as a learning tool in HE, meaning we have been able to identify progressive and regressive factors that can be attributed to the LLM and which can either increase or decrease the LS of students, depending on the concrete use of the LLM and on the setup of the educational and learning environment.

The remainder of the paper is structured as follows: Section 2 discusses the related theoretical background and the theoretical frameworks that are needed to understand the implications of the use of LLMs in HE with regard to teaching and LS, Section 3 discusses the results of the conducted SLR on the impact of LLMs on teaching and learning, Section 4 concludes the paper and gives the contributions of the study, and Section 5 provides a recommendation.

2.0 Methodology

This section details the research approach, search strategy, selection criteria, and quality appraisal undertaken to conduct a SLR on the impact of generative artificial intelligence (GenAI) on teaching and learning in HE. The methodology aligns with rigorous SLR protocols to ensure reliability, reproducibility, and transparency, following guidance from established systematic review frameworks (Kitchenham et al., 2009; Watson & Webster, 2020; Webster & Watson, 2002).

2.1 Research Approach

An SLR was conducted to synthesise empirical findings on the implementations and applications of GenAI in HE, particularly in assessing its influence on pedagogy (teachers' perception and use) and on students' LS (students' perception and use). This approach aimed to identify progressive and regressive factors that impact students' LS. By focusing on recent high-quality, peer-reviewed publications, the SRL provides an up-to-date perspective on the role of GenAI in contemporary HE.

2.2 Search Strategy

A comprehensive search strategy was employed to capture relevant literature across multiple databases (see Table 1). The following search string was used to encompass the broad scope of GenAI in HE contexts:

(generative AND artificial AND intelligence OR generative AND ai OR genai OR gai) AND (assessment OR pedagogic OR student OR teaching AND learning OR teaching OR teacher) OR (llms OR language AND model) OR (academia OR education OR he OR higher AND education)

2.3 Selection Criteria

The strategy followed multiple filtering stages across each database. The selection process involved multiple stages of screening based on predetermined eligibility criteria. Inclusion criteria were restricted to empirical studies, peer-reviewed journal articles, and conference papers in English, published between 2014 and 2024. Exclusion criteria included non-empirical studies, book reviews, editorials, and studies not directly addressing GenAI's pedagogic implications or impact on teaching and learning in HE. This resulted in 109,874 potential papers for inclusion after the first round of filtration.

Following these steps, a total of 1,085 papers were retained after further filtrations based on subject area were applied and were then subjected to individual screening based on title, abstract, and keywords. After the papers underwent further eligibility screening based on the title, abstract and keyword scan of each paper, 155 eligible papers were retained and subjected to final checks.

2.4 Selection of Papers for Inclusion

After quality checks were performed on each papers (see section 2.5), 130 studies were identified as meeting the eligibility requirements across the databases. Table 1 summarises the search process and filtering criteria applied.

Database	**Initial Results**	**Filter 1: English, 2014-2024**	**Filter 2: Relevant Subject Areas**	**Filter 3: Title, Abstract, Keyword Scan**	**Final Papers**
Scopus	2,831	1,416	922	130	121
AIS	802	72	51	7	5

ProQuest	4M+	108,155	Sample of 100	18	4
Web of Science	250,185	231	12	0	0
Total	-	109,874	1,085	155	**130**

Table 1. Selected Databases

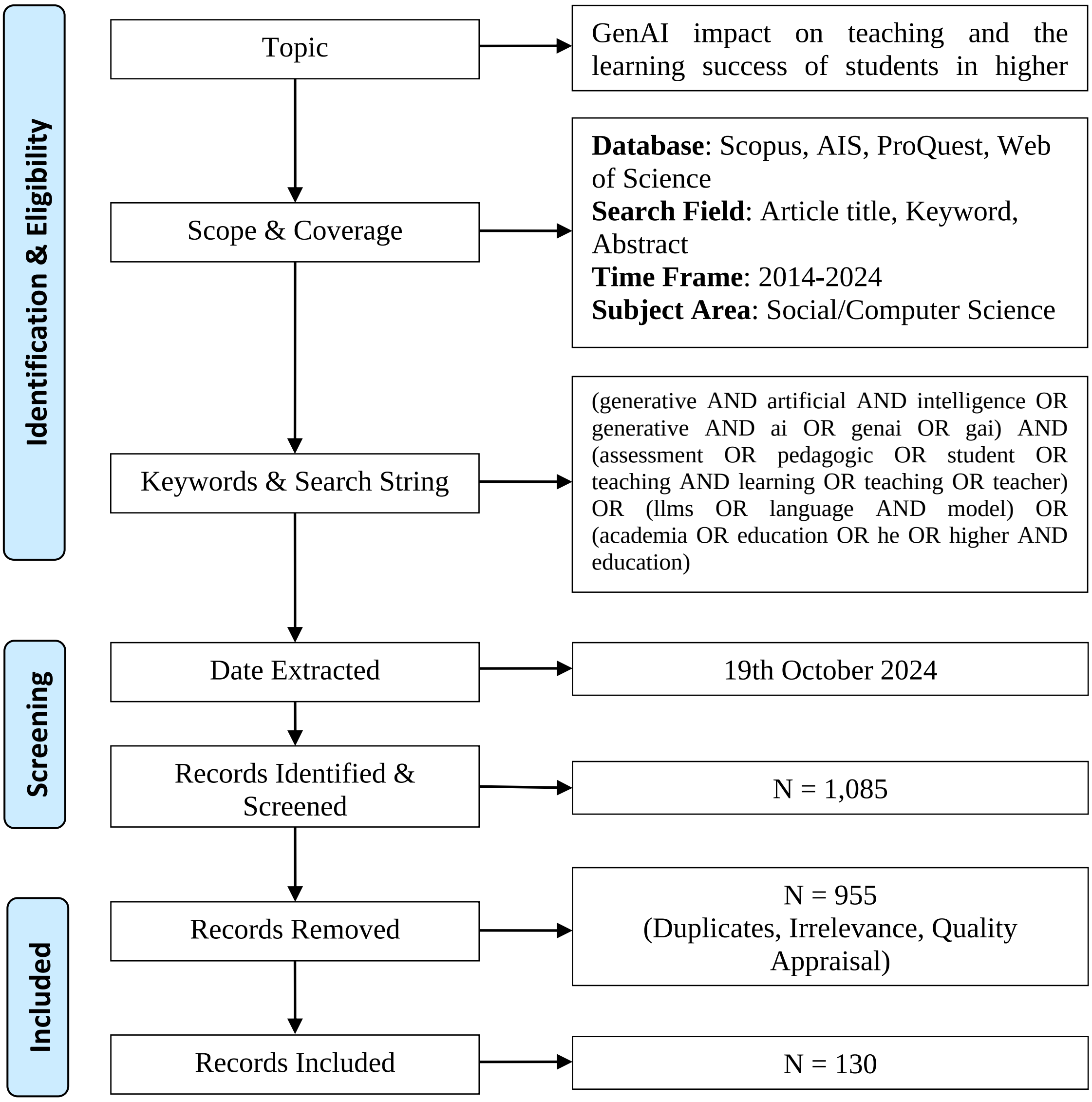

Identification & Eligibility
Screening
Included
Topic
GenAI impact on teaching and the learning success of students in higher
Scope & Coverage
Database: Scopus, AIS, ProQuest, Web of Science
Search Field: Article title, Keyword, Abstract
Time Frame: 2014-2024
Subject Area: Social/Computer Science
Keywords & Search String
(generative AND artificial AND intelligence OR generative AND ai OR genai OR gai) AND (assessment OR pedagogic OR student OR teaching AND learning OR teaching OR teacher) OR (llms OR language AND model) OR (academia OR education OR he OR higher AND education)
Date Extracted
19th October 2024
Records Identified & Screened
N = 1,085
Records Removed
N = 955
(Duplicates, Irrelevance, Quality Appraisal)
Records Included
N = 130

Figure 1. Potential Papers for Inclusion

2.5 Quality Appraisal

To ensure methodological rigour, each paper was subjected to a quality assessment (Mohamed Shaffril et al., 2021; Yang et al., 2021) based on a 4-point scale (X. Li et al., 2024). This scoring system assessed the extent to which each study met key criteria based on the paper theme:

- 0 = Criterion not present: not relevant to the paper theme

- 1 = Slightly present: may allude to the paper theme but only a brief mention
- 2 = Moderately present: has more focus on the paper theme but with some other unrelated content
- 3 = Fully present: fully aligns with the paper theme

Studies were then categorised based on their average percentage score:

Score Range	Quality Category	Description
75%+	High	Strong content throughout
50-74%	Good	Adequate content throughout, but with minor limitations
25-49%	Moderate	Acceptable content throughout, but some significant weaknesses
<24%	Poor	Insufficient content throughout

Table 2. Quality Criteria

Moderate, good and high scoring papers were accepted, while poor scoring papers were omitted. After the quality appraisal process, 130 studies met the quality threshold and were included in the final synthesis. Table 3 provides a breakdown of the quality appraisal outcomes.

Database	Screened Papers	Final High-Quality Papers
Scopus	130	121
AIS	7	5
ProQuest	18	4
Web of Science	0	0
Total	**155**	**130**

Table 3. Quality Screening

Each paper's quality was assessed independently by both authors, and any discrepancies in scores were discussed and reconciled to ensure consistency and reliability in the final selection (see Appendix). This dual-author validation process (Mohammed & Ozdamli, 2024) reinforced the robustness of the selected studies, ensuring a credible foundation for synthesising findings on the educational implications of GenAI.

3.0 Results

3.1 Key Findings

This section examines the multidimensional impact of GenAI on teaching and LS within HE, segmented into essential themes that encompass ethical considerations, skill development, critical engagement, inclusivity, and assessment practices. Each theme and sub-theme reflects the complex interplay of opportunities and challenges presented by GenAI, reinforcing the necessity for structured, ethical integration. These themes were identified through a systematic review of existing literature, focusing on the key dimensions shaping the impact of GenAI in higher education. The analysis categorised findings based on recurring patterns in the papers' scholarly discussions, aligning with critical areas such as ethics, skill development, engagement, inclusivity, and assessment. This thematic structure ensures a comprehensive understanding of both the opportunities and challenges posed by GenAI, facilitating a balanced approach to its integration.

3.2 Ethical and Responsible Use of GenAI

3.2.1 Data Privacy and Security Paradoxes

Data privacy represents a profound ethical concern in GenAI adoption across educational landscapes. As Aad and Hardey (2024) identify, the allure of GenAI in personalising learning is juxtaposed with the risks of data misuse, particularly in systems without robust privacy protocols. The current landscape requires an evolved approach to data security, one that integrates ethical safeguards as core components of AI deployment rather than reactive responses to breaches.

Baek et al. (2024) further assert that GenAI adoption must transcend basic compliance, demanding proactive institutional responsibility to secure data against unauthorised access. The call for regulatory frameworks is reinforced by Ballantine et al. (2024), who highlight the indispensable role of data policies, particularly as educational institutions pivot towards cloud-based GenAI solutions. Collectively, these perspectives underscore that safeguarding data privacy is fundamental to sustainable GenAI integration, ensuring that student trust and institutional integrity remain uncompromised (Chen et al., 2024; Ellis & Slade, 2023).

3.2.2 Accountability and Transparency of Ethical GenAI Applications

The concept of accountability extends beyond functional transparency, emerging as a foundational requisite for legitimacy within educational GenAI applications.

Adarkwah (2024) and Ajevski et al. (2023) articulate that students must understand AI operations, especially in high-stakes areas like assessments, where GenAI's opacity could otherwise create a trust deficit. Transparent GenAI, they argue, enables students and educators alike to engage meaningfully with AI-driven insights, cultivating a sense of shared responsibility and ethical engagement.

Further underscoring this need for transparency, Hostetter et al. (2024) advocate for explainable AI (XAI) models that articulate underlying processes of GenAI-generated feedback. Jensen et al. (2024) propose that transparency policies should be embedded at the policy level, thereby institutionalising accountability as a critical component of GenAI operations. Together, these insights suggest that transparency is not incidental to educational success but central to establishing trust, enabling both students and educators to interact confidently with AI-enhanced educational tools (Jochim & Lenz-Kesekamp, 2024; Kumar et al., 2024; Lee et al., 2024).

3.2.3 Human-Computer Interaction Balance and Collaborative AI-Educator Models

The interaction between human judgement and AI functionalities in educational settings is pivotal to GenAI's responsible deployment. Akpan et al. (2024) argue that GenAI should serve as a complement to human oversight rather than a replacement, preserving the relational dynamics integral to education. Hostetter et al. (2024) and Sullivan et al. (2023) concur, highlighting the educator's role in contextualising GenAI outputs, particularly in disciplines demanding nuanced interpretation.

Sharples (2023) advocates for a collaborative framework where GenAI enhances rather than overshadows the educator's role, fostering a co-active model that balances AI's capabilities with human insight. This collaborative model is further supported by Baek et al. (2024), who emphasise that human-centred AI designs preserve the integrity of teacher-student interactions, transforming GenAI into a support tool rather than an autonomous instructor. This approach underscores that a balanced, integrative GenAI model enhances educational quality and promotes ethical AI engagement (Van Wyk, 2024; 2024; Yeralan & Lee, 2023).

3.2.4 Ethical Dilemmas in AI Decision-Making

The ethical dilemmas surrounding AI decision-making in higher education are becoming increasingly complex, particularly as generative AI tools like ChatGPT are integrated into teaching and learning environments. One of the primary concerns is the potential for AI systems to perpetuate biases in decision-making, as these systems often rely on historical data that may reflect societal prejudices (Chen et al., 2024). In academic settings, this could lead to biased grading systems, unfair evaluations of student work, or the amplification of existing inequalities in access to resources and support (Ajevski et al., 2023). AI's influence on decision-making in admissions, assessments, and resource allocation could inadvertently disadvantage marginalised student groups, thereby challenging the ethical fairness of these processes. Additionally, the transparency and accountability of AI algorithms remain problematic, as many institutions struggle to explain the rationale behind AI-generated decisions, further complicating the trust between students, educators, and the institution (Hostetter et al., 2024).

Moreover, AI in higher education raises concerns about student autonomy and critical thinking. With the growing reliance on AI for research assistance, content creation, and even automated feedback, there is a risk that students may become overly dependent on AI tools, diminishing their capacity for independent learning and reflective thought (Iliyasu et al., 2024). While AI can provide valuable guidance, it cannot replace the nuanced understanding and creativity that human educators offer, which is crucial for fostering deep learning. This reliance may lead to ethical issues around the authenticity of student work and the erosion of academic integrity (Adarkwah, 2024). As such, ethical considerations must guide the implementation of AI technologies in education, ensuring that these tools augment rather than replace essential human elements of teaching and learning. Institutions need to develop frameworks for the ethical use of AI, including clear guidelines and safeguards against bias, while promoting critical AI literacy among students and faculty (Guettala et al., 2024).

3.2.5 Bias, Fairness and Equitable GenAI Frameworks

The conversation on GenAI's potential biases reveals the inherent risks of deploying LLMs that could perpetuate existing societal disparities. Adel et al. (2024) caution that unexamined biases in GenAI systems may marginalise underrepresented groups, highlighting the importance of fairness-oriented AI development. Asad et al. (2024) further posit that fairness metrics and bias assessments are essential for GenAI applications in education, advocating for policies that pre-emptively address discriminatory AI tendencies.

Guettala et al. (2024) reinforce the need for rigorous, fairness-centric GenAI development, suggesting that educational institutions prioritise equity as an ethical imperative. By embedding bias mitigation practices within GenAI, educational environments can ensure that all students, irrespective of background, benefit equally from AI-supported learning environments (Escalante et al., 2023; Pack & Maloney, 2023; Raman et al., 2024).

3.3 GenAI for Skill and Workforce Preparation

3.3.1 Developing Foundational Digital and AI Literacy Skills for Modern Careers

The rapid evolution of AI-integrated workplaces necessitates a foundational literacy in digital and AI competencies. Adarkwah (2024) and Ajevski et al. (2023) argue that embedding GenAI within educational curricula equips students with the skills required for success in tech-driven fields, bridging the digital gap across industries. This perspective aligns with Chen et al. (2024), who posit that AI literacy extends beyond operational knowledge to encompass a strategic understanding of AI’s ethical implications within various sectors.

Alasadi and Baiz (2023) advocate for integrating AI and digital literacy as core educational components, viewing GenAI as an essential tool for fostering workplace readiness. This initiative, as Akpan et al. (2024) underscore, positions GenAI not merely as an educational tool but as a preparatory instrument, building student confidence in AI competencies critical for modern workplaces (Ballantine et al., 2024; Cacho, 2024).

3.3.2 Fostering Students’ Critical and Analytical Skills for Data-Driven Decision-Making

GenAI fosters a robust analytical mindset, encouraging students to scrutinise and interpret data critically. Baek et al. (2024) illustrate how GenAI challenges students to evaluate AI outputs, nurturing a critical approach essential in data-reliant industries.

Ellis and Slade (2023) extend this, highlighting that GenAI instils a verification-based learning culture, prompting students to examine information from multiple perspectives.

Iatrellis et al. (2024) note that GenAI facilitates real-world problem-solving simulations, allowing students to develop skills directly transferable to professional environments. This intersection of theoretical and practical knowledge positions GenAI as a catalyst for analytical skill-building, an indispensable asset in data-driven career paths (Omar et al., 2024; Pack & Maloney, 2023).

3.3.3 Promoting Adaptability, Lifelong Learning and Essential Skills for a Dynamic Workforce

In a rapidly evolving career landscape, adaptability and continuous learning are paramount. Adel et al. (2024) propose that GenAI exposure encourages students to remain agile, continuously enhancing their skills to meet shifting industry demands. Hostetter et al. (2024) reinforce that GenAI's evolving nature supports lifelong learning, instilling resilience and adaptability.

The value of adaptability, as articulated by Guettala et al. (2024), is further demonstrated through GenAI's role in promoting iterative improvement, positioning AI as a model for ongoing skill development. Such adaptability is crucial, as Sharples (2023) suggests, for preparing students to navigate the uncertainty and innovation inherent in AI-driven fields (Wood & Moss, 2024; Yeralan & Lee, 2023).

3.3.4 Enhancing Domain-Specific Competencies and Tailoring GenAI Applications to Disciplinary Needs

GenAI's versatility allows it to address discipline-specific skills, enhancing student proficiency in targeted fields. Akpan et al. (2024) discuss GenAI's applicability in engineering, healthcare, and business, where customised simulations provide hands-on experience. Adarkwah (2024) illustrates GenAI's potential for customisation, which allows students to gain field-relevant expertise and prepares them for sector-specific roles.

In healthcare, Iliyasu et al. (2024) examine GenAI's role in simulating patient interactions, honing medical students' diagnostic abilities. Similarly, Lee et al. (2024) identify GenAI's contributions to legal education, enhancing practical skills through case analysis simulations. These applications underscore GenAI's value in providing sector-specific training that aligns closely with industry requirements (Benuyenah & Dewnarain, 2024; Escalante et al., 2023).

3.3.5 Over-reliance on AI for Faculty Decision-Making

The increasing integration of Generative AI in higher education, while enhancing educational practices, raises concerns about over-reliance on AI for faculty decision-making, particularly in ways that may unintentionally impact student LS. AI tools, such as ChatGPT, offer faculty significant assistance in grading, content generation, and personalised student feedback, streamlining administrative tasks and potentially enhancing the learning environment (Guettala et al., 2024; Ballantine, Boyce, & Stoner, 2024). However, when faculty members depend too heavily on AI-generated recommendations or evaluations, there is a risk that the nuanced understanding of individual students' needs may be overshadowed by the limitations of AI models. These models are not yet capable of fully grasping the complexities of human learning, nor can they adequately interpret emotional and socio-cultural factors that significantly influence student success (Imran et al., 2024). Over-reliance on AI may result in less personalised, human-centric educational experiences, undermining the ability of educators to provide the tailored support that students require to thrive academically.

Moreover, the use of AI for decision-making could inadvertently foster a disengagement from the relational and reflective aspects of teaching, which are critical for promoting student growth and development (Chen et al., 2024). While AI can assist in objective tasks such as grading and content delivery, it is unable to replace the nuanced, empathetic decision-making that is essential to fostering deep learning experiences. Studies show that students benefit significantly from faculty members' emotional intelligence and their ability to adapt teaching strategies based on real-time feedback and contextual understanding (Akpan et al., 2024; Jho & Ha, 2024). Excessive reliance on AI for decision-making risks diminishing these essential human interactions, which may ultimately affect students' academic success, motivation, and engagement in the learning process.

3.3.6 Faculty AI Integration Challenges

The integration of Generative AI into higher education poses significant challenges for faculty members, which, in turn, can impact students' teaching and LS. Many educators face difficulties in effectively adopting AI technologies due to limited technical skills, a lack of sufficient training, and concerns over the reliability and accuracy of AI-generated content (Guettala et al., 2024; Ballantine, Boyce, & Stoner, 2024). This resistance or slow adoption can result in missed opportunities to enhance student learning experiences through personalised, data-driven instructional strategies. Without proper AI integration, students may not benefit from the adaptive learning tools, automated feedback, and tailored content that can significantly improve their engagement, motivation, and overall academic performance (Akpan et al., 2024). Faculty members' apprehensions about the risks of over-reliance on AI or its potential to undermine critical thinking may limit the implementation of innovative pedagogical approaches that can support student success.

Moreover, the lack of standardised guidelines and support systems for integrating AI effectively across various disciplines can lead to inconsistent usage, further hindering its impact on student learning (Imran et al., 2024). As faculty members struggle to balance traditional teaching methods with new AI-driven practices, students may experience uneven educational experiences that hinder their academic growth. Additionally, faculty may lack the resources to assess the full implications of AI tools on student learning outcomes, which can result in a disconnect between AI applications and the actual needs of students (Jho & Ha, 2024). Therefore, without adequate support structures and faculty development programs, the challenges in AI integration can undermine the intended benefits of these technologies, reducing their potential to enhance student LS across higher education institutions.

3.4 Enhancing Critical Thinking and Engagement

3.4.1 Promoting GenAI as a Tool for Critical Reflective Inquiry

GenAI encourages students to engage in critical inquiry, prompting them to evaluate AI-generated insights actively. Adel et al. (2024) argue that GenAI's multi-perspective approach enables students to compare and contrast ideas, fostering a critical examination of complex knowledge structures. This analytical engagement aligns with Bloom's higher-order skills, pushing students beyond mere comprehension to rigorous evaluation. Ellis and Slade (2023) support this perspective, observing that GenAI fosters critical engagement by prompting students to question, verify, and synthesise information. This reflective approach, as Akpan et al. (2024) illustrate, promotes a deeper, more thoughtful interaction with AI-generated content, enhancing students' capacity for independent analysis rather than passive acceptance.

Jeon and Lee (2023) further highlight that critical engagement with GenAI encourages students to reflect on the accuracy and relevance of AI outputs, transforming GenAI from a static tool to an interactive intellectual collaborator. Collectively, these studies suggest that GenAI can be strategically leveraged to cultivate a reflective learning environment, thereby fostering critical inquiry as a core academic skill (Chen et al., 2024; Jochim & Lenz-Kesekamp, 2024).

3.4.2 Supporting Active, Sustained Engagement and Transforming Learning into a Participatory Process

GenAI's interactive capabilities significantly enhance student engagement by creating dynamic, hands-on learning experiences. Baek et al. (2024) explore how GenAI applications, including chatbots and AI-driven simulations, make learning more interactive, thus transforming students from passive recipients into active participants. Hostetter et al. (2024) discuss how instant AI-generated feedback sustains student motivation by allowing real-time adjustments in learning, which promotes a continual engagement loop.

Sharples (2023) contends that GenAI's immersive capacities make challenging subjects more accessible and appealing, thereby capturing student interest and encouraging deeper engagement. These findings collectively indicate that GenAI's interactive nature fosters a participatory educational environment, aligning with experiential learning models that emphasise active, sustained involvement as essential for effective knowledge retention and motivation (Lee et al., 2024; Wood & Moss, 2024).

3.4.3 Stimulating Innovation and Exploration in Learning for Creative Problem-Solving

GenAI supports creative problem-solving by presenting students with tools that allow exploration of multiple approaches to complex issues. Adarkwah (2024) observes that GenAI facilitates creative thinking by offering diverse problem-solving strategies, which encourages students to experiment with innovative solutions. Ajevski et al. (2023) argue that such creative latitude fosters an environment where students feel empowered to navigate varied pathways towards solutions, thus enhancing their creative reasoning skills.

Bai et al. (2024) underscore GenAI’s adaptability in presenting open-ended scenarios, which require students to employ both critical and creative thinking rather than converging on a single answer. Guettala et al. (2024) add that AI-driven simulations push students to tackle intricate problems innovatively, reinforcing the importance of creative flexibility. Collectively, these studies argue that GenAI enriches students' problem-solving skills by exposing them to multiple perspectives and methodologies, thereby enhancing their ability to approach complex challenges with creativity and adaptability (Cacho, 2024; Van Wyk, 2024).

3.4.4 Tailoring Educational Pathways to Individual Needs for Personalised Learning Experiences

One of GenAI’s most transformative applications lies in its capacity to personalise learning experiences, adapting content to suit individual needs. Akpan et al. (2024) discuss how GenAI can generate tailored prompts and challenges that align with each student’s proficiency level, thereby promoting critical engagement with content suited to their skill set. Chen et al. (2024) highlight that personalised GenAI feedback allows students to address specific learning gaps, fostering a more reflective, active learning process.

Imran et al. (2024) examine how GenAI's adaptability allows it to dynamically adjust lesson difficulty based on student responses, ensuring that learning remains challenging without being overwhelming. This personalised approach not only maintains student engagement but also supports the development of critical thinking skills within a supportive, customisable framework. Collectively, these studies suggest that personalised GenAI applications provide tailored learning experiences that enhance analytical skills and motivate students to delve deeper into their studies, promoting both comprehension and engagement (Jho & Ha, 2024; Yeralan & Lee, 2023).

3.4.5 Reduced Human-Led Discussions

The increasing reliance on generative AI tools in higher education has led to concerns regarding the reduction of human-led discussions in the learning environment, which could impact student success. As students increasingly interact with AI-powered systems such as ChatGPT, there is a noticeable shift away from traditional instructor-student and peer-to-peer interactions, which have long been essential for collaborative learning and critical thinking. While generative AI can enhance learning through personalised content delivery and adaptive feedback (Guettala et al., 2024), it may also reduce the depth of human engagement in discussions, potentially hindering the development of essential skills such as problem-solving and creative thinking (Jeon & Lee, 2023). In the absence of human moderators and facilitators, students may become more reliant on AI for information and responses, which could limit their ability to engage in complex, nuanced discussions that are often critical for deep learning.

Moreover, the growing integration of generative AI in higher education raises questions about the quality of student learning outcomes. AI systems, while efficient in providing information, cannot replicate the dynamic and rich exchanges that occur in human-led discussions, which are fundamental to fostering an inclusive and interactive learning environment (Adel et al., 2024). Human instructors offer real-time clarifications, adapt to individual student needs, and challenge students with thought-provoking questions; roles that AI systems are not yet fully equipped to perform. While AI may support student learning by offering instant access to resources, its overuse could result in passive learning environments where students may not develop the critical thinking and collaborative skills necessary for academic success (Cacho, 2024). Thus, a balance must be struck between AI use and maintaining robust human-led interactions to ensure that students continue to develop the necessary skills for success in higher education.

3.4.6 AI-Induced Intellectual Passivity

The integration of generative AI in higher education offers promising advancements in teaching and learning; however, it also raises concerns about the potential for AI-induced intellectual passivity among students. The accessibility of AI tools, such as ChatGPT, can facilitate quick solutions and immediate feedback, yet it may inadvertently reduce critical thinking and deep engagement with learning materials. Studies indicate that when students over-rely on AI to generate responses or complete assignments, they may neglect the development of essential cognitive skills like analysis, evaluation, and independent problem-solving (Bai et al., 2024; Akpan et al., 2024). As students delegate more intellectual tasks to AI systems, they risk becoming passive recipients of knowledge rather than active creators, which undermines their long-term academic success and critical thinking abilities.

This trend of intellectual passivity is particularly concerning in the context of academic writing and research, where students are encouraged to think critically, construct arguments, and engage with diverse perspectives. While generative AI tools can support students in drafting content or brainstorming ideas, the excessive use of such technologies may hinder their ability to refine their reasoning or engage in meaningful reflection (Cacho, 2024; Asad et al., 2024). Furthermore, concerns have been raised regarding the ethical implications of students using AI-generated content without proper understanding or attribution, which can lead to issues of academic integrity (Jensen et al., 2024). To mitigate these risks, educators must emphasise the importance of AI as a supplementary tool, encouraging students to maintain an active role in their learning process while developing their intellectual capabilities in a more meaningful and sustainable way.

3.5 Equitable Access and Inclusivity in GenAI Tools

3.5.1 Ensuring Accessibility for Underrepresented Groups and Reducing Barriers to Inclusive Education

A critical consideration in GenAI integration is ensuring accessibility for students from underrepresented groups, including those with disabilities or limited resources. Aad and Hardey (2024) underscore the need for adaptive GenAI features to avoid inadvertently excluding students with varied learning requirements. Adel et al. (2024) argue that institutional guidelines should prioritise accessibility, ensuring that all students, regardless of physical or cognitive needs, can benefit from AI-driven learning.

Chen et al. (2024) advocate for multimodal GenAI capabilities, allowing accommodations such as screen readers and alternative input methods. Hostetter et al. (2024) add that such accessibility enhancements are essential for creating an inclusive GenAI framework, capable of meeting the needs of diverse learners. Collectively, these studies stress that for GenAI to truly serve as an inclusive educational tool, it must incorporate adaptable features that accommodate a broad spectrum of student needs (Escalante et al., 2023; Jeon & Lee, 2023).

3.5.2 Language and Cultural Biases in GenAI

Language and cultural biases within GenAI systems represent significant inclusivity challenges, as several studies highlight. Bai et al.(2024) argue that language biases embedded in GenAI systems can disadvantage non-native speakers and culturally diverse students, potentially limiting their educational outcomes. Asad et al. (2024) support the need for culturally sensitive AI models to ensure GenAI provides a supportive learning experience across global student populations.

Omar et al. (2024) discuss how GenAI can inadvertently perpetuate cultural stereotypes, emphasising the need for diverse, inclusive training datasets. Similarly, Guettala et al. (2024) advocate for AI models that recognise and respect cultural variations, fostering inclusivity. Together, these findings underscore the importance of cultural and linguistic adaptability in GenAI, advocating for AI systems that support and respect students from diverse backgrounds (Jochim & Lenz-Kesekamp, 2024; Van Wyk, 2024).

3.5.3 Digital Equity and Divide in GenAI Access

Digital equity remains a crucial concern in GenAI adoption, particularly regarding the resource disparities among students. Baek et al. (2024) address the reality that students from low-income backgrounds may lack reliable internet or devices to access GenAI tools, widening the educational gap. Adarkwah (2024) highlights the importance of institutional support mechanisms, such as device loans or on-campus AI resources, to bridge these gaps.

Akpan et al. (2024) emphasise that digital equity is essential to ensure that all students can access GenAI, regardless of socioeconomic status. Benuyenah and Dewnarain (2024) further argue for policy interventions to make GenAI accessible across financial divides, enabling equitable educational opportunities. These findings underscore that for GenAI to fulfil its inclusive potential, institutions must proactively address the digital divide, ensuring equal access to AI resources (Mendez, 2024; Sullivan et al., 2023).

3.5.4 Developing Inclusive Design and Usability for Diverse Demographics

Inclusive design principles are essential for GenAI systems to be effectively usable by a diverse student body. Ellis and Slade (2023) highlight the importance of intuitive GenAI interfaces that accommodate users with varying levels of digital proficiency. Sharples (2023) contends that inclusive GenAI must be developed with input from diverse demographic groups, ensuring that the final product addresses a wide range of accessibility needs.

Chen et al. (2024) argue that engaging minority groups and students with disabilities in the design phase ensures that GenAI systems are truly inclusive. Lee et al. (2024) further emphasise that inclusive GenAI design enhances learning experiences, making AI tools accessible and beneficial for all students. Together, these studies advocate for comprehensive inclusivity in GenAI development, promoting accessible, user-friendly systems that accommodate diverse student needs (Jho & Ha, 2024; Yeralan & Lee, 2023).

3.5.5 AI-Driven Discrimination in Academic Settings (Algorithmic-Bias)

AI-driven discrimination, particularly in the form of algorithmic bias, poses significant challenges to students' LS in higher education, especially as generative AI tools like ChatGPT are integrated into educational settings. Algorithmic bias in AI systems can perpetuate stereotypes, marginalise minority groups, and create inequities in student experiences. For instance, biased AI models may provide skewed or inaccurate feedback, disadvantaging students from underrepresented backgrounds or those whose academic work does not conform to the AI's training data. This can influence how students are assessed, their learning trajectories, and ultimately their academic success. Such biases in AI systems may disproportionately affect students in disciplines like law and business, where specific, contextual, and nuanced understanding is critical (Ajevski et al., 2023; Hostetter et al., 2024).

The presence of algorithmic bias in AI systems challenges the notion of AI as an impartial educational tool and necessitates careful consideration of its role in academic settings. When AI tools are not adequately designed to account for cultural, linguistic, and contextual diversity, students may experience diminished learning opportunities and skewed educational outcomes. For example, generative AI systems used to aid student writing may not adequately address the needs of English as a second language (ESL) learners, leading to lower-quality feedback or less accurate assessments (Escalante et al., 2023). This issue is compounded by a lack of transparency in AI models, making it difficult for educators and students to understand the rationale behind AI-generated content. Therefore, it is crucial that higher education institutions adopt frameworks to address AI bias, ensuring that AI tools enhance rather than hinder student success (Adarkwah, 2024; Baek et al., 2024).

3.6 Assessment and Feedback Enhancement through GenAI

3.6.1 Automated Assessment, Grading and Timely Feedback

GenAI's capacity for automated assessment and grading is transformative, particularly in enhancing efficiency and enabling timely feedback. Adel et al. (2024) discuss the use of GenAI for automating grading processes, arguing that it facilitates rapid feedback and standardised assessments. Akpan et al. (2024) highlight that GenAI alleviates the grading burden for educators, allowing them to focus on more complex pedagogical tasks.

Jeon and Lee (2023) argue that GenAI's scalability is advantageous in institutions with large student cohorts, particularly in grading objective assessments such as multiple-choice tests. Ellis and Slade (2023) further support GenAI's automated grading as effective in maintaining assessment quality while boosting feedback timeliness. These findings suggest that GenAI-driven grading systems offer substantial educational benefits by optimising assessment processes (Chen et al., 2024; Omar et al., 2024).

3.6.2 Personalised Feedback for Supporting Student Growth

GenAI's ability to deliver personalised feedback represents one of its most transformative educational applications, tailoring insights to individual learning needs. Bai et al. (2024) highlight how GenAI analyses student responses to provide feedback that identifies individual strengths and weaknesses, thereby creating a highly tailored learning experience. Sharples (2023) argues that personalised feedback is crucial to effective learning, as it helps students understand not only their mistakes but also the areas they should focus on to improve. This specific, constructive feedback model fosters both academic growth and confidence in students, allowing them to self-direct their progress.

Hostetter et al. (2024) underscore that such personalised feedback can be adapted to fit each student's unique learning style, making it accessible and relevant to their educational journey. Mendez (2024) adds that these adaptive feedback approaches particularly benefit non-native speakers, providing clear guidance on linguistic challenges and academic writing norms. Collectively, these studies highlight the transformative potential of GenAI in offering feedback that is not generic but carefully attuned to the learner's needs, thereby enhancing comprehension, retention, and academic progress (Adarkwah, 2024; Akpan et al., 2024).

3.6.3 Facilitating Ongoing Monitoring and Feedback Loops

Generative AI's formative assessment capabilities extend beyond traditional summative evaluation, offering educators the tools to monitor and support ongoing student development. Ballantine et al. (2024) describe how GenAI can produce diagnostic assessments that provide continuous feedback, allowing educators to detect and address learning gaps in real time. This approach aligns with a formative, student-centred learning model, where educators can intervene and adjust instruction based on real-time insights.

Jochim and Lenz-Kesekamp (2024) argue that interim assessments, quizzes, and adaptive questions generated by GenAI are instrumental in tracking students’ learning trajectories, enabling instructors to offer timely interventions. Cacho (2024) adds that such formative feedback keeps students motivated, as it allows them to gauge their progress without waiting until the end of a course or module. Together, these findings advocate for GenAI as a valuable tool for formative assessment, delivering actionable feedback that supports sustained learning engagement and development (Van Wyk, 2024; Yeralan & Lee, 2023).

3.6.4 Promoting Self-Regulated, Independent Learning and Metacognitive Skills

GenAI facilitates self-regulated learning by empowering students to track their own progress and make informed adjustments to their study strategies. Benuyenah and Dewnarain (2024) highlight that GenAI tools offer self-assessment options, allowing students to measure their performance and become more autonomous in their learning journey. Adarkwah (2024) reinforces this notion, noting that GenAI encourages students to set learning objectives, monitor their own progress, and make necessary modifications based on personalised AI-generated feedback.

Lee et al. (2024) explore how GenAI-driven feedback fosters the development of metacognitive skills, helping students to reflect on their own learning processes, identify areas for improvement, and actively engage in self-directed learning. Omar et al. (2024) add that the constant availability of GenAI feedback allows students to take control of their educational journey, encouraging a proactive approach to growth. These studies collectively affirm that GenAI promotes self-regulation and metacognition, equipping students with the skills needed for lifelong learning and adaptability in a continuously evolving knowledge landscape (Jho & Ha, 2024; Sullivan et al., 2023).

3.6.5 Over-reliance on AI for Assessments

Over-reliance on generative AI tools, such as ChatGPT, for assessments in higher education poses several challenges to students' LS. While AI tools can enhance learning by offering immediate feedback and personalised learning experiences (Guettala et al., 2024; Jho & Ha, 2024), there is a risk that students may come to depend too heavily on these tools, undermining critical thinking and deep learning. Studies suggest that excessive reliance on AI for tasks such as writing and problem-solving can lead to a superficial understanding of course content, as students may prioritise convenience over genuine engagement with the material (Adel et al., 2024; Asad et al., 2024). This dependency may hinder the development of essential cognitive skills, such as analysis, synthesis, and independent problem-solving, which are crucial for academic success and intellectual growth in higher education (Baek et al., 2024).

Furthermore, the overuse of AI in assessments may lead to ethical and academic integrity concerns, with students potentially submitting AI-generated content as their own work (Adarkwah, 2024). Such practices not only compromise the learning process but also jeopardise academic standards, as institutions may struggle to distinguish between student-generated and AI-generated content (Cacho, 2024; Hostetter et al., 2024). For educators, this trend emphasises the need for clear guidelines and policies to balance the advantages of AI integration with the potential risks of diminished learning outcomes. A critical approach to AI usage in education involves fostering collaboration between AI tools and traditional educational methods, ensuring that students benefit from technology while developing the skills necessary for independent learning and success (Amershi et al., 2019; Jeon & Lee, 2023).

3.6.6 Risk of AI Misinterpretation in Feedback

The risk of AI misinterpretation in feedback is a significant concern for student LS, especially in the context of generative AI tools like ChatGPT. As these AI systems are increasingly integrated into higher education, students may rely on AI-generated feedback for assignments and learning activities. However, the complex nature of language models means that AI may sometimes offer misinterpretations of student input or provide feedback that lacks the nuanced understanding required for academic contexts. This misinterpretation could lead to students adopting incorrect conclusions or misconceptions, undermining their educational progress (Adel et al., 2024; Asad et al., 2024). In the absence of human oversight, the inability of AI to understand subtle academic requirements or disciplinary-specific standards could hinder students' critical thinking and problem-solving skills, which are essential for academic success (Baek et al., 2024; Hostetter et al., 2024).

Moreover, while AI feedback has the potential to enhance student engagement and personalise learning, the lack of emotional and contextual sensitivity poses another challenge (Jochim & Lenz-Kesekamp, 2024; Guettala et al., 2024). Students may misinterpret AI-generated feedback as impersonal or disconnected from the learning objectives, which can negatively impact their motivation and overall learning experience. The challenge lies in ensuring that AI feedback complements, rather than replaces, the human element of teaching. Instructors must play an active role in curating and validating AI-provided feedback to maintain the quality and depth of students' learning (Cacho, 2024; Jeon & Lee, 2023).

3.7 Language and Writing Support via GenAI

3.7.1 Providing Assistance in Grammar and Technical Precision in Writing

GenAI is recognised for its substantial contributions to language support, particularly in providing immediate feedback on grammar and style. Adel et al. (2024) discuss how AI-driven grammar checkers can identify technical errors in real-time, enabling students to learn correct language usage independently. This capability is particularly beneficial for non-native speakers who may struggle with nuanced grammatical rules. Ellis and Slade (2023) highlight GenAI's accessibility for students who require frequent language support, allowing them to refine their writing skills autonomously.

Cacho (2024) suggests that the non-intrusive nature of GenAI feedback boosts students' confidence by allowing them to improve incrementally. Mendez (2024) reinforces this point, observing that grammar and style feedback provided by GenAI tools fosters independence, encouraging students to take ownership of their language learning journey. Together, these studies underscore GenAI's value as a supportive tool for foundational writing skills, enhancing technical precision in language while fostering autonomy and confidence (Chen et al., 2024; Hostetter et al., 2024).

3.7.2 Enhancing Vocabulary and Language Diversity for Expanding Lexical Repertoires

GenAI offers diverse linguistic options that allow students to experiment with vocabulary and language structures, promoting richer language expression. Akpan et al. (2024) examine how GenAI applications can suggest synonyms, idiomatic expressions, and varied sentence structures, enabling students to diversify their language use. Baek et al. (2024) argue that GenAI introduces students to advanced academic and professional vocabulary, which is vital for higher-level academic writing.

Bai et al. (2024) find that GenAI vocabulary suggestions help students avoid repetitive phrasing, enhancing language proficiency. Guettala et al. (2024) emphasise that the inclusion of varied linguistic styles within GenAI applications enables students to adapt their language to different contexts, making it a valuable tool for versatile language learning. Collectively, these studies position GenAI as an important resource for developing language diversity, allowing students to expand their lexical repertoires and become more confident, versatile writers (Jeon & Lee, 2023; Sullivan et al., 2023).

3.7.3 Supporting Academic Writing Skills, Structuring Arguments and Enhancing Cohesion

GenAI supports higher-order academic writing skills by guiding students in the organisation and clarity of their arguments. Adarkwah (2024) discusses how GenAI offers structured guidance, assisting students in the logical arrangement of arguments, paragraph coherence, and the development of cohesive essays. Sharples (2023) reinforces that GenAI serves as a virtual writing mentor, providing real-time suggestions on improving clarity, logic, and argumentative flow, which are critical in high-quality academic work.

Ellis and Slade (2023) further note that GenAI tools help students adhere to academic conventions, such as citation styles and referencing, thereby easing common challenges in academic writing. Lee et al. (2024) underscore that GenAI’s analytical capabilities offer constructive feedback on argumentation, encouraging students to develop persuasive, well-supported academic arguments. Collectively, these findings suggest that GenAI is an invaluable resource in academic writing development, supporting students in mastering essential skills for effective scholarly communication (Chen et al., 2024; Yeralan & Lee, 2023).

3.7.4 Facilitating Multilingual and Translation Support and Enhancing Inclusivity for Diverse Linguistic Backgrounds

GenAI's multilingual support capabilities facilitate language inclusivity, assisting students who may struggle with English or are learning new languages. Omar et al. (2024) highlight that GenAI's translation tools help non-native English speakers understand complex material in their native languages, while simultaneously scaffolding their English language acquisition. Guettala et al. (2024) emphasise that these translation features support foreign language learning by enhancing reading comprehension and language retention.

Jho and Ha (2024) discuss how GenAI's translation features can foster cross-linguistic understanding, helping students from diverse linguistic backgrounds access content more effectively. Van Wyk (2024) adds that GenAI's multilingual capabilities enable students to practice translation exercises, which can aid in language retention and build confidence. Together, these findings suggest that GenAI's language inclusivity fosters a supportive learning environment, helping students overcome linguistic challenges and thrive in globalised educational settings (Adarkwah, 2024; Mendez, 2024).

3.7.5 Over-dependence on AI Writing Tools

The over-dependence on generative AI tools, such as ChatGPT, among students in higher education poses several challenges to their LS. As AI writing tools become increasingly accessible and integrated into academic practices, students may rely too heavily on them, which can lead to a decline in critical thinking and independent learning. Research has shown that while these tools offer immediate support in writing tasks, students may lose opportunities to engage in deeper cognitive processes, such as synthesis, analysis, and creative problem-solving (Baek et al., 2024; Benuyenah & Dewnarain, 2024). Over-reliance on AI tools can thus hinder the development of essential academic skills, such as original thought, academic writing proficiency, and self-regulated learning. Students may also struggle to understand the intricacies of their own work, as AI-generated content can mask gaps in comprehension or knowledge (Hostetter et al., 2024).

Moreover, excessive dependence on AI tools can perpetuate a passive learning environment, where students treat these technologies as substitutes for engagement with course materials rather than as supplements (Adarkwah, 2024; Jochim & Lenz-Kesekamp, 2024). This can create long-term barriers to academic growth, as students fail to acquire the self-efficacy and problem-solving abilities needed to succeed in complex academic and professional contexts. While generative AI can certainly enhance efficiency, its misuse in education raises concerns about diminished academic integrity and the potential erosion of essential learning processes, particularly when students turn to these tools for simple answers or shortcuts in assignments (Chen et al., 2024). Thus, educational institutions must guide students in striking a balance between utilising AI tools and fostering independent, reflective learning to ensure that these technologies do not undermine overall academic success.

3.7.6 AI Bias in Language and Translation

AI Bias in language and translation poses significant challenges to students' learning experiences in higher education, particularly in the use of generative AI tools for educational purposes. The inherent biases in AI models can lead to inaccurate or culturally insensitive translations, which might negatively impact students' comprehension and academic success. As generative AI becomes increasingly integrated into learning environments, the accuracy of language models in understanding context, idioms, and cultural nuances becomes crucial. Inaccurate translations or biased language can hinder students' ability to engage with course material effectively, potentially diminishing the educational quality they receive. For instance, when generative AI is employed for tasks such as academic writing assistance, non-native speakers of English may encounter challenges in ensuring that their ideas are accurately represented, especially if the AI fails to account for regional linguistic variations or cultural subtleties. This bias is not limited to language translation but extends to AI-generated feedback on student writing, which can perpetuate stereotypes or reinforce pre-existing biases, undermining the goal of fostering an inclusive learning environment (Akpan et al., 2024; Adel et al., 2024).

Moreover, AI bias in language and translation can exacerbate inequality in higher education by disproportionately affecting students from underrepresented or non-dominant language backgrounds. If AI tools are not designed to be culturally responsive, they may unintentionally reinforce disparities in learning outcomes by providing biased or exclusionary support. For students using generative AI for language learning or academic writing, biased outputs can impede their ability to meet learning objectives or result in inaccurate academic representations. This, in turn, may affect their academic performance, grading, and overall learning experience (Asad et al., 2024). Addressing AI bias is crucial for ensuring that AI tools can be leveraged effectively in higher education, promoting equitable learning experiences. Institutions must therefore implement strategies to evaluate and mitigate bias in generative AI models, ensuring that these technologies support all students, regardless of their linguistic or cultural background (Adarkwah, 2024; Hostetter et al., 2024).

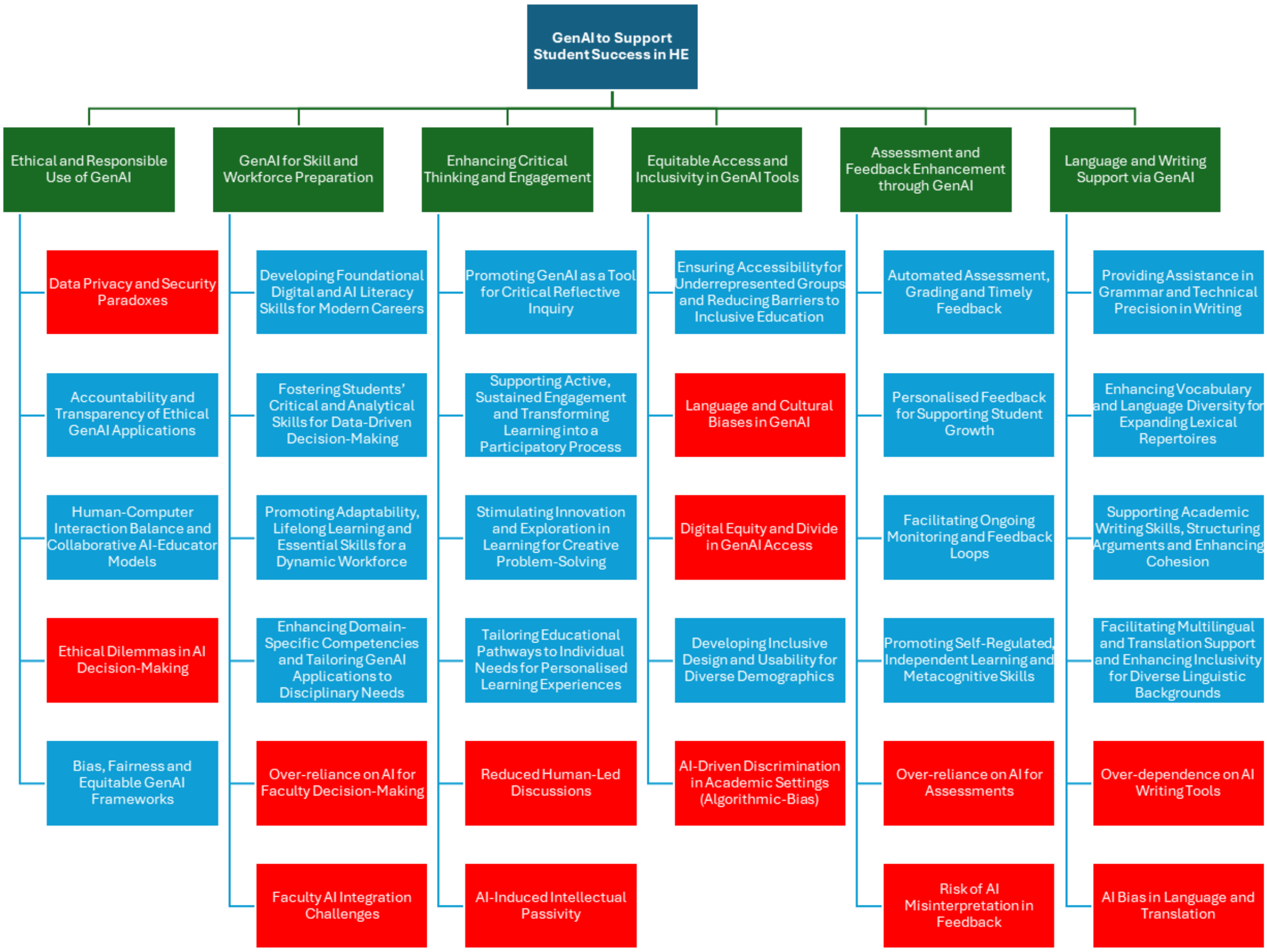

GenAI to Support Student Success in HE
Ethical and Responsible Use of GenAI
Data Privacy and Security Paradoxes
Accountability and Transparency of Ethical GenAI Applications
Human-Computer Interaction Balance and Collaborative AI-Educator Models
Ethical Dilemmas in AI Decision-Making
Bias, Fairness and Equitable GenAI Frameworks
GenAI for Skill and Workforce Preparation
Developing Foundational Digital and AI Literacy Skills for Modern Careers
Fostering Students' Critical and Analytical Skills for Data-Driven Decision-Making
Promoting Adaptability, Lifelong Learning and Essential Skills for a Dynamic Workforce
Enhancing Domain-Specific Competencies and Tailoring GenAI Applications to Disciplinary Needs
Over-reliance on AI for Faculty Decision-Making
Faculty AI Integration Challenges
Enhancing Critical Thinking and Engagement
Promoting GenAI as a Tool for Critical Reflective Inquiry
Supporting Active, Sustained Engagement and Transforming Learning into a Participatory Process
Stimulating Innovation and Exploration in Learning for Creative Problem-Solving
Tailoring Educational Pathways to Individual Needs for Personalised Learning Experiences
Reduced Human-Led Discussions
AI-Induced Intellectual Passivity
Equitable Access and Inclusivity in GenAI Tools
Ensuring Accessibility for Underrepresented Groups and Reducing Barriers to Inclusive Education
Language and Cultural Biases in GenAI
Digital Equity and Divide in GenAI Access
Developing Inclusive Design and Usability for Diverse Demographics
AI-Driven Discrimination in Academic Settings (Algorithmic-Bias)
Assessment and Feedback Enhancement through GenAI
Automated Assessment, Grading and Timely Feedback
Personalised Feedback for Supporting Student Growth
Facilitating Ongoing Monitoring and Feedback Loops
Promoting Self-Regulated, Independent Learning and Metacognitive Skills
Over-reliance on AI for Assessments
Risk of AI Misinterpretation in Feedback
Language and Writing Support via GenAI
Providing Assistance in Grammar and Technical Precision in Writing
Enhancing Vocabulary and Language Diversity for Expanding Lexical Repertoires
Supporting Academic Writing Skills, Structuring Arguments and Enhancing Cohesion
Facilitating Multilingual and Translation Support and Enhancing Inclusivity for Diverse Linguistic Backgrounds
Over-dependence on AI Writing Tools
AI Bias in Language and Translation

Figure 2. **Overview of Progressive (Blue) & Regressive (Red) Issues Impacting LS**

4.0 Discussion

Based on our RQ, the study’s findings highlight several transformative impacts of LLMs on teaching and LS, spanning from enhanced critical engagement to ethical considerations, as well as skill development and personalised learning. However, alongside these advancements, it is essential to critically evaluate both the progressive and regressive dimensions of LLM integration in higher education, ensuring that AI-driven solutions foster educational equity and intellectual growth without exacerbating existing disparities or ethical concerns.

4.1 Ethical and Responsible Use: Shaping Trust and Integrity in Educational AI

One of the most salient findings is the critical need for responsible data management and transparency in LLM integration. While LLMs offer substantial educational value, particularly in personalised feedback and accessibility, they also introduce ethical risks associated with data privacy, transparency, and potential biases. For instance, Aad and Hardey (2024) highlight faculty concerns regarding data misuse, indicating that without robust privacy protocols, institutions risk compromising student trust. This aligns with the ethical imperative to ensure that LLMs are deployed with a commitment to data security and fairness, fostering a learning environment where students feel protected and valued.

The progressive potential of LLMs lies in their ability to enhance trust through explainable AI (XAI) frameworks. Findings from Hostetter et al. (2024) underscore the need for transparent AI decision-making, empowering students to critically engage with AI-generated assessments. However, a regressive issue emerges in cases where opaque AI mechanisms obscure the rationale behind feedback, leading to potential over-reliance on AI-generated content without deeper comprehension or accountability. To mitigate this risk, educators and institutions must balance AI use with pedagogical oversight, ensuring that AI augments rather than dictates the learning process.

4.2 Cognitive and Critical Skills Development: Advancing Analytical Rigor and Reflective Learning

The findings indicate that LLMs significantly enhance cognitive and analytical skills, encouraging students to engage in deeper critical thinking and problem-solving. Through AI-generated content and feedback, students are exposed to diverse perspectives that push them to question, evaluate, and synthesise information, a process that aligns with Bloom’s higher-order skills. Baek et al. (2024) illustrate how GenAI fosters an analytical mindset, training students to interact critically with content rather than passively accept it.

However, while LLMs support cognitive development, a regressive concern arises when students become overly dependent on AI-generated insights, potentially undermining their ability to develop independent analytical skills. Guettala et al. (2024) caution against the passive consumption of AI-produced content, noting that critical engagement must be actively cultivated through structured learning interventions. The challenge, therefore, is to harness LLMs as facilitators of learning rather than substitutes for critical inquiry.

4.3 Skill and Workforce Preparation: Building AI Literacy and Preparing for Technology-Driven Careers

LLMs play an instrumental role in preparing students for technology-integrated careers, equipping them with essential digital and AI literacy skills that are increasingly valuable across industries. The results reveal that exposure to GenAI helps students gain confidence in using AI tools, bridging the gap between academic training and industry demands. Adarkwah (2024) and Chen et al. (2024) advocate for integrating AI literacy as a core component of modern education, positing that familiarity with LLMs provides students with a critical advantage in tech-driven fields.

Nevertheless, the rapid adoption of LLMs in education raises concerns about the digital divide. While some students benefit from early exposure to AI technologies, others may struggle due to a lack of access to digital resources, exacerbating inequalities in workforce preparedness. Akpan et al. (2024) argue that institutions must address these disparities through targeted policies that ensure equitable AI literacy development, particularly for students from underprivileged backgrounds.

4.4 Enhanced Engagement and Personalised Learning: Fostering Self-Regulation and Autonomous Learning

The research highlights the pivotal role of LLMs in fostering student engagement through personalised learning experiences. The results indicate that LLMs can dynamically adjust to individual learning needs, delivering tailored feedback and recommendations that keep students challenged and motivated. Hostetter et al. (2024) and Mendez (2024) demonstrate how such personalised support fosters student self-regulation, empowering learners to take control of their academic progress and engage more deeply with content.

However, a regressive concern emerges in the potential for algorithmic reinforcement of existing learning patterns, which could limit intellectual exploration. Lee et al. (2024) caution that if LLMs prioritise efficiency over diversity of thought, students may receive feedback that overly conforms to established norms rather than encouraging innovative or divergent thinking. Addressing this requires AI frameworks that actively promote creativity, cognitive flexibility, and adaptive learning.

4.5 Inclusivity and Equitable Access: Creating a Supportive Environment for Diverse Learners

While LLMs offer numerous educational benefits, their equitable deployment is essential to ensuring that all students can access these advantages. The research underscores the importance of designing GenAI systems with inclusivity in mind, ensuring that adaptive and multimodal features support students from diverse backgrounds and abilities. Aad and Hardey (2024) and Guettala et al. (2024) discuss the risks of cultural and linguistic biases in AI, suggesting that inclusive GenAI design must address the unique needs of underrepresented groups, including those with disabilities and non-native English-speaking backgrounds.

A regressive issue in AI-driven learning is the risk of reinforcing linguistic and cultural biases, which can lead to marginalisation of certain student groups. If LLMs are trained on datasets that underrepresent diverse linguistic and cultural contexts, they may produce biased outputs that disadvantage non-dominant language speakers. Adarkwah (2024) highlights the need for institutions to critically assess AI models and ensure that inclusivity is embedded within algorithmic frameworks. Moreover, bridging digital divides remains a pressing challenge, requiring proactive measures to guarantee equitable AI access across socio-economic groups.

4.6 Holistic Insights on the Role of LLMs in Higher Education

In summary, the integration of LLMs into HE brings transformative potential, provided that ethical, cognitive, and inclusivity dimensions are prioritised. The findings suggest that when thoughtfully implemented, LLMs can enhance teaching efficacy, promote critical thinking, foster skill development, and support personalised learning experiences. However, achieving these outcomes requires a commitment to responsible AI practices that uphold data security, accountability, and equitable access.

This analysis reinforces that LLMs are not mere instructional tools but complex agents that interact with teaching and learning dynamics. While they offer progressive educational opportunities, their limitations highlight the need for continuous critical evaluation to prevent regressive impacts. The potential impact of LLMs extends beyond LS to include critical thinking, adaptability, ethical awareness, and inclusivity, qualities that are essential for preparing students to navigate an increasingly AI-integrated world. As such, institutions must adopt a balanced approach that leverages AI innovations while safeguarding pedagogical integrity and educational equity.

5.0 Conclusion and Contribution

This study has examined how the integration of LLMs influences teaching efficacy, student engagement, and overall LS in HE, with a focus on the ethical, cognitive, and skill-development dimensions of AI-enhanced learning. The findings suggest that when integrated with a structured and responsible approach, LLMs can significantly enrich educational practices and enhance student outcomes. By enabling personalised learning, fostering critical engagement, and addressing the evolving skill demands of a technology-driven workforce, LLMs emerge as transformative tools in the contemporary HE landscape.

A key conclusion is the necessity of embedding ethical responsibility within LLM applications to foster trust and transparency in educational AI. Without prioritising data security, accountability, and fairness, the benefits of LLMs for teaching efficacy and student engagement may be undermined. This study highlights the importance of robust data privacy protocols and bias mitigation strategies in developing sustainable, trustworthy AI-augmented learning environments. Ensuring the transparency of LLM-generated content is essential for maintaining student trust and encouraging critical engagement with AI-assisted feedback. A strong ethical foundation positions LLMs as collaborative tools that complement human instruction while preserving the integrity of educational practices.

From a cognitive perspective, the study underscores how LLMs enhance analytical and critical thinking skills by fostering deeper engagement with content. Through adaptive feedback and AI-driven learning support, LLMs encourage students to question, evaluate, and synthesise information, cultivating higher-order cognitive skills essential for academic success and real-world problem-solving. Furthermore, by promoting creativity and problem-solving, LLMs enable students to explore diverse approaches to complex questions, fostering intellectual flexibility. These capabilities align with Bloom's higher-order skills, reinforcing the argument that LLMs contribute not only to LS but also to deeper cognitive engagement, empowering students as active, self-directed learners.

In terms of skill development, the findings suggest that LLMs play a crucial role in preparing students for a workforce increasingly shaped by AI. By embedding digital literacy, critical thinking, and adaptability within learning processes, LLMs help bridge the gap between academic training and industry expectations. Exposure to AI-driven tools fosters confidence and competence in technology, equipping students with practical, transferable skills relevant across various professional domains. Thus, the integration of LLMs enhances both immediate LS within the classroom and long-term career readiness, positioning students for success in an AI-driven economy.

The study also highlights the role of LLMs in promoting student engagement through personalised and interactive learning experiences. By tailoring feedback and learning pathways to individual student needs, LLMs foster sustained engagement and encourage self-regulation. The ability to provide real-time, adaptive feedback empowers students to take ownership of their learning, cultivating autonomy and self-discipline; key attributes for lifelong learning. In maintaining high levels of engagement, LLMs not only support immediate LS but also contribute to the development of resilient, self-motivated learners capable of navigating evolving knowledge landscapes.

However, this study also acknowledges several limitations. The rapid evolution of AI means that the findings presented here may require reassessment as LLM capabilities continue to advance. Additionally, while this study offers broad insights into the impact of LLMs across educational contexts, specific challenges within individual institutions, disciplines, and student demographics require further investigation. The ethical and cognitive impacts of LLMs were primarily assessed through secondary research; thus, direct empirical studies on student experiences in LLM-integrated classrooms are necessary to validate these findings.

Future research should focus on empirically assessing the longitudinal impacts of LLMs on teaching efficacy, student engagement, and LS across diverse educational settings. Controlled studies could offer valuable insights into how LLM-driven feedback and personalised learning strategies influence student outcomes over time. Additionally, there is a need for continuous development of ethical guidelines that evolve alongside AI technology, particularly concerning data privacy and inclusivity. Investigating adaptive GenAI models tailored to underrepresented groups, such as non-native speakers and students with disabilities, would provide crucial insights into how LLMs can foster a genuinely inclusive learning environment. Cross-disciplinary research could further illuminate the differential impacts of LLMs across academic fields, such as humanities versus STEM, offering a comprehensive perspective on AI’s role in education.

In conclusion, this study highlights the transformative potential of LLMs like ChatGPT in enhancing teaching efficacy, deepening student engagement, and supporting overall LS in HE. By addressing ethical considerations, supporting cognitive skill development, and equipping students with industry-relevant competencies, LLMs can serve as powerful tools for fostering an inclusive, responsive, and effective educational ecosystem. Achieving these outcomes will require ongoing research, ethical vigilance, and a nuanced approach to diverse student needs. As HE continues to adapt to technological advancements, LLMs hold the potential to reshape the academic landscape, equipping students with the adaptability, ethical awareness, and critical thinking skills essential for success in an AI-driven world.

6.0 References

Aad, S., & Hardey, M. (2024). Generative AI: Hopes, controversies and the future of faculty roles in education. *Quality Assurance in Education*. https://doi.org/10.1108/QAE-02-2024-0043

Adarkwah, M. A. (2024). GenAI-Infused Adult Learning in the Digital Era: A Conceptual Framework for Higher Education. *Adult Learning*, 10451595241271161. https://doi.org/10.1177/10451595241271161

Adel, A., Ahsan, A., & Davison, C. (2024). ChatGPT Promises and Challenges in Education: Computational and Ethical Perspectives. *Education Sciences*, *14*(8), 814. https://doi.org/10.3390/educsci14080814

Ajevski, M., Barker, K., Gilbert, A., Hardie, L., & Ryan, F. (2023). ChatGPT and the future of legal education and practice. *The Law Teacher*, *57*(3), 352–364. https://doi.org/10.1080/03069400.2023.2207426

Akpan, I. J., Kobara, Y. M., Owolabi, J., Akpan, A. A., & Offodile, O. F. (2024). Conversational and generative artificial intelligence and human–chatbot interaction in education and research. *International Transactions in Operational Research*, itor.13522. https://doi.org/10.1111/itor.13522

Alasadi, E. A., & Baiz, C. R. (2023). Generative AI in Education and Research: Opportunities, Concerns, and Solutions. *Journal of Chemical Education*, *100*(8), 2965–2971. https://doi.org/10.1021/acs.jchemed.3c00323

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Asad, M. M., Shahzad, S., Shah, S. H. A., Sherwani, F., & Almusharraf, N. M. (2024). ChatGPT as artificial intelligence-based generative multimedia for English writing pedagogy: Challenges and opportunities from an educator's perspective. *The International Journal of Information and Learning Technology*. https://doi.org/10.1108/IJILT-02-2024-0021

Avison, D., & Elliot, S. (2006). Scoping the discipline of information systems. *Information systems: the state of the field*, *4*(2), 3-18.

Baek, C., Tate, T., & Warschauer, M. (2024). "ChatGPT seems too good to be true": College students' use and perceptions of generative AI. *Computers and Education: Artificial Intelligence, 7*, 100294. https://doi.org/10.1016/j.caeai.2024.100294

Bai, J. Y. H., Zawacki-Richter, O., & Muskens, W. (2024). Re-Examining The Future Prospects Of Artificial Intelligence In Education In Light of the GDPR And Chatgpt. *Turkish Online Journal of Distance Education*, *25*(1), 20–32. https://doi.org/10.17718/tojde.1248901

Ballantine, J., Boyce, G., & Stoner, G. (2024). A critical review of AI in accounting education: Threat and opportunity. *Critical Perspectives on Accounting*, *99*, 102711. https://doi.org/10.1016/j.cpa.2024.102711

Benuyenah, V., & Dewnarain, S. (2024). Students' Intention to Engage With ChatGPT and Artificial Intelligence in Higher Education Business Studies Programmes: An Initial Qualitative Exploration. *International Journal of Distance Education Technologies*, *22*(1), 1–21. https://doi.org/10.4018/IJDET.348061

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (Version 5).

Cacho, R. (2024). Integrating Generative AI in University Teaching and Learning: A Model for Balanced Guidelines. *Online Learning*, *28*(3). https://doi.org/10.24059/olj.v28i3.4508

Chen, K., Tallant, A. C., & Selig, I. (2024). Exploring generative AI literacy in higher education: Student adoption, interaction, evaluation and ethical perceptions. *Information and Learning Sciences*. https://doi.org/10.1108/ILS-10-2023-0160

Cheng, T.-Y. (2023). On the quadrants of the thing-world relations: A critical revision of Hartmut Rosa's resonance theory in terms of thing-world. *The Journal of Chinese Sociology*, *10*(1), 11. https://doi.org/10.1186/s40711-023-00191-8

Delcker, J., Heil, J., Ifenthaler, D., Seufert, S., & Spirgi, L. (2024). First-year students AI-competence as a predictor for intended and de facto use of AI-tools for supporting learning processes in higher education. *International Journal of Educational Technology in Higher Education*, *21*(1), 1-13.

Duong, C. D., Vu, T. N., & Ngo, T. V. N. (2023). Applying a modified technology acceptance model to explain higher education students' usage of ChatGPT: A serial multiple mediation model with knowledge sharing as a moderator. *The International Journal of Management Education*, *21*(3), 100883. https://doi.org/10.1016/j.ijme.2023.100883

Edwards, A. P. (2018). Animals, Humans, and Machines: Interactive Implications of Ontological Classification. In A. L. Guzman (Ed.), *Human-Machine Communication: Rethinking Communication, Technology, and Ourselves* (pp. 29–49). Peter Lang. https://doi.org/10.3726/b14414

Ellis, A. R., & Slade, E. (2023). A New Era of Learning: Considerations for ChatGPT as a Tool to Enhance Statistics and Data Science Education. *Journal of Statistics and Data Science Education*, *31*(2), 128–133. https://doi.org/10.1080/26939169.2023.2223609

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of*

Educational Technology in Higher Education, *20*(1), 57. https://doi.org/10.1186/s41239-023-00425-2

Guettala, M., Bourekkache, S., Kazar, O., & Harous, S. (2024). Generative Artificial Intelligence in Education: Advancing Adaptive and Personalized Learning. *Acta Informatica Pragensia*, *13*(3), 460–489. https://doi.org/10.18267/j.aip.235

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and Communication: A Human–Machine Communication Research Agenda. *New Media & Society*, *22*(1), 70–86. https://doi.org/10.1177/1461444819858691

Hostetter, A. B., Call, N., Frazier, G., James, T., Linnertz, C., Nestle, E., & Tucci, M. (2024). Student and Faculty Perceptions of Generative Artificial Intelligence in Student Writing. *Teaching of Psychology*, 00986283241279401. https://doi.org/10.1177/00986283241279401

Iatrellis, O., Samaras, N., Kokkinos, K., & Panagiotakopoulos, T. (2024). Leveraging Generative AI for Sustainable Academic Advising: Enhancing Educational Practices through AI-Driven Recommendations. *Sustainability*, *16*(17), 7829. https://doi.org/10.3390/su16177829

Iliyasu, Z., Abdullahi, H. O., Iliyasu, B. Z., Bashir, H. A., Amole, T. G., Abdullahi, H. M., Abdullahi, A. U., Kwaku, A. A., Dahir, T., Tsiga-Ahmed, F. I., Jibo, A. M., Salihu, H. M., & Aliyu, M. H. (2024). Correlates of Medical and Allied Health Students' Engagement with Generative AI in Nigeria. *Medical Science Educator*. https://doi.org/10.1007/s40670-024-02181-y

Imran, M., Almusharraf, N., Abdellatif, M. S., & Abbasova, M. Y. (2024). Artificial Intelligence in Higher Education: Enhancing Learning Systems and Transforming Educational Paradigms. *International Journal of Interactive Mobile Technologies (iJIM)*, *18*(18), 34–48. https://doi.org/10.3991/ijim.v18i18.49143

Jensen, L. X., Buhl, A., Sharma, A., & Bearman, M. (2024). Generative AI and higher education: A review of claims from the first months of ChatGPT. *Higher Education*. https://doi.org/10.1007/s10734-024-01265-3

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, *28*(12), 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

Jho, H., & Ha, M. (2024). Towards Effective Argumentation: Design and Implementation of a Generative AI-Based Evaluation and Feedback System. *Journal of Baltic Science Education*, *23*(2), 280–291. https://doi.org/10.33225/jbse/24.23.280

Jochim, J., & Lenz-Kesekamp, V. K. (2024). Teaching and testing in the era of text-generative AI: Exploring the needs of students and teachers. *Information and Learning Sciences*. https://doi.org/10.1108/ILS-10-2023-0165

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, *51*(1), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

Kumar, S., Gunn, A., Rose, R., Pollard, R., Johnson, M., & Ritzhaupt, A. (2024). The Role of Instructional Designers in the Integration of Generative Artificial Intelligence in Online and Blended Learning in Higher Education. *Online Learning*, *28*(3). https://doi.org/10.24059/olj.v28i3.4501

Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., Lekkas, D., & Palmer, E. (2024). The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Computers and Education: Artificial Intelligence*, *6*, 100221. https://doi.org/10.1016/j.caeai.2024.100221

Li, J., Li, J., & Su, Y. (2024). A Map of Exploring Human Interaction Patterns with LLM: Insights into Collaboration and Creativity. In H. Degen & S. Ntoa (Eds.), *Artificial Intelligence in HCI* (pp. 60–85). Springer Nature Switzerland.

Li, X., Li, R., Li, M., Yao, L., Van Spall, H., Zhao, K., Chen, Y., Xiao, F., Fu, Q., & Xie, F. (2024). A Systematic Review and Quality Assessment of Cardiovascular Disease-Specific Health-Related Quality-of-Life Instruments Part I: Instrument Development and Content Validity. *Value in Health*, *27*(8), 1130–1148. https://doi.org/10.1016/j.jval.2024.04.001

Mendez, J. D. (2024). Student Perceptions of Artificial Intelligence Utility in the Introductory Chemistry Classroom. *Journal of Chemical Education*, *101*(8), 3547–3549. https://doi.org/10.1021/acs.jchemed.4c00075

Mohamed Shaffril, H. A., Samsuddin, S. F., & Abu Samah, A. (2021). The ABC of systematic literature review: The basic methodological guidance for beginners. *Quality & Quantity*, *55*(4), 1319–1346. https://doi.org/10.1007/s11135-020-01059-6

Mohammed, F. S., & Ozdamli, F. (2024). A Systematic Literature Review of Soft Skills in Information Technology Education. *Behavioral Sciences*, *14*(10), 894. https://doi.org/10.3390/bs14100894

Omar, A., Shaqour, A. Z., & Khlaif, Z. N. (2024). Attitudes of faculty members in Palestinian universities toward employing artificial intelligence applications in higher education: Opportunities and challenges. *Frontiers in Education*, *9*, 1414606. https://doi.org/10.3389/feduc.2024.1414606

Pack, A., & Maloney, J. (2023). Using Generative Artificial Intelligence for Language Education Research: Insights from Using OPENAI 's CHATGPT. *TESOL Quarterly*, *57*(4), 1571–1582. https://doi.org/10.1002/tesq.3253

Raman, R., Mandal, S., Das, P., Kaur, T., Sanjanasri, J. P., & Nedungadi, P. (2024). Exploring University Students' Adoption of ChatGPT Using the Diffusion of Innovation Theory and Sentiment Analysis With Gender Dimension. *Human*

Behavior and Emerging Technologies, *2024*(1), 3085910. https://doi.org/10.1155/2024/3085910

Sharples, M. (2023). Towards social generative AI for education: Theory, practices and ethics. *Learning: Research and Practice*, *9*(2), 159–167. https://doi.org/10.1080/23735082.2023.2261131

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching*, *6*(1). https://doi.org/10.37074/jalt.2023.6.1.17

Van Wyk, M. M. (2024). Is ChatGPT an opportunity or a threat? Preventive strategies employed by academics related to a GenAI-based LLM at a faculty of education. *Journal of Applied Learning & Teaching*, *7*(1). https://doi.org/10.37074/jalt.2024.7.1.15

Watson, R. T., & Webster, J. (2020). Analysing the past to prepare for the future: Writing a literature review a roadmap for release 2.0. *Journal of Decision Systems*, *29*(3), 129–147. https://doi.org/10.1080/12460125.2020.1798591

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), 13–23.

Wood, D., & Moss, S. H. (2024). Evaluating the impact of students' generative AI use in educational contexts. *Journal of Research in Innovative Teaching & Learning*, *17*(2), 152–167. https://doi.org/10.1108/JRIT-06-2024-0151

Yang, L., Zhang, H., Shen, H., Huang, X., Zhou, X., Rong, G., & Shao, D. (2021). Quality Assessment in Systematic Literature Reviews: A Software Engineering Perspective. *Information and Software Technology*, *130*, 106397. https://doi.org/10.1016/j.infsof.2020.106397

Yeralan, S., & Lee, L. A. (2023). Generative AI: Challenges to higher education. *Sustainable Engineering and Innovation*, *5*(2), 107–116. https://doi.org/10.37868/sei.v5i2.id196

Appendix I: Quality Assessment Scores

Author(s)	Research Type	Paper Title	Intro Criteria	Method Criteria	Results Criteria	Discussion Criteria	Overall Score
Aad, S., & Hardey, M. (2024)	Qualitative	Generative AI: hopes, controversies and the future of faculty roles in education.	2	2	1	2	7/12, 58.33%, Good
Adarkwah, M.A. (2024)	Qualitative	GenAI-Infused Adult Learning in the Digital Era: A Conceptual Framework for Higher Education.	2	2	2	2	8/12, 66.67%, Good
Adel, A., Ahsan, A., & Davison, C. (2024)	Qualitative	ChatGPT promises and challenges in education: Computational and ethical perspectives.	3	2	2	3	10/12, 83.33%, High
Adilov, N., et al. (2024)	Quantitative	ChatGPT and the course vulnerability index.	1	2	2	2	7/12, 58.33%, Good
Ajevski, M., et al. (2023)	Mixed	ChatGPT and the future of legal education and practice.	2	2	2	1	7/12, 58.33%, Good
Akpan, I.J., et al. (2024)	Mixed Methods	Conversational and generative artificial intelligence and human–chatbot interaction in education and research.	3	2	2	2	9/12, 75.0%, High
Al Murshidi, G., et al. (2024)	Quantitative	How understanding the limitations and risks of using ChatGPT can contribute to willingness to use.	2	2	1	2	7/12, 58.33%, Good
Alammari, A. (2024)	Mixed	Evaluating Generative AI Integration in Saudi Arabian Education: A Mixed-Methods Study.	3	3	2	3	11/12, 91.67%, High
Alasadi, E.A., & Baiz, C.R. (2023)	Quantitative	Generative AI in Education and Research: Opportunities, Concerns, and Solutions.	3	3	3	3	12/12, 100.0%, High

Al-Saiari, M.A., et al. (2024)	Qualitative	Investigating the Impact of Training Program on Generative AI Applications in Improving University Teaching.	2	2	2	2	8/12, 66.67%, Good
Al-Sofi, B. B. M. A. (2024)	Mixed	Artificial intelligence-powered tools and academic writing: to use or not to use ChatGPT.	2	2	1	2	7/12, 58.33%, Good
Álvarez-Álvarez, C., & Falcon, S. (2023)	Qualitative	Students’ preferences with university teaching practices: analysis of testimonials with artificial intelligence.	2	1	1	1	5/12, 41.67%, Moderate
Alzubi, A.A.F. (2024)	Quantitative	Generative Artificial Intelligence in the EFL Writing Context: Students' Literacy in Perspective.	3	3	2	3	11/12, 91.67%, High
Araujo, S.M., & Cruz-Correia, R. (2024)	Qualitative	Incorporating ChatGPT in Medical Informatics Education: Mixed Methods Study on Student Perceptions and Experiential Integration Proposals.	2	2	1	2	7/12, 58.33%, Good
Asad, M.M., et al. (2024)	Mixed	ChatGPT as artificial intelligence-based generative multimedia for English writing pedagogy: challenges and opportunities from an educator’s perspective.	2	2	1	2	7/12, 58.33%, Good
Asha, S., & Krishna, S.T. (2024)	Qualitative	Semantics-Based String Matching: A Review of Machine Learning Models.	0	1	0	0	1/12, 8.33%, Poor
Azntorsureanu, I., Voicu-DorobanÈ›u, R., Nisioiu, C.-F., & Ploae, C. (2024)	Mixed	Generative Artificial Intelligence and the Academic Integrity of Graduation Works in Economics.	1	1	1	1	4/12, 33.33%, Moderate
Baek, C., Tate, T., & Warschauer, M. (2024)	Qualitative	“ChatGPT seems too good to be true”: College students’ use and perceptions of generative AI.	3	3	3	3	12/12, 100.0%, High
Baek, T.H., Kim, J., & Kim, J.H. (2024)	Mixed	Effect of disclosing AI-generated content on prosocial advertising evaluation.	1	1	0	1	3/12, 25.0%, Moderat

							e
Bahani, M., El Ouaazizi, A., Avram, R., & Maalmi, K. (2024)	Mixed	Enhancing chest X-ray diagnosis with text-to-image generation: A data augmentation case study.	0	0	0	0	0/12, 0.0%, poor
Bai, J.Y.H., Zawacki-Richter, O., & Muskens, W. (2024)	Mixed	Re-Examining the Future Prospects of Artificial Intelligence in Education in Light of the GDPR and ChatGPT.	3	2	2	3	10/12, 83.33%, High
Bai, S., Gonda, D.E., & Hew, K.F. (2024)	Qualitative	Write-Curate-Verify: A Case Study of Leveraging Generative AI for Scenario Writing in Scenario-Based Learning.	2	2	2	2	8/12, 66.67%, Good
Bai, X., Xie, Y., Zhang, X., Han, H., & Li, J.-R. (2024)	Qualitative	Evaluation of Open-Source Large Language Models for Metal-Organic Frameworks Research.	0	0	0	0	0/12, 0.0%, poor
Ballantine, J., Boyce, G., & Stoner, G. (2024)	Mixed	A critical review of AI in accounting education: Threat and opportunity.	2	2	1	2	7/12, 58.33%, Good
Banh, L., & Strobel, G. (2023)	Mixed	Generative artificial intelligence.	1	1	0	1	3/12, 25.0%, Moderate
Bannister, P., Alcalde PeÃ±alver, E., & SantamarÃa Urbieta, A. (2024)	Qualitative	International Students and Generative Artificial Intelligence: A Cross-Cultural Exploratory Analysis of Higher Education Academic Integrity Policy.	2	2	2	2	8/12, 66.67%, Good
Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024)	Qualitative	Developing evaluative judgement for a time of generative artificial intelligence.	2	2	2	2	8/12, 66.67%, Good
Bennett, L., & Abusalem, A. (2024)	Quantitative	Artificial Intelligence (AI) and its Potential Impact on the Future of Higher Education.	1	1	1	1	4/12, 33.33%, Moderate
Benuyenah, V., & Dewnarain, S. (2024)	Quantitative	Students' Intention to Engage With ChatGPT and Artificial Intelligence in Higher Education Business	3	2	2	3	10/12, 83.33%,

		Studies Programmes: An Initial Qualitative Exploration.					High
Besharat-Mann, R. (2024)	Qualitative	Can I Trust this Information? Using Adolescent Narratives to Uncover Online Information Seeking Processes.	2	2	1	2	7/12, 58.33%, Good
Bower, M., Torrington, J., Lai, J.W.M., Petocz, P., & Alfano, M. (2024)	Mixed	How should we change teaching and assessment in response to increasingly powerful generative Artificial Intelligence?	3	3	2	3	11/12, 91.67%, High
Caccavale, F., Gargalo, C.L., Gernaey, K.V., & KrÃ¼hne, U. (2024)	Qualitative	Towards Education 4.0: The role of Large Language Models as virtual tutors in chemical engineering.	2	2	1	2	7/12, 58.33%, Good
Cacho, R.M. (2024)	Quantitative	Integrating Generative AI in University Teaching and Learning: A Model for Balanced Guidelines.	3	3	2	3	11/12, 91.67%, High
Cesmeli, A. (2024)	Quantitative	The Metaverse: A Brave New 'World'.	1	1	0	1	3/12, 25.0%, Moderate
Chan, C.K.Y., & Tsi, L.H.Y. (2024)	Mixed	Will generative AI replace teachers in higher education? A study of teacher and student perceptions.	3	3	3	3	12/12, 100.0%, High
Chen, K., Tallant, A.C., & Selig, I. (2024)	Qualitative	Exploring generative AI literacy in higher education: student adoption, interaction, evaluation and ethical perceptions.	2	3	2	2	9/12, 75.0%, High
Daher, W., & Diab, H., & Rayan, A. (2023)	Mixed	Artificial Intelligence Generative Tools and Conceptual Knowledge in Problem Solving in Chemistry.	1	2	1	2	6/12, 50.0%, Good
Daher, W., & Gierdien, F. (2024)	Qualitative	Use of Language by Generative AI Tools in Mathematical Problem Solving: The Case of ChatGPT.	1	2	1	2	6/12, 50.0%, Good
De Jesus, F.S., Ibarra, L.M., & Pasion, B.J., et al. (2024)	Mixed	ChatGPT as an Artificial Intelligence Learning Tool for Business Administration Students in Nueva Ecija, Philippines.	3	3	2	2	10/12, 83.33%, High

DeRise, D. (2023)	Quantitative	Will I Even Teach Writing Anymore? An Examination of First-Year Writing Faculty’s Responsibility to Teach About or With ChatGPT.	3	2	2	3	10/12, 83.33%, High
Du, J., & Alm, A. (2024)	Mixed	The impact of ChatGPT on English for academic purposes (EAP) students’ language learning experience: A self-determination theory perspective.	3	2	2	3	10/12, 83.33%, High
Duah, J.E., & McGivern, P. (2024)	Qualitative	How Generative Artificial Intelligence Has Blurred Notions of Authorial Identity and Academic Norms in Higher Education, Necessitating Clear University Usage Policies.	2	2	2	3	9/12, 75.0%, High
Duong, C.D., Vu, T.N., & Ngo, T.V.N. (2023)	Qualitative	Applying a Modified Technology Acceptance Model to Explain Higher Education Students’ Usage of ChatGPT: A Serial Multiple Mediation Model with Knowledge Sharing as a Moderator.	3	2	2	3	10/12, 83.33%, High
Dúo-Terrón, P. (2024)	Quantitative	Generative Artificial Intelligence: Educational Reflections from an Analysis of Scientific Production.	1	1	1	1	4/12, 33.33%, Moderate
Dwivedi, Y. K., Kshetri, N., Hughes, L., et al. (2023)	Quantitative	“So What if ChatGPT Wrote It?” Multidisciplinary Perspectives on Opportunities, Challenges, and Implications of Generative Conversational AI for Research, Practice, and Policy.	3	2	3	3	11/12, 91.67%, High
Eager, B., & Brunton, R. (2023)	Qualitative	Prompting Higher Education Towards AI-Augmented Teaching and Learning Practice.	2	1	2	2	7/12, 58.33%, Good
Elkhodr, M., Gide, E., Wu, R., & Darwish, O. (2023)	Mixed	ICT Students’ Perceptions Towards ChatGPT: An Experimental Reflective Lab Analysis.	3	3	2	3	11/12, 91.67%, High
Ellis, A.R., & Slade, E. (2023)	Quantitative	A New Era of Learning: Considerations for ChatGPT as a Tool to Enhance Statistics and Data Science Education.	3	3	3	3	12/12, 100.0%, High
Escalante, J., Pack, A., &	Mixed	AI-generated feedback on writing: insights into efficacy	3	3	2	3	11/12,

Barrett, A. (2023)		and ENL student preference.					91.67%, High
Essien, A., Bukoye, O.T., O’Dea, X., & Kremantzis, M. (2024)	Mixed Methods	The Influence of AI Text Generators on Critical Thinking Skills in UK Business Schools.	2	2	1	2	7/12, 58.33%, Good
Farhat, F., Chaudhry, B.M., Nadeem, M., Sohail, S.S., & Madsen, D.Ø. (2024)	Qualitative	Evaluating Large Language Models for the National Premedical Exam in India: Comparative Analysis of GPT-3.5, GPT-4, and Bard.	1	1	0	1	3/12, 25.0%, Moderate
Fatahi, S., Vassileva, J., & Roy, C.K. (2024)	Quantitative	Comparing emotions in ChatGPT answers and human answers to the coding questions on Stack Overflow.	0	1	1	1	3/12, 25.0%, Moderate
Fernández Jiménez, A. (2024)	Mixed	Integration of AI Helping Teachers in Traditional Teaching Roles.	2	2	1	2	7/12, 58.33%, Good
Foung, D., Lin, L., & Chen, J. (2024)	Mixed	Reinventing Assessments with ChatGPT and Other Online Tools: Opportunities for GenAI-Empowered Assessment Practices.	3	3	3	3	12/12, 100.0%, High
Friederichs, H., Friederichs, W.J., & März, M. (2023)	Qualitative	ChatGPT in Medical School: How Successful is AI in Progress Testing?	2	2	2	2	8/12, 66.67%, Good
Gao, Z., Cheah, J.-H., Lim, X.-J., & Luo, X. (2024)	Mixed	Enhancing academic performance of business students using generative AI.	2	2	2	2	8/12, 66.67%, Good
Garib, A., & Coffelt, T.A. (2024)	Qualitative	Detecting the anomalies: Exploring implications of qualitative research in identifying AI-generated text for AI-assisted composition instruction.	2	1	1	2	6/12, 50.0%, Good
Guettala, M., Bourekkache, S., Kazar, O., & Harous, S. (2024)	Qualitative	Generative Artificial Intelligence in Education: Advancing Adaptive and Personalized Learning.	3	3	3	3	12/12, 100.0%, High
Guha, A., Grewal, D., &	Quantitative	Generative AI and Marketing Education: What the	2	2	2	2	8/12,

Atlas, S. (2024)		Future Holds.						66.67%, Good
Haindl, P., & Weinberger, G. (2024)	Quantitative	Students' Experiences of Using ChatGPT in an Undergraduate Programming Course.	3	3	3	3	12/12, 100.0%, High
Hamerman, E.J., Aggarwal, A., & Martins, C. (2024)	Quantitative	An Investigation of Generative AI in the Classroom and its Implications for University Policy.	2	2	2	2	8/12, 66.67%, Good
Hao, J., von Davier, A.A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D.J. (2024)	Mixed Methods	Transforming Assessment: The Impacts and Implications of Large Language Models and Generative AI.	3	3	2	3	11/12, 91.67%, High
Hashem, O. A., & Hakeem, M. B. (2024)	Quantitative	Design Education Methodology Using AI.	2	2	1	2	7/12, 58.33%, Good
Hmoud, M., Swaity, H., Hamad, N., Karram, O., & Daher, W. (2024)	Qualitative	Higher Education Students' Task Motivation in the Generative Artificial Intelligence Context: The Case of ChatGPT.	3	2	1	2	8/12, 66.67%, Good
Hostetter, A.B., Call, N., Frazier, G., James, T., Linnertz, C., Nestle, E., & Tucci, M. (2024)	Qualitative	Student and Faculty Perceptions of Generative Artificial Intelligence in Student Writing.	3	3	3	3	12/12, 100.0%, High
Iatrellis, O., Samaras, N., Kokkinos, K., & Panagiotakopoulos, T. (2024)	Quantitative	Leveraging Generative AI for Sustainable Academic Advising: Enhancing Educational Practices through AI-Driven Recommendations.	2	2	1	2	7/12, 58.33%, Good
Ilieva, G., Yankova, T., Klisarova-Belcheva, S., Dimitrov, A., Bratkov, M., & Angelov, D. (2023)	Mixed	Effects of Generative Chatbots in Higher Education.	3	3	2	3	11/12, 91.67%, High
Ilieva, G., Yankova, T., Klisarova-Belcheva, S., Dimitrov, A., Bratkov, M., &	Mixed	Effects of generative chatbots in higher education.	2	2	2	2	8/12, 66.67%, Good

Angelov, D. (2023)								
Iliyasu, Z., Abdullahi, H.O., Iliyasu, B.Z., Bashir, H.A., Amole, T.G., & Abdullahi, H.M. (2024)	Qualitative	C Correlates of Medical and Allied Health Students' Engagement with Generative AI in Nigeria.	2	2	1	2	7/12, 58.33%, Good
Imran, M., Almusharraf, N., Abdellatif, M.S., & Abbasova, M.Y. (2024)	Mixed	Artificial Intelligence in Higher Education: Enhancing Learning Systems and Transforming Educational Paradigms.	2	2	1	2	7/12, 58.33%, Good
Issakov, Y., Omarov, K., Savanchiyeva, A., Kadyrbekova, D., et al. (2024)	Quantitative	Determining the Effectiveness of Using ChatGPT-4 in Organising Excursions.	1	2	2	2	7/12, 58.33%, Good
Ivanov, S., Soliman, M., Tuomi, A., Alkathiri, N.A., & Al-Alawi, A.N. (2024)	Quantitative	Drivers of Generative AI Adoption in Higher Education through the Lens of the Theory of Planned Behaviour.	2	2	1	2	7/12, 58.33%, Good
Jaboob, M., Hazaimeh, M., & Al-Ansi, A.M. (2024)	Qualitative	Integration of Generative AI Techniques and Applications in Student Behavior and Cognitive Achievement in Arab Higher Education.	2	2	1	2	7/12, 58.33%, Good
Jackaria, P.M., Hajan, B.H., & Mastul, A.-R.H. (2024)	Qualitative	A Comparative Analysis of the Rating of College Students' Essays by ChatGPT versus Human Raters.	2	3	3	3	11/12, 91.67%, High
Jadhav, S., Vhatkar, S., & Aalam, Z. (2024)	Qualitative	Bridging the Gap: Exploring the Revolutionary Application of GenAI in Language Teaching and Learning.	3	2	2	3	10/12, 83.33%, High
Jensen, L.X., Buhl, A., Sharma, A., & Bearman, M. (2024)	Qualitative	Generative AI and Higher Education: A Review of Claims from the First Months of ChatGPT.	3	3	3	3	12/12, 100.0%, High
Jeon, J., & Lee, S. (2023)	Mixed	Large Language Models in Education: A Focus on the Complementary Relationship between Human Teachers and ChatGPT.	3	2	2	3	10/12, 83.33%, High
Jho, H., & Ha, M. (2024)	Quantitative	Towards Effective Argumentation: Design and Implementation of a Generative AI-Based Evaluation and Feedback System.	2	3	2	2	9/12, 75.0%, High

Jošt, G., Taneski, V., & Karakatič, S. (2024). (2024)	Mixed	The Impact of Large Language Models on Programming Education and Student Learning Outcomes.	3	3	2	3	11/12, 91.67%, High
Jochim, J., & Lenz-Kesekamp, V.K. (2024)	Qualitative	Teaching and Testing in the Era of Text-Generative AI: Exploring the Needs of Students and Teachers.	3	3	3	3	12/12, 100.0%, High
Johnson, N., Seaman, J., & Seaman, J. (2024)	Mixed	The Anticipated Impact of Artificial Intelligence on Higher Education.	2	2	1	2	7/12, 58.33%, Good
Johnston, H., Wells, R.F., Shanks, E.M., Boey, T., & Parsons, B.N. (2024)	Qualitative	Student Perspectives on the Use of Generative Artificial Intelligence Technologies in Higher Education.	3	3	2	3	11/12, 91.67%, High
Kakhki, M. D., Oguz, A., & Gendron, M. (2024)	Qualitative	Exploring the Affordances of Chatbots in Higher Education: A Framework for Understanding and Utilizing ChatGPT.	3	2	2	3	10/12, 83.33%, High
Kanont et al. (2024)	Quantitative	Generative-AI, a Learning Assistant? Factors Influencing Higher-Ed Students' Technology Acceptance.	2	2	1	2	7/12, 58.33%, Good
Kaplan-Rakowski et al. (2023)	Qualitative	Generative AI and Teachersâ€™ Perspectives on Its Implementation in Education.	3	2	2	3	10/12, 83.33%, High
Kazanidis, I., & Pellas, N. (2024)	Quantitative	Harnessing Generative Artificial Intelligence for Digital Literacy Innovation: A Comparative Study between Early Childhood Education and Computer Science Undergraduates.	3	3	2	3	11/12, 91.67%, High
Kim et al. (2024)	Mixed	Exploring studentsâ€™ perspectives on Generative AI-assisted academic writing.	3	2	2	3	10/12, 83.33%, High
Kong et al. (2024)	Qualitative	A pedagogical design for self-regulated learning in academic writing using text-based generative artificial intelligence tools.	3	3	2	3	11/12, 91.67%, High
Kumar, S., et al. (2024)	Mixed	The Role of Instructional Designers in the Integration of	2	2	1	2	7/12,

		Generative Artificial Intelligence in Online and Blended Learning in Higher Education.					58.33%, Good
Lai, J.W. (2024)	Mixed	Adapting Self-Regulated Learning in an Age of Generative Artificial Intelligence Chatbots.	3	3	2	3	11/12, 91.67%, High
Lang, G., Triantoro, T., & Sharp, J.H. (2024)	Qualitative	Large Language Models as AI-Powered Educational Assistants: Comparing GPT-4 and Gemini for Writing Teaching Cases.	3	2	2	3	10/12, 83.33%, High
Lee, D., Arnold, M., et al. (2024)	Qualitative	The Impact of Generative AI on Higher Education Learning and Teaching: A Study of Educatorsâ€™ Perspectives.	3	3	2	3	11/12, 91.67%, High
Lee, G.-G., & Zhai, X. (2024)	Mixed	Using ChatGPT for Science Learning: A Study on Pre-service Teachers' Lesson Planning.	2	2	1	2	7/12, 58.33%, Good
Lodge, J.M., Thompson, K., et al. (2023)	Mixed	Mapping out a Research Agenda for Generative Artificial Intelligence in Tertiary Education.	3	2	2	3	10/12, 83.33%, High
Magrill, J., & Magrill, B. (2024)	Qualitative	Preparing Educators and Students at Higher Education Institutions for an AI-Driven World.	2	2	2	2	8/12, 66.67%, Good
Mah, C., et al. (2024)	Qualitative	Beyond CheatBots: Examining Tensions in Teachersâ€™ and Studentsâ€™ Perceptions of Cheating and Learning with ChatGPT.	3	3	2	3	11/12, 91.67%, High
Markos, A., Prentzas, J., & Sidiropoulou, M. (2024)	Qualitative	Pre-Service Teachersâ€™ Assessment of ChatGPT’s Utility in Higher Education: SWOT and Content Analysis.	2	2	2	2	8/12, 66.67%, Good
Mayer, C. (2024)	Qualitative	Thriving in an AI-Dominated World: Why Higher Education Must Produce Graduates who are Uniquely Human and Technically Competent.	3	2	2	3	10/12, 83.33%, High
Mendez, J.D. (2024)	Mixed Methods	Student Perceptions of Artificial Intelligence Utility in the Introductory Chemistry Classroom.	2	2	2	2	8/12, 66.67%, Good

Ngo, T.T.A., et al. (2024)	Mixed Methods	ChatGPT for Educational Purposes: Investigating the Impact of Knowledge Management Factors on Student Satisfaction and Continuous Usage.	3	3	2	3	11/12, 91.67%, High
Nguyen, A., et al. (2024)	Mixed	Human-AI Collaboration Patterns in AI-Assisted Academic Writing.	3	3	2	3	11/12, 91.67%, High
Nikolic, S., et al. (2023)	Qualitative	ChatGPT Versus Engineering Education Assessment: A Multidisciplinary and Multi-Institutional Benchmarking and Analysis.	2	2	2	2	8/12, 66.67%, Good
Ning, Y., et al. (2024)	Mixed	Teachersâ€™ AI-TPACK: Exploring the Relationship between Knowledge Elements.	2	2	1	2	7/12, 58.33%, Good
Ogunleye, B., et al. (2024)	Quantitative	Higher Education Assessment Practice in the Era of Generative AI Tools.	2	2	1	2	7/12, 58.33%, Good
Omar, A., Shaqour, A.Z., & Khlaif, Z.N. (2024)	Qualitative	Attitudes of Faculty Members in Palestinian Universities Toward Employing Artificial Intelligence Applications in Higher Education.	2	2	1	2	7/12, 58.33%, Good
Omeh, C.B., et al. (2024)	Qualitative	Application of Artificial Intelligence (AI) Technology in TVET Education: Ethical Issues and Policy Implementation.	2	2	2	2	8/12, 66.67%, Good
Orrù, G., et al. (2023)	Qualitative	Human-Like Problem-Solving Abilities in Large Language Models Using ChatGPT.	2	2	2	2	8/12, 66.67%, Good
Oster, N., Henriksen, D., & Mishra, P. (2024)	Mixed	ChatGPT for Teachers: Insights from Online Discussions.	3	2	2	3	10/12, 83.33%, High
Pack, A., & Maloney, J. (2023)	Quantitative	Using Generative Artificial Intelligence for Language Education Research: Insights from Using OpenAI's ChatGPT.	2	2	2	2	8/12, 66.67%, Good
Pack, A., Barrett, A., & Escalante, J. (2024)	Qualitative	Large Language Models and Automated Essay Scoring of English Language Learner Writing: Insights into	3	3	2	3	11/12, 91.67%,

		Validity and Reliability.					High
Perkins, M., & Roe, J. (2024)	Meta-Analysis	Detection of GPT-4 Generated Text in Higher Education: Combining Academic Judgement and Software to Identify Generative AI Tool Misuse.	3	2	2	3	10/12, 83.33%, High
Pham, T., Nguyen, B., Ha, S., & Ngoc, T.N. (2023)	Qualitative	Digital Transformation in Engineering Education: Exploring the Potential of AI-Assisted Learning.	2	2	1	2	7/12, 58.33%, Good
Picciano, A.G. (2024)	Mixed	Graduate Teacher Education Students Use and Evaluate ChatGPT as an Essay-Writing Tool.	3	3	2	3	11/12, 91.67%, High
Qadir, J. (2024)	Quantitative	Learning 101 Reloaded: Revisiting the Basics for the GenAI Era.	2	2	1	2	7/12, 58.33%, Good
Qasem, F. (2023)	Qualitative	ChatGPT in Scientific and Academic Research: Future Fears and Reassurances.	3	2	2	3	10/12, 83.33%, High
Raman, R., et al. (2024)	Mixed	Exploring University Students' Adoption of ChatGPT Using the Diffusion of Innovation Theory and Sentiment Analysis with Gender Dimension.	2	2	1	2	7/12, 58.33%, Good
Reddy, M.R., et al. (2024)	Qualitative	Implementation and Evaluation of a ChatGPT-Assisted Special Topics Writing Assignment in Biochemistry.	3	3	2	3	11/12, 91.67%, High
Ruiz-Rojas, L.I., et al. (2024)	Qualitative	Collaborative Working and Critical Thinking: Adoption of Generative Artificial Intelligence Tools in Higher Education.	3	3	2	3	11/12, 91.67%, High
SÃ¡ez-Velasco, et al. (2024)	Qualitative	Analysing the Impact of Generative AI in Arts Education: A Cross-Disciplinary Perspective of Educators and Students in Higher Education.	2	2	1	2	7/12, 58.33%, Good
Saihi, et al. (2024)	Qualitative	A Structural Equation Modeling Analysis of Generative AI Chatbots Adoption Among Students and Educators in Higher Education.	2	3	2	3	10/12, 83.33%, High
Salinas-Navarro, J., Sampaio,	Qualitative	Using Generative Artificial Intelligence Tools to Explain	3	3	2	3	11/12,

B., et al. (2024)		and Enhance Experiential Learning for Authentic Assessment.					91.67%, High
Sampaio, B., Salinas-Navarro, J., et al. (2024)	Mixed	Impacts of Generative Artificial Intelligence in Higher Education: Research Trends and Studentsâ€™ Perceptions.	2	2	2	2	8/12, 66.67%, Good
Sharples, M. (2023)	Qualitative	Towards Social Generative AI for Education: Theory, Practices, and Ethics.	3	2	2	3	10/12, 83.33%, High
Smerdon, D. (2024)	Commentary	AI in Essay-Based Assessment: Student Adoption, Usage, and Performance.	3	2	2	3	10/12, 83.33%, High
Soliman, M., et al. (2024)	Qualitative	Modelling Continuous Intention to Use Generative Artificial Intelligence as an Educational Tool Among University Students.	3	2	2	3	10/12, 83.33%, High
Steele, J.L. (2023)	Qualitative	To GPT or Not GPT? Empowering Our Students to Learn with AI.	2	2	1	2	7/12, 58.33%, Good
Sullivan, M., et al. (2023)	Qualitative	ChatGPT in Higher Education: Considerations for Academic Integrity and Student Learning.	3	2	2	3	10/12, 83.33%, High
Sun, L., et al. (2024)	Qualitative	Does Generative Artificial Intelligence Improve the Academic Achievement of College Students? A Meta-Analysis.	3	3	2	3	11/12, 91.67%, High
Tam, T., et al. (2023)	Mixed Methods	Nursing Education in the Age of Artificial Intelligence-Powered Chatbots (AI-Chatbots): Are We Ready Yet?	2	2	1	2	7/12, 58.33%, Good
Thanh, T., et al. (2023)	Quantitative	Race with the Machines: Assessing the Capability of Generative AI in Solving Authentic Assessments.	2	2	2	2	8/12, 66.67%, Good
Tiwari, S., et al. (2024)	Quantitative	What Drives Students Toward ChatGPT?	3	2	2	3	10/12, 83.33%, High

Tlili, A., et al. (2023)	Qualitative	What if the Devil Is My Guardian Angel: ChatGPT as a Case Study of Using Chatbots in Education.	3	2	2	3	10/12, 83.33%, High
Tupper, S., et al. (2024)	Qualitative	Field Courses for Dummies: To What Extent Can ChatGPT Design a Higher Education Field Course?	2	2	1	2	7/12, 58.33%, Good
Tzeng, J. Y. (2023)	Qualitative	Ignorance-oriented Instruction for Future Learning: Principles and Practices.	2	2	1	2	7/12, 58.33%, Good
Valova, A., et al. (2024)	Quantitative	Students' Perception of ChatGPT Usage in Education.	3	2	2	3	10/12, 83.33%, High
van den Berg, J., & du Plessis, M. (2023)	Mixed	ChatGPT and Generative AI: Possibilities for Its Contribution to Lesson Planning, Critical Thinking and Openness in Teacher Education.	2	2	1	2	7/12, 58.33%, Good
Van Wyk, R. (2024)	Mixed Methods	Is ChatGPT an Opportunity or a Threat? Preventive Strategies Employed by Academics Related to a GenAI-Based LLM at a Faculty of Education.	2	2	1	2	7/12, 58.33%, Good
VankÃºÅ¡, M. (2024)	Quantitative	Generative Artificial Intelligence on Mobile Devices in the University Preparation of Future Teachers of Mathematics.	2	2	1	2	7/12, 58.33%, Good
Vosevich, K., & Hutson, J. (2024)	Mixed	Absent Presence: The Human Influence in AI-Generated Content in the Age of Technoculture.	2	2	1	2	7/12, 58.33%, Good
Wölfel, M., Shirzad, M. B., Reich, A., & Anderer, K. (2023)	Quantitative	Knowledge-Based and Generative-AI-Driven Pedagogical Conversational Agents: A Comparative Study of Grice's Cooperative Principles and Trust.	2	2	1	2	7/12, 58.33%, Good
Waluyo, B., & Kusumastuti, S. (2024)	Qualitative	Generative AI in Student English Learning in Thai Higher Education: More Engagement, Better Outcomes?	3	2	2	3	10/12, 83.33%, High
Wang, C. (2024)	Mixed	Exploring Students' Generative AI-Assisted Writing Processes: Perceptions and Experiences from Native and	3	2	2	3	10/12, 83.33%,

		Nonnative English Speakers.						High
Wang, D., et al. (2024)	Qualitative	ChatGPT or Bert? Exploring the Potential of ChatGPT to Facilitate Preservice Teachersâ€™ Learning of Dialogic Pedagogy.	2	2	1	2	7/12, 58.33%, Good
Wang, K., Ruan, Q., Zhang, X., Fu, C., & Duan, B. (2024)	Quantitative	Pre-Service Teachersâ€™ GenAI Anxiety, Technology Self-Efficacy, and TPACK: Their Structural Relations with Behavioral Intention to Design GenAI-Assisted Teaching.	3	3	2	3	11/12, 91.67%, High
Waring, D. (2024)	Qualitative	Artificial Intelligence and Graduate Employability: What Should We Teach Generation AI?	3	2	2	3	10/12, 83.33%, High
Wood, A., & Moss, B. (2024)	Mixed	Evaluating the Impact of Students' Generative AI Use in Educational Contexts.	3	2	2	3	10/12, 83.33%, High
Wu, D., Zhang, S., Ma, Z., Yue, X. G., & Dong, R. K. (2024)	Qualitative	Unlocking Potential: Key Factors Shaping Undergraduate Self-Directed Learning in AI-Enhanced Educational Environments.	3	2	2	3	10/12, 83.33%, High
Xing, Y. (2024)	Mixed	Exploring the Use of ChatGPT in Learning and Instructing Statistics and Data Analytics.	3	2	2	3	10/12, 83.33%, High
Yan, W., Nakajima, T., & Sawada, R. (2024)	Qualitative	Benefits and Challenges of Collaboration between Students and Conversational Generative Artificial Intelligence in Programming Learning: An Empirical Case Study.	3	2	2	3	10/12, 83.33%, High
Yang, S., Dong, Y., & Yu, Z. G. (2024)	Mixed	ChatGPT in Education: Ethical Considerations and Sentiment Analysis.	2	2	1	2	7/12, 58.33%, Good
Yang, X., Wang, Q., & Lyu, J. (2024)	Mixed	Assessing ChatGPT’s Educational Capabilities and Application Potential.	2	2	1	2	7/12, 58.33%, Good
Yeralan, S., & Lee, L. A. (2023)	Mixed	Generative AI: Challenges to higher education.	2	2	1	2	7/12, 58.33%,

							Good
Yilmaz, R., & Karaoglan Yilmaz, F.G. (2023)	Qualitative	The Effect of Generative Artificial Intelligence (AI)-Based Tool Use on Students' Computational Thinking Skills, Programming Self-Efficacy, and Motivation.	3	3	2	3	11/12, 91.67%, High
Zhai, X., Nyaaba, M., & Ma, W. (2024)	Qualitative	Can Generative AI and ChatGPT Outperform Humans on Cognitive-Demanding Problem-Solving Tasks in Science?	3	3	2	3	11/12, 91.67%, High

Appendix II: Key Findings of GenAI to Support Student Success in HE

Theme	**Sub-theme**	**Description**	**Representative Authors**
Ethical and Responsible Use of GenAI	Data Privacy and Security Paradoxes	Emphasises the importance of safeguarding data privacy and instituting robust security protocols to build student trust and prevent data misuse.	Aad & Hardey (2024)
	Accountability and Transparency of Ethical GenAI Applications	Advocates for transparent GenAI operations, particularly in assessments, to foster trust and ensure students understand AI-generated outputs.	Hostetter et al. (2024); Jensen et al. (2024)
	Bias, Fairness and Equitable GenAI Frameworks	Highlights the need for fairness-oriented AI to mitigate potential biases, ensuring equitable learning outcomes for all students.	Adel et al. (2024); Asad et al. (2024)
	Human-Computer Interaction Balance and Collaborative AI-Educator Models	Discusses the importance of balancing AI capabilities with human oversight, enhancing the relational dynamics between students and educators.	Akpan et al. (2024); Sharples (2023); Hostetter et al. (2024)
	Ethical Dilemmas in AI Decision-Making	Explores ethical concerns around bias, autonomy, and fairness in AI-driven decision-making and the need for responsible institutional frameworks.	Chen et al. (2024); Adarkwah (2024); Guettala et al. (2024)
GenAI for Skill and Workforce Preparation	Developing Digital and AI Literacy Skills for Modern Careers	Positions GenAI as a tool for building foundational digital and AI skills critical for future workforce readiness.	Adarkwah (2024); Chen et al. (2024)
	Fostering Students’ Critical and Analytical	Explores how GenAI fosters critical thinking and data-	Baek et al. (2024); Ellis

	Skills for Data-Driven Decision-Making	driven decision-making, essential for tech-driven careers.	& Slade (2023)
	Promoting Adaptability, Lifelong Learning and Essential Skills for a Dynamic Workforce	Emphasises GenAI’s role in developing adaptability and continuous learning skills needed in a dynamic job market.	Adel et al. (2024); Guettala et al. (2024)
	Enhancing Domain-Specific Competencies and Tailoring GenAI Applications to Disciplinary Needs	Demonstrates GenAI’s versatility in providing field-specific training across disciplines such as healthcare and engineering.	Akpan et al. (2024); Iliyasu et al. (2024); Lee et al. (2024)
	Over-reliance on AI for Faculty Decision-Making	Raises concerns about over-reliance on GenAI in faculty decision-making, potentially undermining human insight and student needs.	Guettala et al. (2024); Imran et al. (2024)
	Faculty AI Integration Challenges	Highlights barriers to effective GenAI integration due to limited faculty training, inconsistent use, and lack of support systems.	Imran et al. (2024); Jho & Ha (2024)
Enhancing Critical Thinking and Engagement	Promoting GenAI as a Tool for Critical Reflective Inquiry	Illustrates how GenAI encourages students to critically engage with and evaluate AI-generated insights, fostering independent analysis.	Adel et al. (2024); Ellis & Slade (2023)
	Supporting Active, Sustained Engagement and Transforming Learning into a Participatory Process	Describes GenAI’s interactive nature, which sustains student engagement by transforming learning into a participatory process.	Baek et al. (2024); Sharples (2023)
	Stimulating Innovation and Exploration in Learning for Creative Problem-Solving	Highlights GenAI’s role in promoting innovation by enabling students to explore diverse approaches to problem-solving.	Adarkwah (2024); Ajevski et al. (2023)
	Tailoring Educational Pathways to Individual Needs for Personalised Learning Experiences	Examines GenAI’s capacity to tailor content and feedback to each student’s learning needs, promoting self-directed learning.	Akpan et al. (2024); Chen et al. (2024)
	Reduced Human-Led Discussions	Examines how increased GenAI use may reduce traditional human-led interactions essential for collaborative and deep learning.	Jeon & Lee (2023); Cacho (2024)
	AI-Induced Intellectual Passivity	Warns of intellectual passivity resulting from over-reliance on AI, undermining critical thinking and academic development.	Bai et al. (2024); Akpan et al. (2024)
Equitable Access and	Ensuring Accessibility for Underrepresented	Stresses the need for adaptive GenAI features to make	Aad & Hardey (2024);

Inclusivity in GenAI Tools	Groups and Reducing Barriers to Inclusive Education	learning inclusive for students with diverse needs.	Chen et al. (2024)
	Language and Cultural Biases in GenAI	Emphasises the importance of inclusive GenAI design that respects linguistic and cultural diversity to foster equitable learning experiences.	Bai et al. (2024); Asad et al. (2024)
	Digital Equity and Divide in GenAI Access	Calls for institutional policies to bridge the digital divide, ensuring all students can access AI-supported learning tools.	Baek et al. (2024); Adarkwah (2024)
	Developing Inclusive Design and Usability for Diverse Demographics	Advocates for user-friendly, accessible GenAI interfaces developed with input from diverse demographic groups.	Ellis & Slade (2023); Chen et al. (2024)
	AI-Driven Discrimination in Academic Settings (Algorithmic-Bias)	Explores risks of algorithmic bias in GenAI, which can reinforce stereotypes and disadvantage underrepresented students.	Ajevski et al. (2023); Escalante et al. (2023)
Assessment and Feedback Enhancement through GenAI	Automated Assessment, Grading and Timely Feedback	Highlights GenAI's efficiency in automating grading, providing immediate feedback to students and reducing educator workload.	Adel et al. (2024); Ellis & Slade (2023)
	Personalised Feedback for Supporting Student Growth	Describes GenAI's ability to provide feedback tailored to each student's strengths and weaknesses, enhancing learning outcomes.	Bai et al. (2024); Sharples (2023)
	Facilitating Ongoing Monitoring and Feedback Loops	Explores GenAI's role in providing continuous feedback, enabling instructors to monitor and adjust instruction based on student progress.	Ballantine et al. (2024); Cacho (2024)
	Promoting Self-Regulated, Independent Learning and Metacognitive Skills	Highlights how GenAI fosters autonomous learning by encouraging students to self-assess and track their progress.	Benuyenah & Dewnarain (2024); Adarkwah (2024)
	Over-reliance on AI for Assessments	Explores how over-reliance on GenAI for assessments may hinder deep learning and academic integrity.	Adel et al. (2024); Asad et al. (2024)
	Risk of AI Misinterpretation in Feedback	Raises concerns about AI misinterpreting student inputs and delivering inaccurate or misleading feedback.	Adel et al. (2024); Hostetter et al. (2024)
Language and Writing Support via GenAI	Providing Assistance in Grammar and Technical Precision in Writing	Examines how GenAI tools help students improve writing accuracy through immediate grammar and style feedback.	Adel et al. (2024); Cacho (2024)
	Enhancing Vocabulary and Language	Describes GenAI's ability to enrich student vocabulary	Akpan et al. (2024);

	Diversity for Expanding Lexical Repertoires	and encourage diverse language use, supporting advanced writing.	Baek et al. (2024)
	Supporting Academic Writing Skills, Structuring Arguments and Enhancing Cohesion	Explores how GenAI assists students in structuring arguments, improving cohesion, and adhering to academic conventions.	Adarkwah (2024); Ellis & Slade (2023)
	Facilitating Multilingual and Translation Support and Enhancing Inclusivity for Diverse Linguistic Backgrounds	Highlights GenAI’s multilingual features that aid non-native speakers and support cross-linguistic understanding.	Omar et al. (2024); Jho & Ha (2024)
	Over-dependence on AI Writing Tools	Examines how overuse of GenAI tools can reduce independent learning and mask gaps in comprehension.	Baek et al. (2024); Benuyenah & Dewnarain (2024)
	AI Bias in Language and Translation	Identifies risks of biased or inaccurate translations that can misrepresent students' work and reinforce exclusion.	Akpan et al. (2024); Adel et al. (2024)

Moral Judgment and Generative AI in the Creative Industries

Sian Joel-Edgar, Yu-Chun Pan & Alice Helliwell
Northeastern University London, UK
sian.joel-edgar@nulondon.ac.uk; y.pan@nulondon.ac.uk;
alice.helliwell@nulondon.ac.uk

Abstract

This paper combines literature analysis and focus group discussion to explore the role of moral judgments in generative AI use within the creative industries, focusing on whether professionals perceive it as morally acceptable. Utilising literature to inform our focus group questions, we sought to understand how creatives use generative AI, if at all, the ethical barriers to adoption, the perceptions of the generative AI creative output and the broader implications of generative AI use. In our focus group, we found there was a range of generative AI use cases, and how it was used had a bearing on whether it was deemed morally acceptable. Ethical barriers ranged from individual moral objections (e.g. a sense of copying others' work), collective moral objections (negative impact on creative education and industry as a whole), and broader ethical concerns about energy usage. The qualitative analysis and literature review have helped to form a theoretical framework which we aim to empirically test.

Keywords: Artificial Intelligence, Creativity, Morality, Ethics, Creative

1.0 Introduction

The use of Artificial Intelligence (AI) has become integral to everyday work practices. In the creative industries, it has now become omnipresent in the creative process and in creative output (Anantrasirichai, Nantheera & Bull, 2022). It can be seen as a tool, a collaborator, or an innovator. In the production of art, design, literature, and media, the integration of generative AI technologies like ChatGPT has significant potential to transform creative processes, enabling new forms of expression, and enhancing artistic collaborations. Interactions between creators and AI systems can facilitate innovative ideas, generate unique content, and open up new creative possibilities.

AI has the potential for introducing new ingenious approaches, processes and outputs but it also challenges notions of human originality, creativity and creative agency (both individual and as part of a collective). According to Puntoni et al. (2020), the use of AI technologies are often perceived as a neutral tool however the use of AI can stir emotional responses. These include feelings of exploitation from personal data collection, the sense of well-being through personalisation, and concerns about self-integrity.

Existing research on morality and AI adoption in creativity has looked at personal values and their role in technology adoption (Hemingway and Maclagan, 2004), however, there is fragmented literature on moral acceptance and its role in disclosing the use of AI. Moreover, there is limited literature on the role of a person's perceived ethical stance and moral acceptance for both the creative's own moral agency and their identity as part of a creative community and the impact that AI adoption would have on the creative industry as a whole. In the following paper, we review the literature on AI in art and design and the theoretical concepts of morality. This literature has in turn informed a question bank we posed to creative industry professionals in a focus group. The results of the literature and the focus group have informed a theoretical framework that we hope to build upon to explore the important concept of whether a creative individual feels it is ethical to create with generative AI and what mechanisms can support an ethical stance.

2.0 Literature Review

2.1 AI in Art and Design

AI has been used in many different ways in the creative process and its role has been hotly discussed by creative scholars. Ivcevic and Grandinetti (2024) have described how AI has been used as the 4 Cs. Mini Creativity and supporting the learning process such as "Poetry Machine," an AI tool designed to teach poetry writing to secondary school students (Kangasharju, 2022). Little Creativity and the enhancement of creative tasks often to augment Human-AI Co-Creation. For example, DragonflyAI markets itself to use AI for creative collaboration such as proposing novel combinations of colour or appealing design trends based on its vast dataset.[1] Pro Creativity which supports the creation of new artwork such as Refik Anadol's AI-driven art installations. Big Creativity which involves transforming the domain, for example the use of SORA for XR and Adobe adding SORA to its suite of creative tools (Weatherbed, 2024). Additionally, they discuss the role that AI plays in process optimisation and the organisational aspects of creative processes.

Generative AI, a field within AI, is responsible for generating fresh and potentially unique content (van Dis et al., 2023). It can be viewed as both a creative and rational tool, depending on its usage and the surrounding circumstances. In November 2022,

[1] See https://dragonflyai.co/.

OpenAI introduced ChatGPT, which swiftly garnered acclaim for its innovative approach to generating AI-based content (Dwivedi et al, 2023). ChatGPT provides unique text in response to user queries by harnessing a huge collection of textual data. The outputs closely mimic human-generated content. There has been widespread usage of ChatGPT in a variety of fields, such as research (Joel-Edgar and Pan, 2024), software development, poetry, essays, corporate communication, and legal agreements (Zhuo et al, 2023). Scholars from a variety of fields have expressed interest in this, and the general public has started discussions on the implications of ChatGPT and generative AI, looking at both the potential advantages and potential disadvantages (Dwivedi et al, 2023).

Given the growing popularity of AI tools such as LLMs and their sophisticated natural language processing capabilities for activities like text synthesis, language translation, and answering inquiries in a variety of creative contexts (e.g. writing content for the Guardian newspaper - Pavlik, 2023), it is crucial to examine perceptions of this emerging technology. The successful integration of AI in creative industries relies on professionals accepting and understanding its potential (Mogavi et al, 2023). For creative industry professionals, they are instrumental in developing synergies between AI and creative output.

ChatGPT and other LLMs are not the only generative AI tool that is not being used in creative industries. AI image generators, including Generative Adversarial Networks (GANs) and text-to-image models such as OpenAI’s DALL-E, Stable Diffusion, and Midjourney (e.g. see Anantrasirichai & Bull, 2002; Danesi, 2024) are being used to generate images in creative industries, either used whole-sale, or used as part of a broader creative process. A small number of practising artists have been working with AI tools for several years),[2] though this has now expanded to a wide array of creative professionals, with generative AI now built into commonly used creative tools like Adobe Photoshop.[3] We are also seeing an increase in the use of AI tools for music generation, which is making its way into the music industry (Zhang, Yan & Briot, 2023; Zhu et al., 2023). There is a proliferation of generative AI songwriting tools, to complement the growing use of AI technology for other audio tasks (Henkin, 2023).

[2] See e.g. https://aiartists.org/

[3] See https://helpx.adobe.com/uk/acrobat/using/ai-summaries-acrobat-home.html

The work proposed here builds on previous studies that sought to understand the role of AI in creative industries (Chowdhury et al, 2022). However, the work proposed looks specifically at the adoption of newer generative AI tools. The huge potential offered by LLMs warrants specific and in-depth research, however, due to its relative novelty there is a scarcity of studies in the current literature concerning generative AI adoption in regard to creative industries professionals (Qadir, et al, 2023).

The creative industry relies on creativity, originality, and intellectual property, encompassing sectors that create, produce, and distribute creative goods and services and has been predominantly seen as a human endeavour (de Cock Buning, 2018). The growing use of generative AI in this industry raises many questions regarding output quality, job market impact, biases in training data, intellectual property rights, copyright issues, and ethics. Professionals in the creative industry face the challenge of balancing automation and creativity. Though generative AI automates some portions of the creative process, freeing up time for more difficult activities, excessive automation runs the danger of eroding creativity and originality, values that are crucial in these businesses

The adoption of emerging technologies like generative AI is a subject of extensive research. Factors such as ease of use, perceived usefulness, trust, security, personal innovativeness, compatibility, and moral judgement influence users' intentions to adopt new technologies (Venkatesh, 2003). The impact of these factors varies based on the technology and user group, emphasising the complex dynamics involved. In the creative industry, where originality is paramount, concerns grow regarding the level of technical assistance and the preservation of ownership over work assisted by generative AI. Moral judgments, shaped by cultural, social, religious, and personal factors, also influence individuals' perception of "originality" and their willingness to embrace generative AI (Inie, Falk and Tanimoto, 2023).

2.2 Moral and Ethical Implications

We are interested here in how creative professionals make moral judgements about AI use in creative practices. Moral judgements assess the moral correctness of actions, behaviour, intentions, or outcomes based on personal or societal principles (Reynolds and Ceranic, 2007). These judgments reflect one's moral convictions and guide ethical decision-making and interactions with others (Jennings et al 2015; Malle 2021). They provide a framework for evaluating the consequences, fairness, justice, and ethical implications of actions (Sullivan and Wamba, 2022). According to Malle (2021), moral judgement encompasses evaluations, norm judgments, moral wrongness judgments, and blame judgments. In the proposed research, the constructs of each of these four elements will be explored in the perceptions of creative industry professionals.

Moral judgement is the process in which people make decisions about what is right and wrong based on reasoning and value (Kohlberg and Hersh, 1977). Kohlberg and Power's (1981) work on moral development builds on that of Piaget (1932) in which he emphasises reasoning behind moral choices. While Kohlberg focused on justice based reasoning centred on abstract rules and universal principles, Gilligan (1982) highlighted the ethics of care, which prioritises empathy, relational context, and the responsibilities individuals have toward others. In our research, we apply Kohlberg's justice-based reasoning and Gilligan's ethics of care to creative industry professionals to allow us to explore how they balance abstract principles like intellectual property rights with relational dynamics such as collaboration and cultural sensitivity in their moral judgments.

Credibility has been described from a source credibility perspective as the perception of a communicator's trustworthiness, expertise and attractiveness (Hovland, Janis and Kelly, 1953). Each of these aspects in turn can be defined. Trustworthiness can be described as a source that is reliable, honest and unbiased. Expertise can be described as the source being accurate and knowledgeable. Attractiveness (McCroskey and Teven, 1999) has been described as how appealing or engaging the source is. Source credibility theory is widely used to understand how people form opinions and make decisions about the credibility of the source information often in media and advertising. We apply it to our study to understand the credibility of the creative output.

Creativity can be considered in terms of creative persons, processes, "press" (relating to the environment) and product (Rhodes, 1961). In terms of the processes that are undertaken in creative industries, creativity can be described as socially constructed (Vygotsky, 1978), shaped and influenced by social interactions with others (Sawyer, 2007; Burr 1995). The creative product is formed through a network of collaborators, who provide feedback, instigate ideas, create elements of the overall artwork and provide validation (Csikszentmihalyi, 1990). AI can be seen as fitting within the social construction as outlined by Vygotsky (1978) and Burr (1995). AI may have the potential to act as a co-creator in the creative process, enhancing and supporting creative output (McCormack et al., 2019). By providing insights drawn from vast datasets, offering feedback, and suggesting ideas, AI can assist creatives in generating and refining concepts within a social and iterative framework (Runco & Jaeger, 2012). Although AI can be seen as a beneficial collaborator, and as complementary to human creativity (Amabile, 1996), there have been moral objections to its use. Brynjolfsson and McAfee (2014) for example, highlight AI overshadowing human aspects of creativity, and that AI lacks intentionality and emotional responses. Furthermore, whilst AI has been shown to enhance individual creative processes, it has been suggested that it reduces overall creativity in a population (Doshi & Hauser, 2024). This may be a cause for particular concern for those working in creative industries, particularly those with a particularly keen concern around market competition.

Responsibility is a central concern for creative works, as how we attribute responsibility (positive or negative) is key for questions of creative attribution and authorship (e.g. see Mag Uidhir, 2013). The willingness to assign responsibility for work has been found to be asymmetrical between humans and AI. People have been found to be less willing to credit AI systems, when they would credit a human for doing the same level of work (Bankins et al. 2022), and indeed, this includes in creative work as Formosa et al (2024) found a similar effect with the use of ChatGPT for creative writing assistance.

Further to this, there is the well-known problem of the responsibility gap (Matthias, 2004; Coeckelbergh, 2020). This gap occurs when we cannot attribute moral responsibility to an AI, but we also cannot easily attribute responsibility to a human actor. Moral responsibility is typically understood as requiring knowledge and control (Coeckelbergh, 2020). As most AI systems seem to be able to meet these requirements as they are thought to lack key features such as agency, autonomy, or consciousness (which are thought necessary for control and knowledge respectively, see Coeckelbergh 2020, p. 111), even when they have some level of autonomy. As such, when using AI, we may be concerned that any moral issues cannot be properly attributed to anyone. This may be an issue that comes up for users of AI systems. We may see users unwilling to credit AI systems (as suggested above), yet they may also struggle to attribute responsibility to humans that make use of these systems.

In the case of creative works, we also have what we might call creative or artistic responsibility at play, which concerns how we want to credit creators for their work. We are already seeing a similar ‘responsibility gap’ at play with the use of AI in these domains, as seen in discussions of the copyrightability of AI works. This has been evidenced in current legal approaches to AI. In the UK and the USA there remains uncertainty over how to attribute legal authorship for AI generated works. In the US in 2023 the Copyright Office clarified that "Most fundamentally, the term "author," which is used in both the Constitution and the Copyright Act, excludes non-humans." and that " When an AI technology determines the expressive elements of its output, the generated material is not the product of human authorship." (Copyright Office, 2023). As Erickson writes, "In other words, the human-made aspects of AI-generated works, such as "prompt instructions"', are eligible for copyright protection, while any output from, such as images in a text-to-image model like Midjourney, are not." (Erickson, 2024). Under UK law, it is not yet clear who could be awarded copyright of AI outputs. According to Erickson, the UK Copyright Designs and Patents Act (CDPA 1988) allows copyright protection in works which have been generated by a computer in some circumstances, when there is no human author to attribute the work to. A person will be vested with a lower standard of copyright in such a case, and this will be the person who made arrangements for the work's creation (Erickson, 2023). However, as Erickson points out, it is not clear whether this person must meet a threshold of originality as is typical in copyrighted works, and whether prompting a system like ChatGPT or text-to-image generators would count in such a case.

This relates to concerns from authors, researchers, artists, as well as users of generative AI tools that the works generated are reproductions of works from the training data, or that they are plagiarised in some way.[4] Of course, this relates to the copyrightability of AI outputs, leading some companies (such as Adobe)[5] to assure users of their generative tools that they will not be liable for their works (Erickson, 2024). There is however also a moral concern here, as plagiarism is typically judged to be morally wrong (in our shared social context), due to its perceived connection to dishonesty (East, 2010), whether there is a legal question of profit or not.

Finally, there are increasing ethical concerns regarding the use of AI in general, including environmental concerns, privacy concerns and concerns of bias in AI outputs (Stahl, 2021). Whilst these are not particular to generative AI systems or their creative uses, they nevertheless may be a key consideration for those considering the use of AI in their industry.

[4] See, for example the lawsuit against Anthropic: https://www.theguardian.com/technology/article/2024/aug/20/anthropic-ai-lawsuit-author

[5] See report of this from: https://www.fastcompany.com/90906560/adobe-feels-so-confident-its-firefly-generative-ai-wont-breach-copyright-itll-cover-your-legal-bills

3.0 Research Design and Methodology

To understand the role morality plays in AI adoption and the role of a creative professional's perceived ethical stance in the adoption of AI, a small focus group method was selected to provide in-depth, participatory insights. This drew out the views of art and design professionals to discuss their direct experience and ethical views about generative AI. The focus group contained a range of creative industry professions, including architecture, fashion, music and graphics.

To analyse the focus group a Framework Analysis (Richie and Spencer, 2002) approach was used. Initial reading of the transcripts and notes taken enabled familiarisation with the themes emerging from the focus group discussion. Using generative AI, themes were drawn out from the transcript and notes (Joel-Edgar and Pan, 2023). The generative AI themes were cross-referenced with a human who was familiar with the transcript and notes, to produce these final themes:

1. Perceptions of Generative AI

- Awareness and Familiarity
- Cautious Optimism
- Perceived Utility
- Synthetic Risks

2. Applications and Current Uses

- Productivity Enhancement
- Creative Industry Application
- Collaborative Utility

3. Moral and Ethical Considerations

- Data Ownership and Copyright
- Bias, Representation and Inclusivity
- Blockchain and Provenance

4. Levels of Morality: Individualised, Industrial, Societal

- Individualised view
- View of others
- Environmental Impact
- Democratisation
- Homogenisation

- Job Displacement and Economic Effects
- Educational Concerns

5. Future Prospects and Regulatory Needs

- Innovative Possibilities
- Psychological and Social Risks
- Regulatory Imperatives

For each theme and sub-theme, (e.g. educational concerns), a matrix was drawn up in which the comments from each focus group participant that related to the theme were added and summarised. A summary per theme was then documented to produce the findings from the Focus group.

4.0 Initial Findings

4.1 Perceptions of Generative AI

Many participants initially viewed generative AI through specific branded tools like ChatGPT, Midjourney, and Claude. There was a mixture of excitement and wariness. For example, one participant commented that: "*I kind of see it in that context as like a new pigment. It's the new palette. It's a new medium for artists, and it's taking it out of the way that I interact with it on a very basic level, and using creativity to kind of propel it forward, which is really exciting*". Whereas others commented that: "*I feel like it's not the best option because...it's an algorithm, giving you products based on what has been fed*".

4.2 Applications and Current Uses

Participants frequently cited generative AI's broad reach, including applications in text, image, and video generation, with some associating it closely with tools they already use, like Photoshop. One participant noted that: "*So [we] use ChatGPT 10-12 times a day for a variety of different things. Can be turning random thoughts in between meetings into agendas, improving the tone of writing for a particular document, or pulling research together. It's getting 80% of what we're working on done...*". Even though it was acknowledged that some creatives avoided using it, one participant commented on its omnipresence and that people are using it without knowing they do so: "*I use AI for Photoshop as well, like the generative tool…I think people unknowingly have been using it without even thinking, 'Oh, this is AI.'*" It was acknowledged that generative AI was widely used across diverse fields for purposes ranging from enhancing productivity to assisting creative processes. ChatGPT was noted as helping to convert ideas into structured agendas, refine tone for different audiences, and develop customer journey maps. Examples were given of how AI can be used in Photoshop for tasks like background removal, while more advanced uses involve ideation in creative industries, such as generating visuals for music platforms or transforming energy data into visual displays.

4.3 Moral and Ethical Considerations

A number of ethical considerations were discussed. Bias in AI algorithms raised significant moral concerns, as AI outputs sometimes reflect racial, gender, and socioeconomic biases. Participants recounted incidents where AI tools produced stereotypical images. For example, " *Sometimes that information can give you biassed results... If you search, like, to generate an image of a teacher in a classroom, it would show you specifically like a female, or if you search like a pilot, it will specifically give you like a man... it doesn't have feelings, it doesn't consider morals.*" Participants emphasised the importance of data diversity to mitigate these biases.

Copyright issues were discussed, particularly around AI models trained on publicly available but potentially copyrighted materials. To mitigate copyright issues, one participant gave an example of how the music industry deals with sampling in the creative process: "*I feel like the music industry with sampling has quite a good model for how we should be using it, and in terms of copyright...Blockchain plays a really interesting role in the future of AI, in giving it provenance, and having a record of where it comes from. So...if they're using it in a responsible, ethical way, and saying where the source is from and showing provenance and actually putting blockchain into that, then we can trust them.*"

4.4 Levels of Morality: Individualised, Industrial, Societal

There were strong reactions both in favour and against someone creative using generative AI. One participant noted:" *I specifically say I absolutely avoid it because of ethical concerns. Generative AI is trained on copyrighted materials... Someone has made their work, and then their style is taken*". Strong negative views of others "copying" ideas were presented. One person gave the following example from their career: "*they asked me to get other brand designs similar to their brand image and copy* *those, just tweak certain elements so that it looks exactly the same. I already felt bad because I thought small brands, small businesses, who are putting their whole passion into their designs. I wouldn't be surprised if they were using AI now to just take an image from another brand and be like, okay, just tweak this a little bit*". Others were more positive: "*I think if they're [someone creative] using it in a progressive and innovative way, fantastic.*". Another example supported this view: "*The value can go to the idea now, because with AI, anyone can execute in a few years' time... I think the brands and agencies who are doing interesting activities like [identifiable name redacted] in Japan are doing some amazing things with AI,*" Referring to AI as having a democratising effect.

Collaborative uses included developing experience briefs and working on research projects that visualise sustainable infrastructure. The group reflected on how AI shapes collaborative work, highlighting the benefits of using AI in team settings to clarify ideas and generate shared visuals quickly. For example, one participant noted that: "*It's incredibly collaborative for visualised sound. Musicians on that platform can upload their track, choose a preset, and see how their song interacts with art and AI*". Another example given of AI use on a collaborative creative process was this: "*In the studio, particularly as you work through creating experience briefs, you want to match the tone of the idea to the audience and the context. You can add in some images for training... It's a very good briefing tool—picture worth a thousand words—to illustrate your thinking.*". Some ethical implications arose when discussing data sourcing and ownership with some users emphasising the need to collaborate responsibly, ensuring that creators retain rights to their contributions.

4.5 Future Prospects and Regulatory Needs

Generative AI's impact on society sparked varied responses. Some participants believed AI could democratise creativity by enabling more people to execute complex projects, such as films or artwork, but expressed worries about homogenisation in creative output. AI-driven applications in industries like fashion and arts raised concerns about job displacement and an over-reliance on AI for content creation, potentially diminishing human creativity. In education, participants feared AI could encourage shortcuts, resulting in a loss of deep learning and critical thinking skills. Concerns over environmental impact were strong, with AI's energy consumption viewed as a "deal-breaker" for some, noting AI's significant water and energy use in data processing. For example: "*The thing that I don't like about AI is that it creates a distance between us and the resources it consumes. The actual server processing all this stuff is, like, in some mountain...we have no clue how much energy consumption is going in...even a single response on ChatGPT can consume a whole glass of water. So, really, is it convenient, actually, that it's worth it*".

5.0 Discussion

Generative AI's versatility, from productivity to creativity, aligns with its classification as both a creative and rational tool. Focus group participants cited practical uses such as transforming ideas into agendas and generating visuals for creative briefs, mirroring literature that emphasises AI's applications in diverse contexts (Dwivedi et al., 2023; Zhuo et al., 2023). For example, one participant remarked that AI accomplished "*80% of what we're working on*," indicating its potential to increase productivity.

However, concerns about the potential over-reliance on AI and its impact on originality echo the literature's focus on balancing automation with human creativity (de Cock Buning, 2018). Participants worried about AI homogenising creative output and replacing deep learning, consistent with warnings about excessive automation undermining values critical to creative industries (Inie, Falk, & Tanimoto, 2023).

Ethical challenges, including biases in AI outputs and copyright issues, featured in the focus group discussions. Participants highlighted cases of AI reinforcing stereotypes, such as associating gender roles with specific professions, which align with documented issues of bias in training data (van Dis et al., 2023). To address this, participants emphasised data diversity.

Concerns about copyright were also mentioned, with participants expressing discomfort about AI models using copyrighted materials without appropriate attribution. These discussions reflect the literature's emphasis on the importance of intellectual property rights in creative industries. For example, participants suggested blockchain as a mechanism to trace and verify the provenance of AI-generated content, aligning with proposals for using technology to ensure ethical use and attribution.

Participants expressed varied moral judgments about generative AI, reflecting its dual role as both an enabler and disruptor. Some saw AI as a democratising tool, allowing broader access to complex creative processes, consistent with the literature's perspective on AI's potential for inclusivity (van Dis et al., 2023). However, others voiced strong concerns about originality and ethics, particularly regarding the misuse of AI to replicate or appropriate creative work, mirroring debates in the literature about moral correctness in technology use (Inie, Falk, & Tanimoto, 2023).

Participants also raised concerns about AI's broader societal impact, such as job displacement and diminished critical thinking skills in education. These concerns resonate with literature discussing the challenges of integrating automation in industries reliant on originality and intellectual engagement (de Cock Buning, 2018). The environmental impact of AI, highlighted in the focus groups, underscores the need for sustainability in AI development, an area less explored in the reviewed literature but crucial for ethical integration.

Participants consistently linked generative AI adoption to the agency of creative professionals, reinforcing the literature's argument that professionals are key to integrating AI into workflows effectively. While some viewed AI as a threat to originality, others praised its potential to complement creative processes, such as generating visuals to clarify ideas. This reflects the literature's emphasis on balancing technical assistance with preserving the core values of creativity and intellectual property.

The focus group findings, supported by the literature, underscore the complex dynamics of generative AI adoption in creative industries. While participants recognised AI's potential to enhance workflows and democratise creativity, they also highlighted ethical, intellectual property, and societal concerns. Addressing these challenges requires inclusive data practices, frameworks for ethical usage, and sustainable development models. Future work should focus on equipping creative professionals to navigate these complexities, ensuring generative AI serves as a tool for empowerment rather than an eroder of originality.

6.0 Conclusion

The exploration of generative AI in the creative industries reveals a multifaceted dynamic, where its potential is both celebrated and critiqued. The insights from focus group discussions and literature highlight the dual role of AI as a powerful enabler of innovation and productivity while simultaneously posing ethical, moral, and societal challenges that cannot be ignored. Generative AI has the capacity to streamline workflows, democratise access to creative tools, and foster innovative collaborations that were previously unimaginable. The focus group participants emphasised its practical applications, from generating visuals for creative briefs to automating repetitive tasks, which aligns with broader research acknowledging AI's capacity to contribute to creative processes.

However, these benefits are accompanied by concerns over originality, intellectual property, and the homogenisation of creative outputs. The notion of originality, which is a cornerstone of the creative industries, is challenged by AI's ability to replicate and reinterpret existing work, often trained on datasets that may include copyrighted materials without explicit consent. This raises significant questions about authorship and accountability, echoing the sentiments of participants who voiced unease about AI eroding the integrity of creative contributions. The ethical implications of generative AI extend beyond copyright issues, with biases in AI algorithms emerging as a significant concern. Addressing such biases will require deliberate efforts to diversify datasets and implement transparent practices in AI development. The societal implications of generative AI also extend to concerns about job displacement, the potential decline of critical thinking skills, and its impact on education. All the concerns suggested the need for a thoughtful approach to AI integration.

The initial findings reflected the complexity of generative AI adoption in the creative industries, where its benefits must be carefully weighed against its potential risks. The successful integration of AI will require robust frameworks that address ethical, moral, and societal considerations while empowering creative professionals to harness its potential effectively and appropriately.

This research is not without limitations. The study relies on data collected from a small focus group of creative industry professionals. While this method provides rich qualitative insights, the findings may not be representative of the broader population of creative professionals across diverse industries, cultures, or geographical locations. The perspectives captured are limited to those of the participants and may not encompass the full spectrum of attitudes, experiences, or ethical concerns. Although the focus group included a range of creative professions (e.g., architecture, fashion, music, and graphics), it may not fully represent all sectors of the creative industries, such as film, gaming, or fine arts. The findings may, therefore, omit perspectives from other creative areas.

7.0 Future Work

Therefore, future research should focus on bridging the gaps identified in the literature as well this research, particularly around the idea of credibility in the creative context. The role that explainability has in influencing credibility and the impact credibility has on the moral judgement of creative industry professionals. This has resulted in the proposed theoretical framework (figure 1). This is based on the Source Credibility Theory (Hovland, Janis and Kelly, 1953) in which explainability influences the credibility model, the attitude (moral judgement) and the outcome (disclosure of use). We extend the credibility model by using the elements of Source Credibility Theory: trustworthiness, expertise and attractiveness but interpreting attractiveness as creativity. We interpret creativity as the perceived novelty of the AI output. We intend to use this research model to survey creative industry professionals and quantitatively ascertain whether this research model is supported.

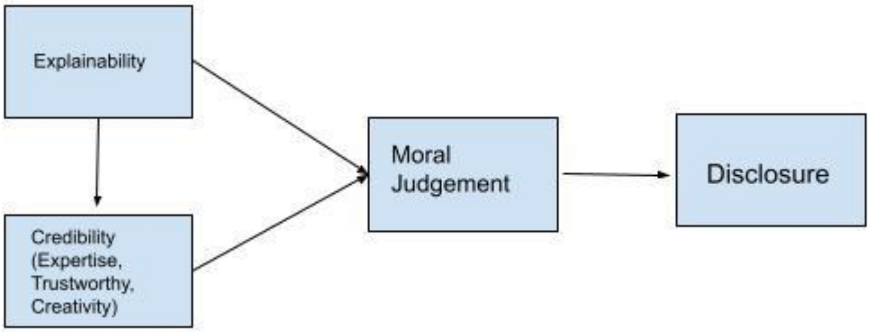


Figure 1. Proposed Research Model

References

Amabile, T. M. (2018) *Creativity in context: Update to the social psychology of creativity*. Routledge.

Anantrasirichai, N. and Bull. D. (2022) “Artificial intelligence in the creative industries: a review.” *Artificial intelligence review* 55, no. 1 : 589-656.

Brynjolfsson, E. and McAfee, A. (2014) *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & company

Burr, V. (1995) *An Introduction to Social Constructionism*. London: Routledge.

Chowdhury, S., Budhwar, P., Dey, P. K., Joel-Edgar, S. and Abadie, A. (2022) "AI-employee collaboration and business performance: Integrating knowledge-based view, socio-technical systems and organisational socialisation framework." *Journal of Business Research* 144: 31-49.

Coeckelbergh, M. (2020). 'A-responsible machines and unexplainable decisions' AI Ethics, Cambridge MA: MIT Press, pp. 109-123.

Copyright Office (2023) *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, A Rule by the Copyright Office,* Library of Congress on 03/16/2023. Available at: https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence

Csikszentmihalyi, M. (1990) *Flow: The Psychology of Optimal Experience*. New York: Harper & Row.

Danesi, M. (2024) AI in Marketing and Advertising. In: AI-Generated Popular Culture. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-54752-2_7

De Cock Buning, M. (2018) Artificial Intelligence and the creative industry: new challenges for the EU paradigm for art and technology by autonomous creation. In Research Handbook on the Law of Artificial Intelligence. Edward Elgar Publishing.

Dowling, M. and Lucey, B. (2023) ChatGPT for (finance) research: The Bananarama conjecture.

Doshi, A. R., and Hauser, O. P. (2024) "Generative AI enhances individual creativity but reduces the collective diversity of novel content" *Science Advances* 10(28) DOI:10.1126/sciadv.adn5290

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Baabdullah, A.M., Koohang, A., Raghavan, V., Ahuja, M. and Albanna, H., (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, p.102642.

East, J. (2010) Judging plagiarism: a problem of morality and convention." *Higher Education* 59, 69–83. https://doi.org/10.1007/s10734-009-9234-9

Erickson, K. (2024) Copyright protection in AI-generated works. [Blog], Creative Industries Policy and Evidence Centre. Available at: https://pec.ac.uk/blog_entries/copyright-protection-in-ai-generated-works/

Gilligan, C. (1982) *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press.

Hemingway, C. A., and Maclagan, P. W. (2004) Managers' personal values as drivers of corporate social responsibility. *Journal of business ethics* 50: 33-44.

Henkin, C. (2023) "Orchestrating the future: AI in the Music Industry". *Forbes,* Dec 5, 2023. Available at: https://www.forbes.com/sites/davidhenkin/2023/12/05/orchestrating-the-future-ai-in-the-music-industry/

Hovland, C. I., Janis, I. L. and Kelley, H. H. (1953) *Communication and Persuasion: Psychological Studies of Opinion Change*. New Haven, CT: Yale University Press

Inie, N., Falk, J. and Tanimoto, S., (2023) Designing Participatory AI: Creative Professionals' Worries and Expectations about Generative AI. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-8).

Ivcevic, Zorana, and Mike Grandinetti. (2024) “Artificial intelligence as a tool for creativity." *Journal of Creativity* 34, no. 2: 100079.

Jennings, P.L., Mitchell, M.S. and Hannah, S.T., (2015). The moral self: A review and integration of the literature. *Journal of Organizational Behavior,* 36(S1), pp.S104-S168.

Joel-Edgar, S. and Pan, Y. C. (2024). Generative AI as a Tool for Thematic Analysis: An Exploratory Study with ChatGPT. *UK Academy for Information Systems Conference Proceedings 2024*. 8.

Kangasharju, Arja, Liisa Ilomäki, Minna Lakkala, and Auli Toom. (2022) "Lower secondary students’ poetry writing with the AI-based poetry machine." *Computers and Education: Artificial Intelligence* 3: 100048.

Kohlberg, L. and Hersh, R. H. (1977) "Moral development: A review of the theory." *Theory into practice* 16, no. 2 : 53-59.

Kohlberg, L. and Power. C. (1981) "Moral development, religious thinking, and the question of a seventh stage." *Zygon: Journal of Religion and Science* 16, no. 3 (1981).

Mag Uidhir, C. (2013). *Art and Art-Attempts.* Oxford University Press.

Malle, B. F. (2021). Moral judgments. Annual Review of Psychology, 72, 293–318.

Matthias, A. (2004) ‘The responsibility gap: Ascribing responsibility for the actions of learning automata’, *Ethics and Information Technology,* 6(3), pp. 175-183.

McCormack, Jon, Oliver Bown, Alan Dorin, Jonathan McCabe, Gordon Monro, and Mitchell Whitelaw. (2014) "Ten questions concerning generative computer art." *Leonardo* 47, no. 2: 135-141.

McCroskey, James C., and Jason J. Teven. (1998) "Goodwill: A Reexamination of the Construct and Its Measurement." *Communication Monographs* 66, no. 1: 90–103. https://doi.org/10.1080/03637759909376464.

Mogavi, R.H., Deng, C., Kim, J.J., Zhou, P., Kwon, Y.D., Metwally, A.H.S., Tlili, A., Bassanelli, S., Bucchiarone, A., Gujar, S. and Nacke, L.E., (2023). Exploring User Perspectives on ChatGPT: Applications, Perceptions, and Implications for AI-Integrated Education. *arXiv preprint* arXiv:2305.13114.

Pavlik, J. V. (2023). Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism & Mass Communication Educator*, 10776958221149577.

Piaget, J. (1932) *The Moral Judgment of the Child*. Translated by Marjorie Gabain. London: Kegan Paul, Trench, Trubner & Co.

Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing*, 85(1), 131-151.

Reynolds, S.J. and Ceranic, T.L., (2007). The effects of moral judgment and moral identity on moral behavior: an empirical examination of the moral individual. *Journal of applied psychology,* 92(6), p.1610.

Rhodes, M. (1961). An analysis of creativity. *The Phi Delta Kappan*, 42(7), 305-310.

Ritchie, J. and Spencer, L., (2002). Qualitative data analysis for applied policy research. In *Analyzing qualitative data* (pp. 173-194). Routledge.

Runco, M, A., and Garrett J. J.,. (2012)"The standard definition of creativity." *Creativity research journal* 24, no. 1 : 92-96.

Sawyer, R. K. (2007) *Group Genius: The Creative Power of Collaboration*. New York: Basic Books..

Stahl, B.C. (2021) “Ethical issues of AI”. *Artificial Intelligence for a better future: An ecosystem perspective on the ethics of AI and emerging digital technologies,* Springer Nature. pp.35-53.

Sullivan, Y. W., & Wamba, S. F. (2022). Moral Judgements in the Age of Artificial Intelligence. *Journal of Business Ethics*.

Qadir, J., (2023). May. Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE Global Engineering Education Conference* (EDUCON) (pp. 1-9). IEEE.

Van Dis, E. A., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature*, 614(7947), 224-226.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.

Vygotsky, Lev S. (1978) *Mind in Society: The Development of Higher Psychological Processes*. Edited by Michael Cole, Vera John-Steiner, Sylvia Scribner, and Ellen Souberman. Cambridge, MA: Harvard University Press.

Weatherbead, J. (2024) Adobe Launches AI Video Generation Tools in Premiere Pro with Firefly Model. *The Verge*. Published October 14, 2024. https://www.theverge.com/2024/10/14/24268695/adobe-ai-video-generation-firefly-model-premiere-pro.

Zhang, N., Yan, J. and Briot, J.P., (2023). Artificial intelligence techniques for pop music creation: A real music production perspective. *SSRN.*

Zhuo, T.Y., Huang, Y., Chen, C. and Xing, Z., (2023). Exploring ai ethics of chatgpt: A diagnostic analysis. arXiv preprint arXiv:2301.12867.

Zhu, Y., Baca, J., Rekabdar, B. and Rawassizadeh, R., (2023). A Survey of AI Music Generation Tools and Models. *arXiv preprint* arXiv:2308.12982.

Key Resources and Barriers for Generative AI Implementations: Insights from Service-Driven Organizations

Tom Essers
University of Twente, Enschede, The Netherlands

Baran Gülbey
University of Twente, Enschede, The Netherlands

Jelmer Hofman
University of Twente, Enschede, The Netherlands

Valerie Seinstra
University of Twente, Enschede, The Netherlands

Stephan Wildenberg
University of Twente, Enschede, The Netherlands

Pauline Weritz*
University of Twente, Enschede, The Netherlands, p.weritz@utwente.nl

Abstract

Like the service industry, academia provides intangible outputs (education, knowledge dissemination, research) that directly benefit students, researchers, industries, and society. With the increasing importance of digital technologies, the sector experiences challenges. Generative AI holds promise for creating content, automating tasks, and enhancing organizational efficiency, but implementing it effectively has its challenges. This study investigates the key resources and barriers to the implementation of Generative AI in service-driven organizations. Through a systematic review of recent research, key resources are identified, such as organizational understanding, financial strength, strong technology skills, and positive expectation of results as enablers of successful GenAI implementation. At the same time, barriers like data privacy issues, lack of knowledge, regulatory concerns, and data quality appeared. Our framework shows the most important resources for each phase of GenAI implementation, from early adoption planning to the actual use of GenAI. This study offers practical guidance to help organizations prepare for GenAI, manage critical resources, and overcome common challenges for a smoother, more effective implementation.

Keywords: Key resources, Barriers, GenAI, Generative AI, Service industry, SLR.

1.0 Introduction

Artificial Intelligence (AI) is a system's ability to correctly interpret external data, learn from such data, and use those learnings to achieve specific goals and tasks through flexible adaptation (Dennehy et al., 2023). Since the quick rise in popularity of AI and releasing ChatGPT on the 30th of November 2022 by the organization OpenAI, generative AI (GenAI) has made its introduction. Feuerriegel (2023) defines the term generative AI as the computational techniques that can generate seemingly new, meaningful content such as human-like written text, producing music, images, and audio, or solving complex problems. Specific tools, such as Dall-E 2 (image generation), GPT-4 (text generation), and Copilot (code generation), are already estimated to automate the jobs of 300 million knowledge workers.

However, understanding how to implement the technology is still a question for organizations (Lawrence et al., 2023). Specifically, governmental organizations or service-driven organizations have more challenges with creating value by using AI (Wang et al., 2021). Organizations are investigating how to utilize this technology to increase the efficiency of processes, improve customer service, and innovate quickly (Holmström & Carroll, 2024). Potential innovative applications of GenAI can be for conversational interfaces, such as chatbots and virtual assistants, or content creation (Holmström & Carroll, 2024). With the fact that GenAI is still relatively new, and service-driven organizations are having trouble implementing this technology, research regarding the implications of these implementations exists, however, some challenges remain (Li et al., 2024). Exploring how GenAI impacts academia is intriguing because it challenges traditional norms of knowledge creation, dissemination, and engagement. Studying these impacts advances our understanding of GenAI and ensures that its adoption aligns with the mission of academia enriching society through knowledge.

This study aims to address the gap by performing a systematic literature review to gather existing knowledge on the topic by shedding light on key resources that impact the implementation of GenAI in service-driven organizations. Therefore, the research question is: *How do key resources and barriers impact the implementation of GenAI in service-driven organizations?*

We conducted a systematic literature review (SLR) to investigate how key resources and barriers impact the implementation of GenAI. With the SLR, we provide an evidence-based paper with results related to the implementation, knowledge of how to implement, and advice for organizations where focus is needed. These contributions pave the way for a deeper exploration of AI implementation dynamics and its theoretical underpinnings. With the findings, we address the conceptual (un)clarity in GenAI research, discuss unexpected findings in resource allocation, and identify overlapping factors between the findings and the TAM 3.0 model (Venkatesh & Bala, 2008). Further, we explore the practical implications and offer guidance for GenAI implementations in the service industry such as academia. Second, we support the scale-up of GenAI projects by identifying the key resources and offering insights on how to mitigate post-deployment risks. Next to these implications, we provide avenues for future research.

2.0 Conceptual Background

2.1 GenAI as the Study Context

AI technologies lead to the customization of services to customers and employees. Other opportunities for the use of AI are accelerating execution or operations, reducing costs, increasing engagement, increasing innovation, and fortifying trust (Davenport & Mittal, 2023). Nevertheless, understanding the technology is core to using it for the proposed objectives of the organization. Nowadays, research divides AI into two categories, namely traditional generations of predictions and GenAI to create content (Wessel et al., 2023). Predictive AI is a subfield whose purpose is to make predictions about future events using historical data and machine learning models. Predicting under which circumstances a situation could occur makes it possible for the organization to define its strategy or even adjust where necessary. GenAI, on the other hand, is the subfield of AI that is capable of generating new content (Feuerriegel et al., 2023). This content can be in the form of text, videos, sound, images, or other data. GenAI is capable of redesigning organizational processes to increase efficiency and potentially improve customer experiences (Wessel et al., 2023). In the context of this study, we define GenAI as *'a subfield of AI that is capable of generating new content in the form of text, videos, sound, images, or other data.'*

While GenAI can be utilized in all sorts of organizations, service-driven organizations are especially interesting, due to the nature of services (Weritz et al., 2024). Delivering a service to another organization brings customer experience complexity that is less present within good-driven exchanges (Vargo & Lusch, 2016). The organization delivering the service is more involved in the different phases in which a solution is used by another organization. The potential of GenAI is prominent in improving customer experience, for example, in the form of a chatbot (Wessel et al., 2023).

2.2 Phases of GenAI Implementation

To integrate GenAI in organizations, a clear pattern and three common phases were identified in the SLR, namely adoption, development & deployment, and use. Many papers focus on the adoption phase, this is expected, given that GenAI is still relatively new for most organizations (Nair et al., 2024). Organizations are realizing the impact of using GenAI and are investigating how to adopt this new technology. To better structure the findings, the framework by Reim et al. (2020) was utilized, which offers a roadmap for AI implementation. While multiple sources were reviewed, Reim's roadmap aligns closely with the phases of GenAI integration identified. This framework helps to understand how key resources impact each phase.

The adoption phase corresponds to Reim's first two steps: "Understanding AI and Organizational Capabilities" and "Understanding the Current Business Model and Potential for Innovation" (Reim et al., 2020). During this phase, organizations analyze their resources and ability to use the technology to determine how prepared they are for GenAI. The focus on adoption in the literature analysis aligns with Reim's model, highlighting the significance of determining risks and identifying critical resources early (Reim et al., 2020). Organizations are currently finding out the obstacles they will face and how GenAI may be integrated into their operations.

The development & deployment phase in this study aligns with Reim's third step, "Developing and Refining Capabilities" (Reim et al., 2020). Although there were fewer studies discussing development than adoption, the research showed that development requires a solid infrastructure. Reim's theorem highlights how crucial it is to build these competencies, which aligns with the real-world difficulties that organizations have when putting them into practice (Reim et al., 2020). In line with Reim's last phase, the use phase concentrates on making sure staff members can use GenAI efficiently in their day-to-day duties (Reim et al., 2020). This step is a critical step within the process, as low intention of use can lead to staff not using the implemented GenAI solution.

2.3 Resource-Based View

The resource-based view (RBV) is a strategic framework that emphasizes the importance of internal resources in achieving and sustaining competitive advantage. The theory suggests that organizations can outperform competitors if they effectively utilize resources that are valuable, rare, inimitable, and non-substitutable (Barney, 1991). These might include tangible assets (e.g., physical infrastructure, technology) and intangible assets (e.g., knowledge, reputation, culture) as well as organizational capabilities, such as the ability of an organization to deploy these resources effectively. The absence of these key resources is also considered to impact the implementation of GenAI. Physical resources are defined as tangible resources, for example, equipment, raw materials, buildings, and manuals. Human resources are defined as the key people and skills. Intellectual resources include intellectual property, codified systems and processes, and the intangible know-how of the teams. Lastly, financial resources include cash and lines of credit.

3.0 Methodology

3.1 Literature Identification and Data Extraction

The methodology describes the execution process of the SLR. The first step of the SLR was to find the keywords for the search that together make up the search query. Through an iterative approach, these keywords were updated to find more potentially relevant articles. Scopus was used as the database due to its comprehensive coverage of literature and advanced search facilities. Only articles from after 2018 are included, as the introduction of the first GPT in 2018 marked the first signs of applications of GenAI in practice. This aligns with the available research into implementations of GenAI, as the amount of research available before 2018 is limited. This query retrieved results based on the following search string: *TITLE-ABS-KEY (("generative AI" OR "GenAI" OR "artificial intelligence" OR "AI") AND ("implementation" OR "adoption" OR "integration") AND ("key factors" OR "success factors" OR "critical factors" OR "best practices" OR "drivers" OR "facilitators" OR "enablers" OR "barriers" OR "key resources" OR "resources" OR "capabilit*") AND ("business" OR "corporation" OR "company" OR "enterprise" OR "firm" OR "corporate") AND ("service*")))*

3.2 Critical Analysis

A three-round approach was used to select and screen our articles. In these rounds, articles were screened based on the inclusion and exclusion criteria. Articles that did not suffice these criteria were excluded from the review. The inclusion and exclusion criteria were defined as follows: 1) Article needs to mention key resources or factors, 2) Article needs to take place in the service industry, 3) Article needs to cover the implementation of GenAI, 4) Articles need to be peer reviewed 5) Articles need to be written after 2018, 6) No books or theses, 7) Scopus impact factor (FWCI) > 1. In the first round, the title and abstract were screened. This resulted in the exclusion of 261 articles. Next, the 118 remaining articles were screened using the introduction and conclusion sections. This resulted in the exclusion of 88 articles. In the last round, the last 30 articles were completely read and screened. This resulted in the inclusion of 16 articles in this research. The full selection process is displayed in Figure 1.

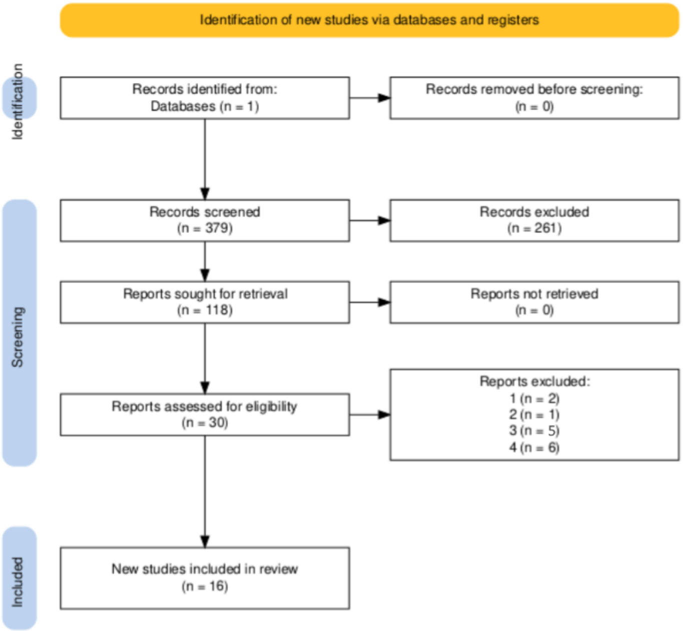


Figure 1. PRISMA structure.

The 16 included articles were coded in three phases using the grounded theory approach (Wolfswinkel et al., 2013). Grounded Theory involves systematic data collection and analysis to develop theories grounded in the data itself. Open coding was used in the first stage. In this phase, the data was broken down into discrete parts and they were labeled or coded to identify key concepts or themes (Wolfswinkel et al., 2013). All accepted articles were scanned and analyzed on key resources mentioned. To minimize bias in the selection process, all articles were reviewed by three different colleagues, with the first round being validated by a second colleague in the following round.

Axial coding was used to reassemble the data to highlight the relationships between codes and identifies categories. It links categories and subcategories to understand their interactions (Wolfswinkel et al., 2013). There, we combined key resources and barriers in categories to find underlying relationships. In the next stage, selective coding was used to refine the identified patterns and relationships. In this last stage of coding, the key resources were connected to the three GenAI implementation phases that were identified (adoption, development & deployment, and use). Based on these three steps of coding, a theoretical framework was developed, showing how key resources and barriers impact each of the three phases of GenAI implementation. Figure 1 displays the data structure, with first-order codes such as awareness of GenAI, second-order codes (axial codes) such as intellectual or human resources and the aggregated dimensions (selective codes) as key resources, barriers for the implementation phases.

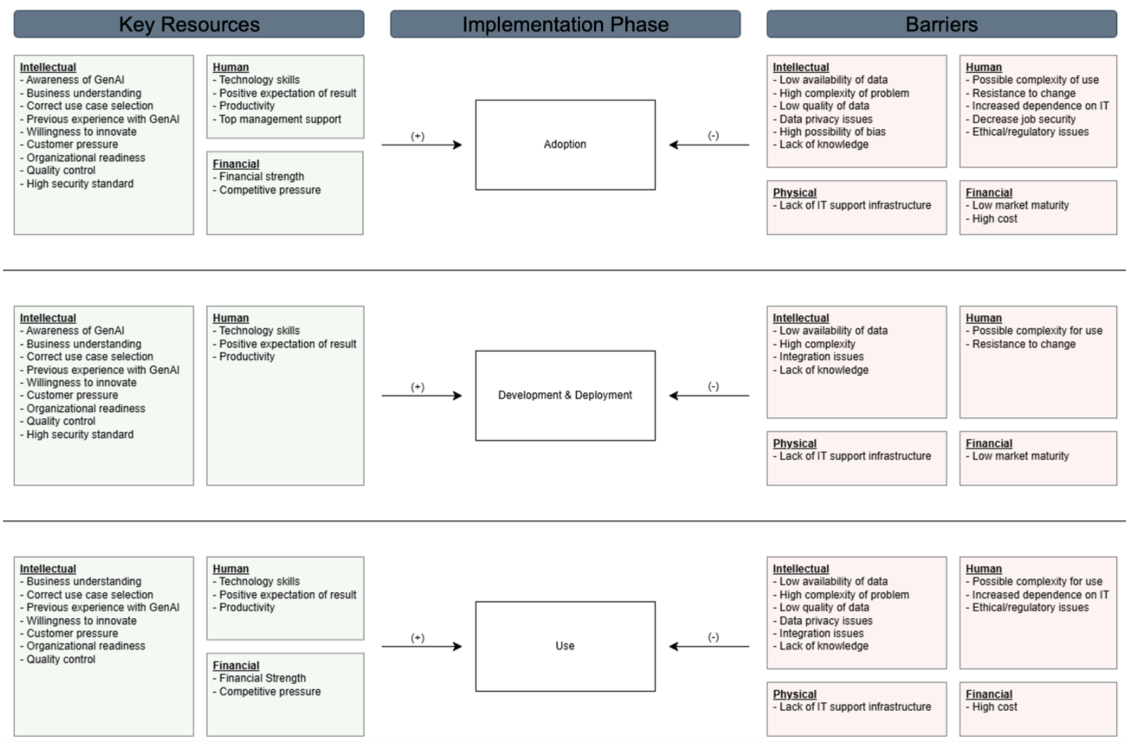


Figure 2. Data structure.

4.0 Results

4.1 Key Resources

4.1.1 Intellectual Resources

Understanding and awareness of GenAI and knowing the possible ways it can be used within the organization have a positive impact on the adoption of GenAI solutions. When employees and managers recognize the importance of AI as a possible tool for business transformation and a way to use it to improve their processes, they are more likely to talk about it to their senior managers (Mogaji & Nguyen, 2021). Awareness of the potential, and the issues of implementing a solution are relevant for this decision-making, but also for development & deployment, and use of it (Pai & Chandra, 2022). Business understanding includes the understanding of business knowledge and its placement within the sector. This includes understanding the key success indicators, and possible ways that an organization wants to improve (Skuridin & Wynn, 2024). Having this understanding has a positive impact on the adoption, development & deployment, and use of AI, as it allows organizations to better define ways that AI can be useful for them (Islam et al., 2022). Examples of business understanding are organizations within some sectors that have found that the implementation of chatbots is useful to improve certain processes (Sandu & Gide, 2019) and improve business interaction (Bhattacharyya, 2024). Proper identification of possible ways to improve has also shown increased adoption rates, especially when communicated well to managers (Mogaji & Nguyen, 2021).

Choosing the right use case is crucial for successful AI development & deployment. It is important to identify where AI can add value and assess whether tools like a conversational user interface are suitable for the task. Selecting the right use case has a positive impact on the development & deployment, and use of AI (Skuridin & Wynn, 2024; Labanava et al., 2022).

Organizations can be pressured into implementing GenAI services by the demands of their customers. When customers want to maintain a competitive position in the market, related organizations feel the need to adapt to these desires. This can accelerate technological adoption and is also the case for the adoption of GenAI (Wael, 2023). This was visible for organizations within the education sector, where students are expecting the software to be integrated with GenAI solutions (Sandu & Gide, 2019). For other sectors, providing a chatbot service has already become the norm (Bhattacharyya, 2024), and is an expected feature by many customers (Mogaji & Nguyen, 2021).

A high-security standard is one of the most important resources when dealing with (sensitive) data. As GenAI is trained on data and generates new data once in use, this should be taken into account. When security standards are high, it has a positive impact on the adoption and deployment of a GenAI solution. While this is often complex, it is reflected within organizations, as they often find cybersecurity to be one of the critical success factors of any GenAI solution (Skuridin & Wynn, 2024; Belanche et al., 2024).

Organizational readiness refers to the available knowledge, support, and assets for adopting a new technology. It includes a wide variety of different elements that together show how advanced an organization is. When an organization has a high readiness, it positively impacts the adoption, development & deployment, and use of GeAI solutions (Wael, 2023). Organizational readiness is a combination of multiple other resources that we have mentioned, however, it is important to mention it separately as the combination impacts the adoption, development & deployment, and use.

Expert knowledge and previous experience have a positive impact on the adoption, development & deployment, and use of GenAI. Making an informed decision to implement a GenAI solution is easier when GenAI has previously been implemented within the organization (Pai & Chandra, 2022). Additionally, having experts on the team makes it easier to develop a GenAI solution (Skuridin & Wynn, 2024; Damij & Bhattacharya, 2022), but also to use it once it is in place. Especially when earlier implementations were viewed as positive, they are more likely to recommend using it again.

Quality control refers to the strategy and execution to ensure high quality of a created GenAI solution. When this strategy and execution are in place, there is a positive impact on the adoption, development & deployment, and use of the product (Bhattacharya & Sinha, 2022; Damij & Bhattacharya, 2022). An example of such quality control is the monitoring of the performance of a chatbot (Skuridin & Wynn, 2024). It makes sense that the use of GenAI solutions is positively impacted by this, as it is expected that quality control improves the performance of the solution.

Organizations that show a willingness to innovate in new technologies are more likely to adopt GenAI solutions. This comes from an organizational culture mindset but also from individuals within the organization who show interest in new technologies (Cubric, 2020). From the use of AI perspective, this also makes sense, as the willingness to innovate is in line with a willingness to consume GenAI services after they are created (Bhattacharyya, 2024). This willingness to innovate here is seen as a ‘push-factor’. Just like ‘previous experience with AI', positive experiences increase the willingness to innovate even further (Bhattacharya & Sinha, 2022).

4.1.2 Financial Resources

Organizations adopt new technologies such as GenAI as a strategic response to competing businesses. Pressures from competitors in the same industry stimulate the adoption of new technologies in order to increase the market share and get ahead of other businesses. Industry 4.0 technologies, including GenAI, are currently being implemented because of this (Wael, 2023).

As for any technological innovation, finance is the core resource needed to purchase or build a system. The financial strength of an organization is crucial as it determines whether the organization will be able to use the GenAI solution (Pai & Chandra, 2022). Financial strength thus positively impacts the adoption of GenAI solutions. Additionally, when the organization is sure of possible economic benefits from a GenAI solution, this has a positive impact on the use of AI (Damij & Bhattacharya, 2022).

4.1.3 Human Resources

When users expect a positive outcome of the implementation of the GenAI, they are more motivated to use the technology. The adoption and use of chatbots is positively impacted by the expectation that the chatbot will give accurate, complete, and up-to-date responses (Antonio et al., 2022). Users also have more intention to use the GenAI if they believe that the technology can solve currently existing problems (Islam et al., 2022). Pai & Chandra (2022) indicate that the potential positive impact has a significant influence on the adoption of GenAI technologies.

The idea that GenAI can boost productivity has a positive effect on the adoption of it. If managers and employees believe that the productivity of their work increases, they are more likely to use the GenAI solution (Antonio et al., 2022). Improved GenAI as a chatbot increases the user's engagement and might be beneficial to organizational productivity (Bhattacharya & Sinha, 2022).

The expertise and skills in terms of technology impact the ability to use technologies such as GenAI. The implementation of GenAI often requires some technical knowledge about machine learning, AI, ETL (data extraction, transformation, and loading) processes and data protection (Labanava et al., 2022). However, a lack of experts within the organization can hinder the implementation, some indicate that it is important to have or find specialized people to help with the implementation (Skuridin & Wynn, 2024). Other papers also note that having technical skills fosters the adoption, and development & deployment of GenAI solutions (Damij & Bhattacharya, 2022; Sandu & Gide, 2019; Wael, 2023).

Support from the top management has a positive impact on the adoption of AI and the intention to use AI. The support from the top management also positively impacts the motivation of the other staff (Islam et al., 2022). Another reason why top management support has a positive effect on the adoption of GenAI, is that they are often the decision maker and providers of financial support, they also decide about the strategic direction of the organization. The adoption of GenAI is not possible without their intervention. Therefore, top management support has a positive impact on the adoption of GenAI (Wael, 2023).

4.2 Barriers

4.2.1 Lack of Intellectual Resources

AI systems, especially those used in public services, bring up important problems in the cases of trust, fairness, and transparency. Addressing these concerns is crucial to gaining public trust and ensuring that AI platforms are used responsibly and ethically (Damij & Bhattacharya, 2022; Labanava et al., 2022). In academic environments, for example, students have expressed discomfort about data privacy, fearing that they might lose their personal information (Bhattacharya & Sinha, 2022, Sandu & Gide, 2019). This fear is shared by many users of GenAI solutions, who are more skeptical about how their data is used and stored. Since these AI systems rely on large datasets to continually collect new data during use and are still being trained using the new data, privacy concerns can significantly lower the willingness to adopt such technologies (Antonio et al., 2022).

In some cases, AI solutions can be less effective than traditional methods. When a technology is difficult to understand, organizations are more likely to avoid adopting it. GenAI tools are generally easy to use, but they can be complex to implement and need a high level of expertise. This complexity can negatively impact adoption (Wael, 2023). Also, if design thinking isn't applied correctly, it can reduce trust in the technology and lower its effectiveness (Labanava et al., 2022).

Bias in AI systems is a critical issue that can arise from the data used to train. The data selected for AI models directly impacts the decisions and answers provided to customers. If biased or unrepresentative data is used, the AI's outputs can reflect those biases, leading to unfair or inaccurate results. This challenge is particularly concerning in systems that rely on large datasets for decision-making (Feuerriegel et al., 2023). This concern needs to be thought of in the adoption phase, otherwise it could lead to issues in the later phases.

Integrating AI into existing systems is a big challenge. Aligning new technology with old systems can cause delays and raise costs. Projects like chatbots are hard because they need to work with many different software and IT tools. This complexity can slow down the development & deployment of the system. This way the benefits of AI could be limited if the systems don't work together properly (Skuridin & Wynn, 2024).

A common barrier to AI adoption is the lack of understanding about the potential capabilities of machine learning (ML) techniques, particularly for solving specific problems. This knowledge gap can lead to unrealistic expectations of what the technology can achieve, which hinders proper adoption and use (Bhattacharya & Sinha, 2022; Cubric, 2020). In some cases, users may not fully comprehend how AI tools like large language models (LLMs) and generative AI can impact their work or industry. However, some recognize the future potential of AI and are taking steps to educate themselves, such as learning prompt engineering, to better prepare for using AI (Bhattacharyya, 2024). The difference between those who are trying to close the knowledge gap and those who are not shows how important education and knowledge are to the successful adoption and usage of AI (Sandu & Gide, 2019).

A lack of access to large, structured datasets can impact the performance of the AI. In many cases, AI struggles with unstructured data, which is common in fields like healthcare. Most healthcare data is unstructured, difficult to share, and project datasets often includes confidential information. This makes them hard to collect and maintain. Furthermore, issues like reproducibility and generalizability happen, often called the "black-box" problem, making it even more challenging to develop effective AI models (Bhattacharya & Sinha, 2022; Cubric, 2020; Christoph Szedlak et al., 2021).

Low data quality can significantly impact the success of AI models. Making sure that systems like chatbots function with minimal errors is essential to their effectiveness (Savastano et al., 2024). One major resource influencing the adoption of GenAI solutions is the accuracy and quality of the data available. High-quality data improves implementation, but poor data quality can result in ineffective models, which could lead to organizations opting not to adopt the technology (Fotis Kitsios et al., 2023). Additionally, the limitations of AI algorithms in recognizing and extracting implicit emotional cues from interactions can further challenge the effectiveness of models, especially in areas like chatbots (Damij & Bhattacharya, 2022). Low-quality data can therefore hinder the proper functioning and broader acceptance of AI technologies (Bhattacharya & Sinha, 2022; Sandu & Gide, 2019).

4.2.2 Lack of Financial Resources

High costs limit the integration of AI significantly in all phases. Most costs occur in the deployment phase where deadlines are often extended due to implementation issues (Savastano et al., 2024). These issues include the higher complexity of system integration and data cleaning than initially expected. Furthermore, data labeling can be an exceedingly expensive effort (Bhattacharya & Sinha, 2022; Cubric, 2020).

Low market maturity refers to the underdevelopment or insufficient readiness of technologies in a given industry or sector, which can hinder the implementation of GenAI (Christoph Szedlak et al., 2021). According to Christoph Szedlak et al., (2021), this lack of maturity presents a significant barrier to the adoption of AI-based technologies. This paper highlights that this immaturity, along with high costs, poses a major obstacle. Similarly, (Damij & Bhattacharya, 2022) notes that despite AI's potential, its adoption, especially in sectors like mental health, remains limited due to the early phase of research and application, which further limits acceptance. This slow market development results in hesitation in both businesses and public sectors.

4.2.3 Lack of Human Resources

The implementation of GenAI technologies can lead to concerns about job security, particularly among people who fear that AI will replace their roles (Cubric, 2020). According to Cubric (2020), many nurses felt their job security was threatened by the introduction of AI-driven systems, which could automate tasks now performed by humans. This can also be seen in other sectors (Cubric, 2020). Employees may see AI as a tool that could make their roles redundant, leading to resistance to adoption and increased workplace tension. These fears show the need for clear communication and strategies that highlight the collaborative potential of AI, rather than its role as a replacement for human workers (Cubric, 2020).

Regulatory issues pose a significant challenge to the adoption of GenAI (Damij & Bhattacharya, 2022). According to (Damij & Bhattacharya, 2022), concerns about the lack of confidentiality protections in commercial chatbots highlight the need for government regulations and ethical guidelines to prevent the misuse of sensitive information. In highly regulated sectors like healthcare, Fotis Kitsios et al. (2023) states that the stricter the regulatory environment, the lower the rate of GenAI adoption usually is, as organizations face more legal issues. These legal issues, combined with the complexity of compliance, create significant barriers for organizations looking to implement GenAI, especially in industries where data security and privacy are high priorities (Damij & Bhattacharya, 2022; Fotis Kitsios et al., 2023).

Increased dependence on non-human technologies, such as GenAI and chatbots, can lead to several negative consequences (Bhattacharyya, 2024; Sandu & Gide, 2019). According to (Damij & Bhattacharya, 2022), over-reliance on AI-driven chatbots, particularly for critical information like mental health guidance, can be risky, especially for vulnerable users. This growing reliance on AI could not only decrease trust in human advisors but also limit users' critical thinking and decision-making autonomy, raising concerns about the balance of humans and machines (Cubric, 2020).

The complexity of GenAI solutions can be seen as a significant barrier to adoption, as professionals may resist using AI if they find it too complicated (Bhattacharyya, 2024; Bhattacharya & Sinha, 2022). According to (Damij & Bhattacharya, 2022), involving users in the design and evaluation of AI solutions, such as chatbots, is critical for social acceptance, especially in public services. Without this involvement, users may feel alienated or overwhelmed by the complexity (Damij & Bhattacharya, 2022).

Resistance to change is a significant barrier to the successful implementation of AI technologies like GenAI. According to Savastano et al. (2024), users often resist adopting new technologies, making it challenging to convince them to use AI solutions. Similarly, Skuridin & Wynn (2024) state that user skepticism about the capabilities of AI tools, combined with resistance to work with them, further decreases the adoption rate. This resistance shows the need for a user-centered approach to AI implementation, addressing concerns and building trust in the benefits.

4.2.4 Lack of Physical Resources

The lack of support infrastructure presents a significant barrier to the implementation of GenAI (Cubric, 2020). According to Cubric (2020), having the right infrastructure in place is crucial for scaling AI technologies across industries. Furthermore, Bhattacharyya (2024) highlights the importance of critical resources like reliable devices, stable internet access, and a consistent power supply for the effective use of AI-driven systems. When these resources are unavailable, they create significant barriers to the adoption and use of AI (Cubric, 2020).

5.0 Discussion

5.1 Theoretical Contribution

The findings of this study led to several theoretical advancements and offer a foundation for future research in the field. The contribution thereby is threefold: First, while executing research related to new technology, such as AI, interpretations can differ. Defining the concept of AI has a broad context, and next to the definition of the concept, the structure of the papers differs. Some of the papers refer to the technology itself as an objective, while others refer to AI as a method to use it for achieving the identified objective. With this in hindsight, papers relate more to the process of implementation or to the product of AI itself.

Second, findings in this research were different than anticipated beforehand; for example, the financial key resource ‘high costs’ was not found during the development & deployment phase. This was unexpected as many organizations go over budget at this phase due to complexity. During this phase, there is a high demand for extensive data management, high-quality infrastructure, and well-managed integration with existing business processes (Cheung et al., 2024). Boston Consulting Group found that when GenAI projects are scaled from experimental to full production, this often results in higher costs than expected due to issues with data quality, security needs, and infrastructure upgrades (Cheung et al., 2024). Another surprising element is that training is not seen as a key resource for the implementation of AI. Sharma & Yetton (2007) indicate that training after the development phase leads to a higher user acceptance and thus to a successful implementation.

Third, some of the factors identified in the TAM 3.0 model overlap with the resources that were identified in this research: 'Previous experience with AI', 'Positive expectation of result' and 'Willingness to innovate' from our paper, are similar to 'Perceived usefulness', 'Intention to use', and 'Experience' from the TAM 3.0 framework. Venkatesh and Bala (2008) have developed the TAM 3.0 model, which provides a theoretical framework with resources that influence the acceptance of technology in organizations. However, the other elements identified in the TAM 3.0 framework were not found in the literature that was studied for this research.

5.2 Practical Implications

The framework created in this paper can be very useful for organizations that are planning to implement GenAI in the service industry. Managing such a project well is essential since this technology impacts several critical areas of business performance. The proposed framework offers a systematic approach that highlights specific challenges at each phase of implementation, therefore increasing the likelihood of a successful implementation. Using this framework in the adoption phase can increase the chance of successful development & deployment by ensuring that the project is both viable and aligned with organizational goals. This importance can also be found in literature (Fountaine et al., 2019). The framework's focus on development & deployment is especially relevant given that Deloitte's study from 2024 states that 68% of organizations struggle to scale GenAI from the pilot phase due to budget constraints and unexpected technical complexity (Deloitte Survey, 2024). Having the key resources from our framework known upfront can help organizations go from the pilot phase to productive use. Lastly, our framework addresses post-deployment risks, which can impact performance over time (Cheung et al., 2024). According to Boston Consulting Group, organizations with higher GenAI maturity often experience better returns because they actively monitor and adjust their AI models to address these challenges and optimize their use (Cheung et al., 2024). Our framework helps manage the key resources and decrease these risks.

5.3 Future Research Avenues

Future studies could build on the SLR by investigating more on the implementation phases and the relevant key resources. First, future research could break down the phases of GenAI implementation in more detail (adoption, development & deployment, and use). This study showed that resources like *low availability of data* appear mostly in the adoption phase, but it is possible they also impact development & deployment and use. Moreover, the relationship between the variables could be explored more to understand the links and validate findings. Second, one gap that is seen in the results is the absence of certain potentially key resources, like *training personnel*. Effective training should be a positive resource in implementing GenAI, as well-trained employees are better able to understand and use AI tools effectively. Likewise, a lack of training could easily become a barrier to the successful use of AI. This could also mean that there may be other missing key resources that also play a role in the success or failure of GenAI integration. Finally, future research could focus on researching more and a broader range of (IS) literature.

6.0 References

Antonio, R., Tyandra, N., Nusantara, L. T., Anderies, and Gunawan, A. S. (2022) *Study Literature Review: Discovering the Effect of Chatbot Implementation in E-commerce Customer Service System Towards Customer Satisfaction,* In IEEE Xplore, https://doi.org/10.1109/iSemantic55962.2022.9920434.

Barney, J. (1991) *Firm Resources and Sustained Competitive Advantage,* Journal of Management, 17(1), 99-120.

Bhattacharya, S. and Sinha, K. (2022) *The role of artificial intelligence in banking for leveraging customer experience.* Australasian Accounting, Business and Finance Journal, 16(5), 89-105.

Bhattacharyya, S. (2024) *Study of adoption of artificial intelligence technology-driven natural large language model-based chatbots by firms for customer service interaction.* Journal of Science and Technology Policy Management.

Cubric, M. (2020) *Drivers, Barriers and Social Considerations for AI Adoption in Business and Management: A Tertiary Study,* Technology in Society, 62, 101257–101257, https://doi.org/10.1016/j.techsoc.2020.101257.

Damij, N. and Bhattacharya, S. (2022) *The Role of AI Chatbots in Mental Health Related Public Services in a (Post)Pandemic World: A Review and Future Research Agenda,* In Northumbria Research Link, https://doi.org/10.1109/temsconeurope54743.2022.9801962.

Davenport, N. and Mittal, N. (2023) *All in on AI: How Smart Companies Win Big with AI,* TJ Books Ltd., Padstow, Cornwall, UK, ISBN: 978-1-64782-469-3.

Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y. K., Mäntymäki, M., and Pappas, I. O. (2023) *Artificial intelligence (AI) and information systems: perspectives to responsible AI. Information* Systems Frontiers, 25(1), 1-7.

Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P. (2023) *Generative AI,* Business & Information Systems Engineering, 66(1), 111–126, https://doi.org/10.1007/s12599-023-00834-7.

Holmström, J. and Carroll, N. (2024) *How Organizations Can Innovate with Generative AI,* Business Horizons, https://doi.org/10.1016/j.bushor.2024.02.010.

Islam, M., Mamun, A. A., Afrin, S., Quaosar, A., and Uddin, M. A. (2022) *Technology Adoption and Human Resource Management Practices: The Use of Artificial Intelligence for Recruitment in Bangladesh,* South Asian Journal of Human Resources Management, 9(2), 324–349, https://doi.org/10.1177/23220937221122329.

Kitsios, F., Kamariotou, M., Syngelakis, A. I., and Talias, M. A. (2023) *Recent Advances of Artificial Intelligence in Healthcare: A Systematic Literature Review,* Applied Sciences, 13(13), 7479–7479.

Labanava, A., Dreyling, R. M., Mortati, M., Liiv, I., and Pappel, I. (2022) *Capacity Building in Government: Towards Developing a Standard for a Functional Specialist in AI for Public Services,* In Communications in Computer and Information Science, 503–516, https://doi.org/10.1007/978-981-19-8069-5_34.

Lawrence, C., Cui, I., and Ho, D. (2023) *The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies,* https://doi.org/10.1145/3600211.3604701.

Li, M. M., Dickhaut, E., Bruhin, O., Wache, H., and Weritz, P. (2024) *More Than Just Efficiency: Impact of Generative AI on Developer Productivity,* In AMCIS Proceedings.

Mogaji, E. and Nguyen, N. P. (2021) *Managers' Understanding of Artificial Intelligence in Relation to Marketing Financial Services: Insights from a Cross-Country Study,* International Journal of Bank Marketing, 40(6), 1272–1298, https://doi.org/10.1108/ijbm-09-2021-0440.

Nair, M., Svedberg, P., Larsson, I., & Nygren, J. M. (2024). *A comprehensive overview of barriers and strategies for AI implementation in healthcare: mixed-method design.*

Pai, V. and Chandra, S. (2022) *Exploring Factors Influencing Organizational Adoption of Artificial Intelligence (AI) in Corporate Social Responsibility (CSR) Initiatives,* Pacific Asia Journal of the Association for Information Systems, 14, 82–115, https://doi.org/10.17705/1pais.14504.

Reim, W., Åström, J., and Eriksson, O. (2020*) Implementation of Artificial Intelligence (AI): A Roadmap for Business Model Innovation,* AI, 1(2), 180–191, https://doi.org/10.3390/ai1020011.

Sandu, N. and Gide, E. (2019) *Adoption of AI-Chatbots to Enhance Student Learning Experience,* Higher Education in India, https://doi.org/10.1109/ithet46829.2019.8937382.

Savastano, M., Biclesanu, I., Anagnoste, S., Laviola, F., and Cucari, N. (2024) *Enterprise Chatbots in Managers' Perception: A Strategic Framework to Implement Successful Chatbot Applications for Business Decisions,* Management Decision, https://doi.org/10.1108/md-10-2023-1967.

Skuridin, A. and Wynn, M. (2024) *Chatbot Design and Implementation: Towards an Operational Model for Chatbots,* Information, 15(4), 226, https://doi.org/10.3390/info15040226.

Szedlak, C., Leyendecker, B., Reinemann, H., Kschischo, M., and Pötters, P. (2021) *Risks and Benefits of Artificial Intelligence in Small-and-Medium Sized Enterprises,* In 4th European International Conference on Industrial

Engineering and Operations Management, https://doi.org/10.46254/eu04.20210175.

Vargo, S. L. and Lusch, R. F. (2015) *Institutions and axioms: an extension and update of service-dominant logic,* Journal of the Academy of Marketing Science, 44(1), 5–23, https://doi.org/10.1007/s11747-015-0456-3.

Venkatesh, V. and Bala, H. (2008) *Technology Acceptance Model 3 and a Research Agenda on Interventions,* Decision Sciences, 39(2), 273–315, https://doi.org/10.1111/j.1540-5915.2008.00192.x.

Wael, A. L. (2023) *Drivers of Generative Artificial Intelligence to Fostering Exploitative and Exploratory Innovation: A TOE Framework,* Technology in Society, 75, 102403, https://doi.org/10.1016/j.techsoc.2023.102403.

Wang, C., Teo, T. S. H., and Janssen, M. (2021) *Public and private value creation using artificial intelligence: An empirical study of AI voice robot users in Chinese public sector,* International Journal of Information Management, 61, 102401, https://doi.org/10.1016/j.ijinfomgt.2021.102401.

Weritz, P., Wache, H., and Honigsberg, S. (2024) *How Digital Readiness Relates to the Intention to Use Generative AI in Workplace Service Systems,* In AMCIS Proceedings.

Wessel, M., Adam, M., Benlian, A., and Thies, F. (2023) *Generative AI and its transformative value for digital platforms,* Journal of Management Information Systems.

Wolfswinkel, J. F., Furtmueller, E., and Wilderom, C. P. M. (2013) *Using grounded theory as a method for rigorously reviewing literature,* European Journal of Information Systems, 22(1), 45–55.

Stakeholder Identification and Involvement in Policy Development and Implementation:

A Community Approach to Co-Creating Digital Economy Policy in a Developing Country

Ubongabasi Kingsley Omon (University of Salford)

Research In progress

Abstract

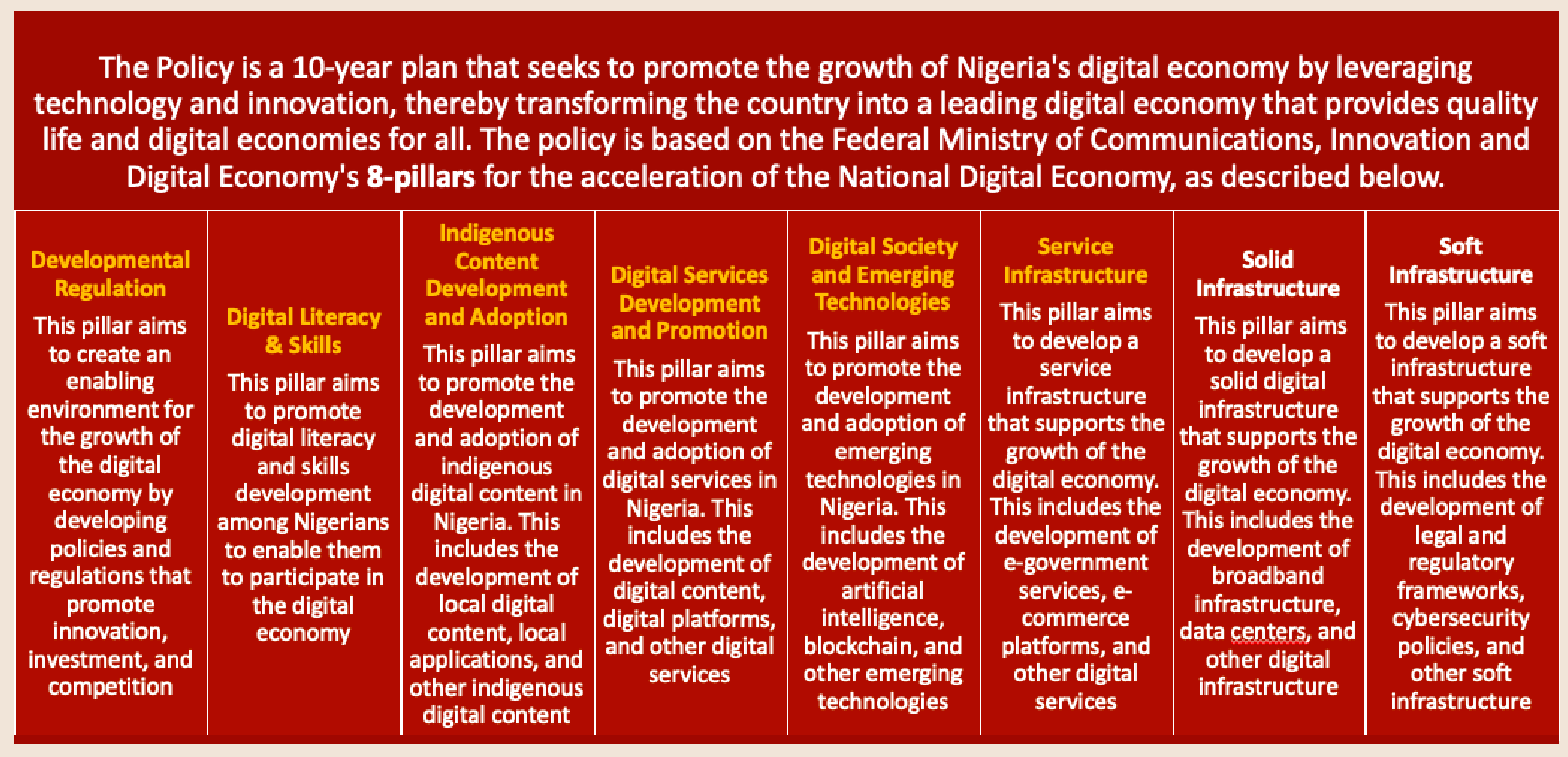
The research investigates stakeholder identification, selection, and involvement in digital economy policy development and implementation, focusing on Nigeria's National Digital Economy Policy and Strategy (2020-2030). It addresses the gap in effective stakeholder engagement in developing digital policies in sovereign developing countries. The study employs the Critical Systems Heuristics (CSH) framework to analyse stakeholder roles and involvement. The overarching aim of the PhD thesis is to develop a framework for beneficiary-focused policies applicable in developing nations – given their distinct power dynamics characteristics that often question the motivation for acting on behalf of the public. Through interviews with policymakers, implementers, and beneficiaries, the research explores the rationale behind stakeholder selection, power dynamics, expertise utilisation, and perceptions of legitimacy. The findings will contribute to a new framework incorporating evidence-backed criteria for identification, selection, and involvement, thereby offering guidelines for effective stakeholder engagement in developing national digital economy policies.

Keywords: Digital Economy, Startups, Critical Systems Heuristics, Stakeholder Engagement, Public Policy and Strategy, Policy Beneficiaries, Developing Countries, Nigeria

1 Introduction

1.1 Justification of the Study

This research investigates the intersectionality of public policy and digital technology, focusing on Nigeria's vibrant tech-driven start-up scene as a case study. Specifically, the study explores the motivations behind stakeholder involvement in developing Nigeria's National Digital Economy Policy and Strategy (2020-2030) (NDEPS, henceforth). It questions why certain stakeholders were selected and how their involvement influenced the policy's creation and implementation. The research aims to uncover the rationale for choosing specific stakeholders, including government institutions, tech hubs, start-ups, and related stakeholders, while excluding others. It seeks to understand the criteria used to determine ideal participants and how these choices shaped the NDEPS – and its eventual impact on tech hubs and digital tech-driven start-ups in the country.



The Policy is a 10-year plan that seeks to promote the growth of Nigeria's digital economy by leveraging technology and innovation, thereby transforming the country into a leading digital economy that provides quality life and digital economies for all. The policy is based on the Federal Ministry of Communications, Innovation and Digital Economy's **8-pillars** for the acceleration of the National Digital Economy, as described below.

Developmental Regulation	Digital Literacy & Skills	Indigenous Content Development and Adoption	Digital Services Development and Promotion	Digital Society and Emerging Technologies	Service Infrastructure	Solid Infrastructure	Soft Infrastructure
This pillar aims to create an enabling environment for the growth of the digital economy by developing policies and regulations that promote innovation, investment, and competition	This pillar aims to promote digital literacy and skills development among Nigerians to enable them to participate in the digital economy	This pillar aims to promote the development and adoption of indigenous digital content in Nigeria. This includes the development of local digital content, local applications, and other indigenous digital content	This pillar aims to promote the development and adoption of digital services in Nigeria. This includes the development of digital content, digital platforms, and other digital services	This pillar aims to promote the development and adoption of emerging technologies in Nigeria. This includes the development of artificial intelligence, blockchain, and other emerging technologies	This pillar aims to develop a service infrastructure that supports the growth of the digital economy. This includes the development of e-government services, e-commerce platforms, and other digital services	This pillar aims to develop a solid digital infrastructure that supports the growth of the digital economy. This includes the development of broadband infrastructure, data centers, and other digital infrastructure	This pillar aims to develop a soft infrastructure that supports the growth of the digital economy. This includes the development of legal and regulatory frameworks, cybersecurity policies, and other soft infrastructure

Table 1: An overview of the NDEPS and its eight pillars; this study is concerned with the first six pillars only

The study uses Critical Systems Heuristics (CSH) to analyse the "*what is*" question regarding stakeholder selection and will ultimately propose a "*what ought to be*" model. This will be achieved through interviews, document analysis, and thesis formulation, aiming to improve future policy development by understanding the dynamics of stakeholder engagement.

The study is crucial due to the scarcity of research examining the policy development processes and impact of national digital policies on digital tech-driven enterprises, considering that existing studies often overlook the nuanced dynamics within local ecosystems – especially in developing countries, like Nigeria, where the digital technology ecosystem is yet to mature (Ajala, 2023; Balogun & Adjei, 2018; Kayser, 2023). Consequently, little empirical evidence exists on how national digital economy policies directly impact these emerging ventures (Cao & Shi, 2021; Omon et al., 2024).

Narrowed down further, the case study revolves around Lagos-based Co-creation Hub (CcHub), arguably Africa's most consequential and impactful digital technology innovation and incubation center, to understand how the NDEPS has influenced CcHub's efforts in catalysing the founding and growth of digital tech-driven start-up Small, Medium and Micro Enterprises (SMMEs) (Ajala, 2023). Start-ups, particularly those driven by digital technology, are essential for economic growth and diversification (Atiase et al., 2020). For developing countries like Nigeria, understanding how tech policy can promote the founding, incubation, and growth of more digital tech-driven start-ups is crucial for economic development and job creation (Daraojimba et al., 2023; Ngene et al., 2021; Tikoudi, 2023).

1.2 Research Gap: Problematising the Thesis

Several studies have highlighted the role of technology incubation centres in facilitating the founding, incubation, and nurturing of SMMEs in Nigeria and beyond (Abubakar-Sadeeq et al., 2021; Ikebuaku & Dinbabo, 2018; Osabohien et al., 2022). Relatedly, other researchers have enunciated the benefits SMMEs can derive from such incubation hubs (Akhuemonkhan et al., 2014; Bubou & Okrigwe, 2011; Ofili, 2014; Xu, 2023). Furthermore, scholars have outlined the roles and outcomes of Nigeria's National Board for Technology Incubation's (NBTI) establishment of Technology Incubation Centres (TICs) across the six geopolitical zones of the country (Adelowo et al., 2015; Akpoviroro et al., 2021; Kaggwa, 2024).

While this rich tapestry of scholarly works has clearly painted the picture of Nigeria's technology incubation centre's relationship with government policies and strategies, on the one hand, and policy outcomes among the SMMEs, on the other hand, no study has yet linked the effects or impacts of the NDEPS regarding the founding, incubation, and nurturing of digital tech-driven start-ups since its implementation in 2020. Relatedly, no studies have investigated, assessed, or evaluated the relationship between digital economy policy development processes – specifically stakeholder identification, selection, and involvement, and the impact of the policy on beneficiaries (Aminullah et al., 2024; Dahlman et al., 2016; Landini & Marini, 2021). These gaps in research are notable, considering that the (Yaba) Lagos cluster is considered the most vibrant and successful of all TICs in the country (Adegbite & Adegbite, 2021; Akhuemonkhan et al., 2014; Osiakwan, 2017), and CcHub is consistently singled out as the most impactful tech innovation and incubation centre in Nigeria and beyond (Atiase et al., 2020; Friederici, 2018; Madichie & Hinson, 2022).

Additionally, understanding how the NDEPS, which has been operational since it’s 2020 launch (Commission, 2020), has impacted the founding, incubation, and nurturing of digital tech-driven start-ups in the Yaba cluster is crucial because empirical data will help policymakers to identify and evaluate where there are cogs in the wheels of progress, pivot quickly, and address any problems detected. However, to correctly appraise the impact, a need arises to first identify the stakeholders that participated in the development of the NDEPS, discover why they were involved, and understand the role they played in the processes.

1.3 Significance of the Study

The study's significance lies in its potential to provide data-driven insights into the identification, selection, and involvement of critical stakeholders in the development of digital economy policies in a developing nation, like Nigeria. Equally important, the study aims to provide insights into the relationship between the policy development processes and the anticipated outcomes, specifically regarding supporting the founding, incubation, and nurturing of digital tech-driven start-ups in Nigeria. The research outcome is articulated in three distinct, but related categories:

1.3.1 Policy

Contribute to a bottom-up, evidence-backed approach towards national digital economy policy formulation and implementation in developing countries, thereby ensuring that digital economy policies in developing countries are truly beneficial and contribute to sustainable development.

1.3.2 Theory

Develop a framework incorporating evidence-backed criteria for critical stakeholder identification, selection, and involvement, thereby offering guidelines for effective stakeholder engagement in developing national digital economy policies.

1.3.3 Practice

Contribute to policymakers' understanding of how to develop homegrown digital economy policies & strategies, taking into account the needs and realities of the beneficiaries and other critical stakeholders.

Furthermore, the study will contribute to the existing literature concerning the necessary stakeholders developing countries require for the development of original, (or the domestication of foreign) digital economy policies targeted at digital tech-driven start-ups. This contribution will enrich the literature and providing a basis for future research.

1.4 Study Aim

The overarching goal of the study is to ***develop a stakeholder identification, selection, and involvement framework*** suitable for formulating beneficiary-stakeholder-focused digital economy policies in developing countries.

1.5 Study Objective

To achieve the aim above, the study will examine the NDEPS, using the Critical Systems Heuristics (CSH) framework (Hutcheson et al., 2024; Ulrich & Reynolds, 2010) to critically analyse various stakeholders' roles and degree of involvement in the policy formulation and implementation processes and, on the back of that, co-create a beneficiary-stakeholder-focused digital economy policy development framework.

1.6 Research Questions

The study will aim to answer one main research question: ***how are critical stakeholders identified and selected, and to what degree are they involved in the development and implementation of a digital economy policy in a developing country, and how should they?***

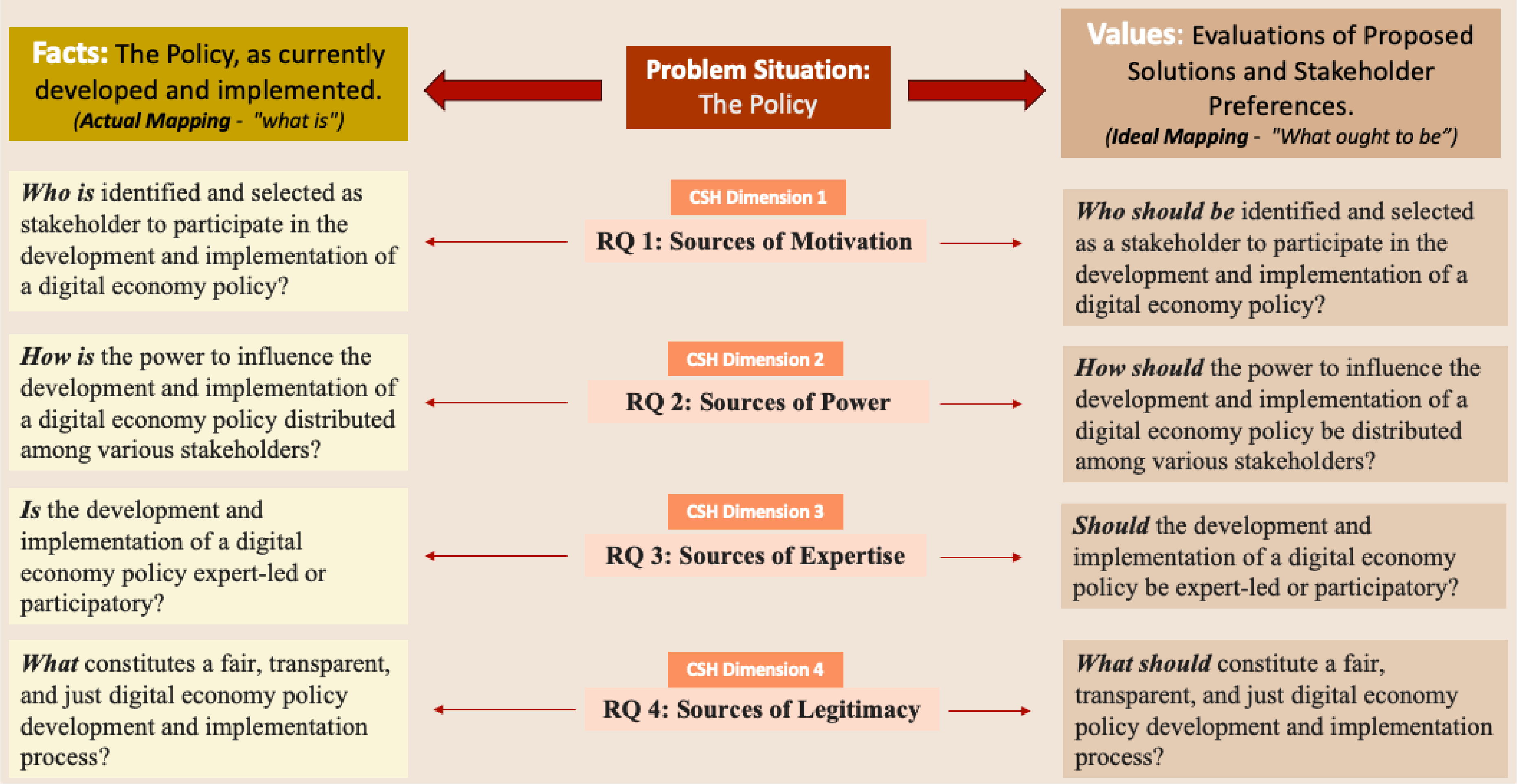
Additionally, below are four supporting research questions *implicitly* aligned to the CSH framework's four dimensions and their respective "*what is*" and "*what ought to be*" dynamics:

- **RQ 1:** What is the ***rationale*** for identifying and selecting various stakeholders, and to what degree did those reasons influence respective participants' involvement in the development and implementation of the NDEPS?

- **RQ 2:** How did the ***distribution of power*** among stakeholders impact their ability to influence the development and implementation of the NDEPS?

- **RQ 3:** How were diverse ***stakeholder expertise*** represented and utilised in the development and implementation of the NDEPS, and through what mechanisms was that expertise evaluated to ensure their appropriateness?

- **RQ 4:** How do varying ***stakeholder perceptions*** of procedural fairness, transparency, and stakeholder representation in the development and implementation of the NDEPS influence the policy's impact and effectiveness?



Facts: The Policy, as currently developed and implemented.
(Actual Mapping - "what is")
Problem Situation: The Policy
Values: Evaluations of Proposed Solutions and Stakeholder Preferences.
(Ideal Mapping - "What ought to be")
Who is identified and selected as stakeholder to participate in the development and implementation of a digital economy policy?
CSH Dimension 1
RQ 1: Sources of Motivation
Who should be identified and selected as a stakeholder to participate in the development and implementation of a digital economy policy?
How is the power to influence the development and implementation of a digital economy policy distributed among various stakeholders?
CSH Dimension 2
RQ 2: Sources of Power
How should the power to influence the development and implementation of a digital economy policy be distributed among various stakeholders?
Is the development and implementation of a digital economy policy expert-led or participatory?
CSH Dimension 3
RQ 3: Sources of Expertise
Should the development and implementation of a digital economy policy be expert-led or participatory?
What constitutes a fair, transparent, and just digital economy policy development and implementation process?
CSH Dimension 4
RQ 4: Sources of Legitimacy
What should constitute a fair, transparent, and just digital economy policy development and implementation process?

Figure 1: Four RQs described as Boundary Questions and mapped against Actual Mapping and Ideal Mapping Boundary Categories. (Source: Own Creation)

1.7 Who are Critical Stakeholders and Why Study Them?

Critical stakeholders are vital in public policy development (Blanchard et al., 2015; Onwujekwe et al., 2015; Wagg & Simeonova, 2021). They are individuals, organisations, or government bodies whose support or opposition can significantly influence a policy's creation and impact. These stakeholders have a vested interest in the policy area and can exert considerable influence on its outcome.

Key characteristics of critical stakeholders include:

- **Directly affected**: These groups experience the policy's consequences firsthand (e.g., start-ups).
- **Power and influence**: They can shape public opinion or exert political pressure (e.g., the Nigerian government).
- **Expertise**: They possess specialized knowledge relevant to the policy issue (e.g., policymakers).
- **Legitimacy**: They are recognised as credible representatives of a particular segment of society (e.g., start-ups).

Furthermore, the study adopts a community approach to digital economy policy framework development primarily because the literature indicates that critical stakeholders – including beneficiaries, and beyond government officials and technical experts, are crucial to policy development, hence the emphasis on active participation of these critical stakeholders throughout the entire process. This approach ensures that those affected by the policy have a say in its design and implementation, thereby fostering a sense of ownership and collaboration, and leading to more effective and impactful policies.

2 Study Methodology

This section briefly enunciates the research methodology. The Research Onion framework diagram presents the systematic research approach adopted.

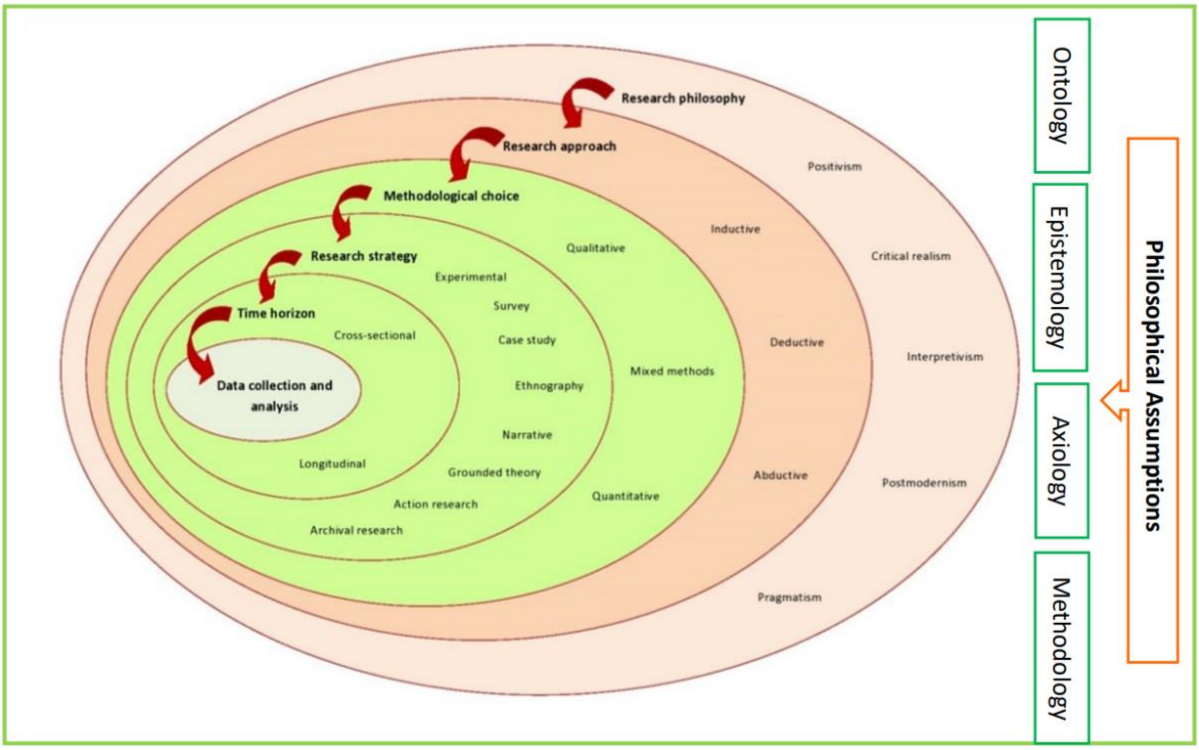


Figure 2: Mark Saunders' Research Onion (Saunders & Tosey, 2012); developed further by Dissanayake (2023) to include the four Philosophical Assumptions

The study is grounded in interpretivism, acknowledging the subjective and socially constructed nature of reality. An inductive approach is employed, deriving theories and insights from the acquired data. The study adopts a qualitative research method to explore the nuances and complexities of the phenomena. The case study method is utilised, allowing for a detailed analysis of a specific event within its real-life setting.

Furthermore, a cross-sectional time horizon strategy is employed, offering a snapshot of the present status of the NDEPS, its development processes pre-2020 implementation, and its impact on start-ups. Data collection rely on primary sources, such as semi-structured interviews with stakeholders from FMCIDE, NITDA, CcHub, and selected start-ups, and secondary sources, such as policy documents and academic literature. The collected data is subjected to a thematic analysis to identify patterns and themes. Ethical considerations are prioritised throughout the study, ensuring anonymity, confidentiality, and academic integrity.

2.1 Study Participants

Due to time and resource constraints, the study limited its participants to those deemed most critical following a pilot study. This included key figures from the Federal Ministry of Communications and Digital Economy (FMCIDE) and the National Information Technology

Development Agency (NITDA), responsible for the NDEPS policy development and implementation, respectively. Others are CcHub, and five digital tech-driven start-ups incubated at CcHub. This focused approach aimed to gather sufficient data for the research while acknowledging the challenges in securing broader stakeholder involvement.

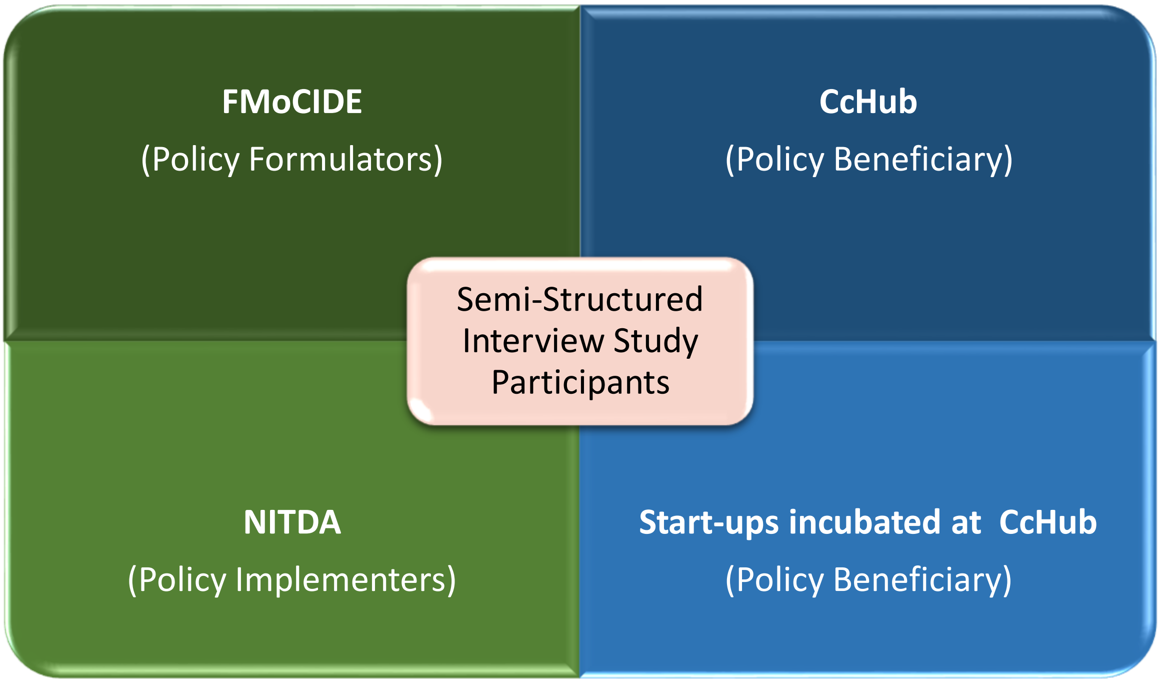
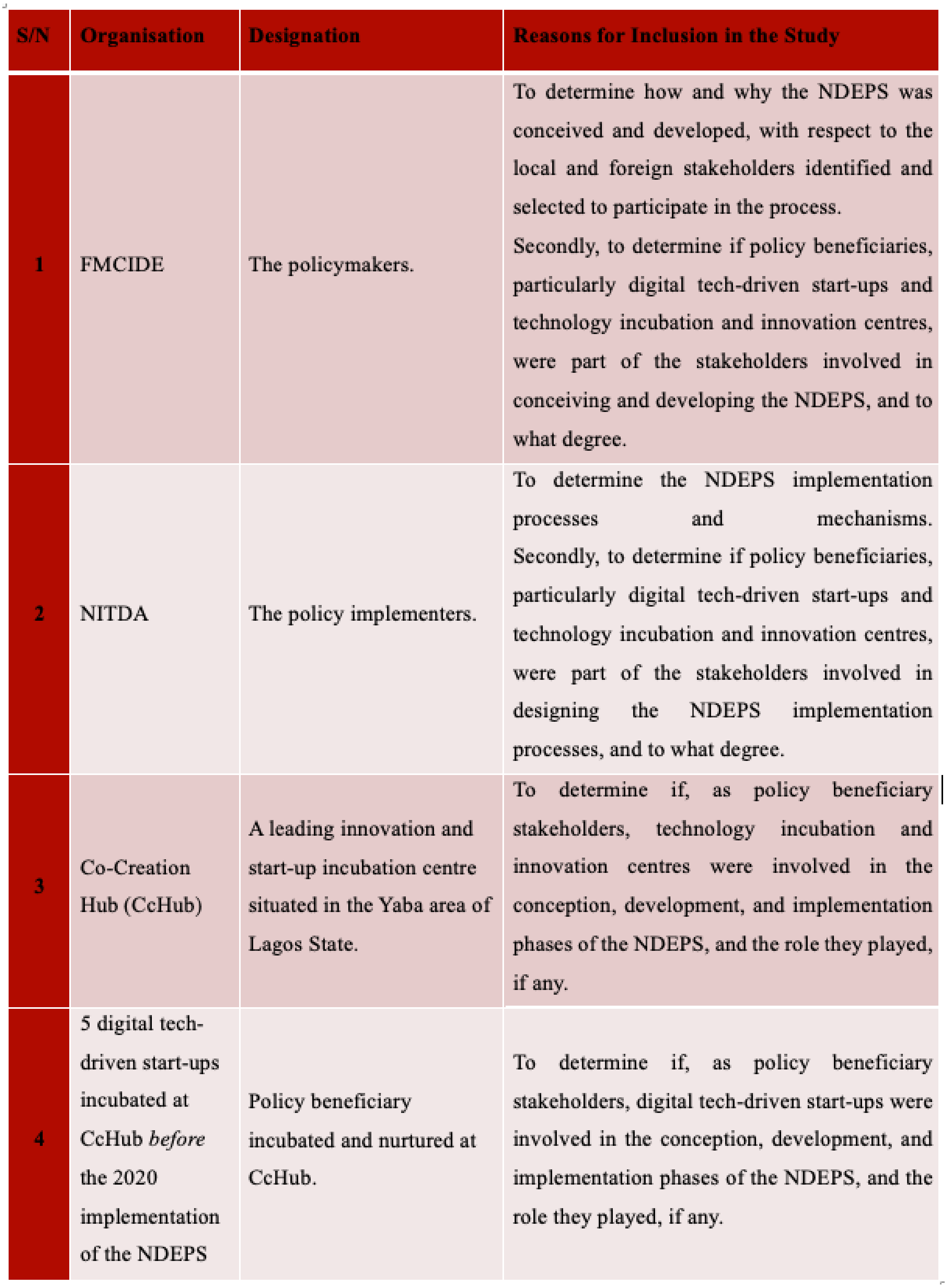


Figure 3: Matrix showing the study participants and respective designations in relation to the NDEPS. (Source: Own Creation)

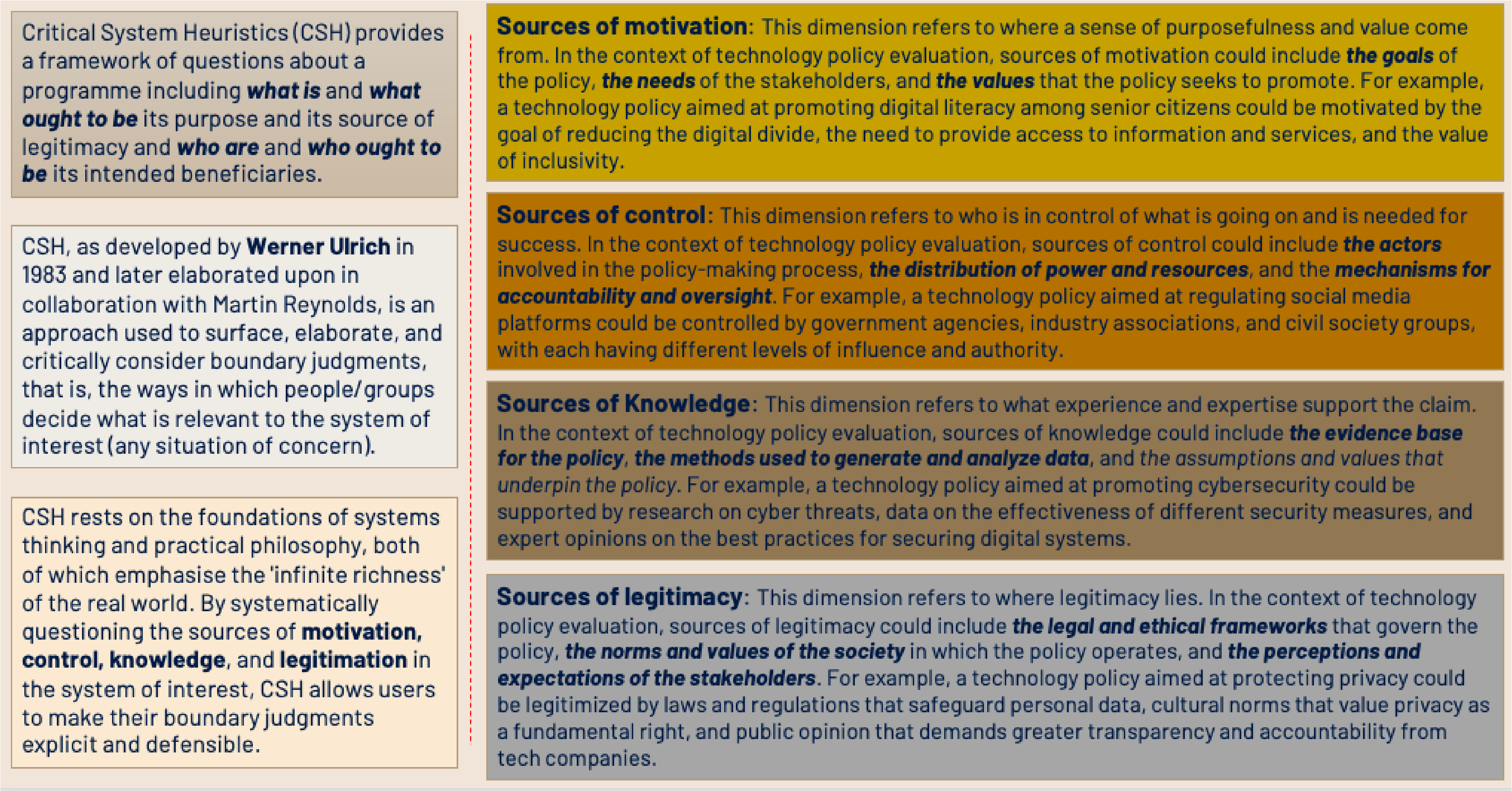


S/N	Organisation	Designation	Reasons for Inclusion in the Study
1	FMCIDE	The policymakers.	To determine how and why the NDEPS was conceived and developed, with respect to the local and foreign stakeholders identified and selected to participate in the process. Secondly, to determine if policy beneficiaries, particularly digital tech-driven start-ups and technology incubation and innovation centres, were part of the stakeholders involved in conceiving and developing the NDEPS, and to what degree.
2	NITDA	The policy implementers.	To determine the NDEPS implementation processes and mechanisms. Secondly, to determine if policy beneficiaries, particularly digital tech-driven start-ups and technology incubation and innovation centres, were part of the stakeholders involved in designing the NDEPS implementation processes, and to what degree.
3	Co-Creation Hub (CcHub)	A leading innovation and start-up incubation centre situated in the Yaba area of Lagos State.	To determine if, as policy beneficiary stakeholders, technology incubation and innovation centres were involved in the conception, development, and implementation phases of the NDEPS, and the role they played, if any.
4	5 digital tech-driven start-ups incubated at CcHub *before* the 2020 implementation of the NDEPS	Policy beneficiary incubated and nurtured at CcHub.	To determine if, as policy beneficiary stakeholders, digital tech-driven start-ups were involved in the conception, development, and implementation phases of the NDEPS, and the role they played, if any.

Table 2: Table showing the rationale for including respective participants in the study. (Source: Own Creation)

3 Theoretical Framework: Critical Systems Heuristics

Critical Systems Heuristics (CSH) is a time-tested framework for analysing complex systems and is particularly useful in policy evaluation. Developed by Werner Ulrich in 1983 (Ulrich & Reynolds, 2010), CSH goes beyond traditional methods by emphasising critical reflection and inclusivity. It uses four dimensions and 12 questions to investigate the current realities and ideal situations regarding stakeholders and their relationship with the phenomena under review.



Critical System Heuristics (CSH) provides a framework of questions about a programme including ***what is*** and ***what ought to be*** its purpose and its source of legitimacy and ***who are*** and ***who ought to be*** its intended beneficiaries.

CSH, as developed by **Werner Ulrich** in 1983 and later elaborated upon in collaboration with Martin Reynolds, is an approach used to surface, elaborate, and critically consider boundary judgments, that is, the ways in which people/groups decide what is relevant to the system of interest (any situation of concern).

CSH rests on the foundations of systems thinking and practical philosophy, both of which emphasise the 'infinite richness' of the real world. By systematically questioning the sources of **motivation, control, knowledge**, and **legitimation** in the system of interest, CSH allows users to make their boundary judgments explicit and defensible.

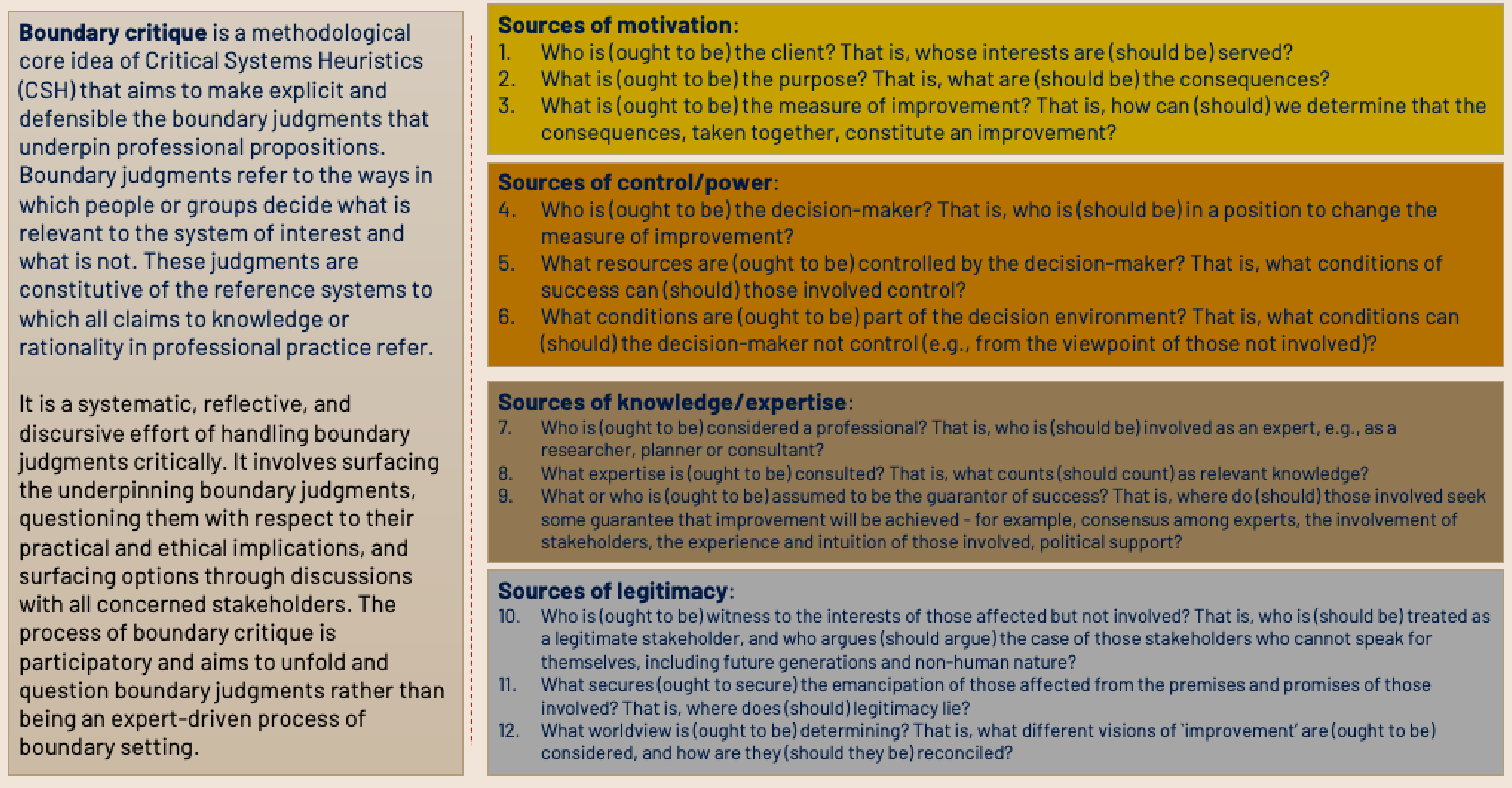
Sources of motivation: This dimension refers to where a sense of purposefulness and value come from. In the context of technology policy evaluation, sources of motivation could include ***the goals*** of the policy, ***the needs*** of the stakeholders, and ***the values*** that the policy seeks to promote. For example, a technology policy aimed at promoting digital literacy among senior citizens could be motivated by the goal of reducing the digital divide, the need to provide access to information and services, and the value of inclusivity.

Sources of control: This dimension refers to who is in control of what is going on and is needed for success. In the context of technology policy evaluation, sources of control could include ***the actors*** involved in the policy-making process, ***the distribution of power and resources***, and the ***mechanisms for accountability and oversight***. For example, a technology policy aimed at regulating social media platforms could be controlled by government agencies, industry associations, and civil society groups, with each having different levels of influence and authority.

Sources of Knowledge: This dimension refers to what experience and expertise support the claim. In the context of technology policy evaluation, sources of knowledge could include ***the evidence base for the policy***, ***the methods used to generate and analyze data***, and *the assumptions and values that underpin the policy*. For example, a technology policy aimed at promoting cybersecurity could be supported by research on cyber threats, data on the effectiveness of different security measures, and expert opinions on the best practices for securing digital systems.

Sources of legitimacy: This dimension refers to where legitimacy lies. In the context of technology policy evaluation, sources of legitimacy could include ***the legal and ethical frameworks*** that govern the policy, ***the norms and values of the society*** in which the policy operates, and ***the perceptions and expectations of the stakeholders***. For example, a technology policy aimed at protecting privacy could be legitimized by laws and regulations that safeguard personal data, cultural norms that value privacy as a fundamental right, and public opinion that demands greater transparency and accountability from tech companies.

Figure 4: A Brief Overview of the Critical Systems Heuristics Framework



Boundary critique is a methodological core idea of Critical Systems Heuristics (CSH) that aims to make explicit and defensible the boundary judgments that underpin professional propositions. Boundary judgments refer to the ways in which people or groups decide what is relevant to the system of interest and what is not. These judgments are constitutive of the reference systems to which all claims to knowledge or rationality in professional practice refer.

It is a systematic, reflective, and discursive effort of handling boundary judgments critically. It involves surfacing the underpinning boundary judgments, questioning them with respect to their practical and ethical implications, and surfacing options through discussions with all concerned stakeholders. The process of boundary critique is participatory and aims to unfold and question boundary judgments rather than being an expert-driven process of boundary setting.

Sources of motivation:

1. Who is (ought to be) the client? That is, whose interests are (should be) served?
2. What is (ought to be) the purpose? That is, what are (should be) the consequences?
3. What is (ought to be) the measure of improvement? That is, how can (should) we determine that the consequences, taken together, constitute an improvement?

Sources of control/power:

4. Who is (ought to be) the decision-maker? That is, who is (should be) in a position to change the measure of improvement?
5. What resources are (ought to be) controlled by the decision-maker? That is, what conditions of success can (should) those involved control?
6. What conditions are (ought to be) part of the decision environment? That is, what conditions can (should) the decision-maker not control (e.g., from the viewpoint of those not involved)?

Sources of knowledge/expertise:

7. Who is (ought to be) considered a professional? That is, who is (should be) involved as an expert, e.g., as a researcher, planner or consultant?
8. What expertise is (ought to be) consulted? That is, what counts (should count) as relevant knowledge?
9. What or who is (ought to be) assumed to be the guarantor of success? That is, where do (should) those involved seek some guarantee that improvement will be achieved - for example, consensus among experts, the involvement of stakeholders, the experience and intuition of those involved, political support?

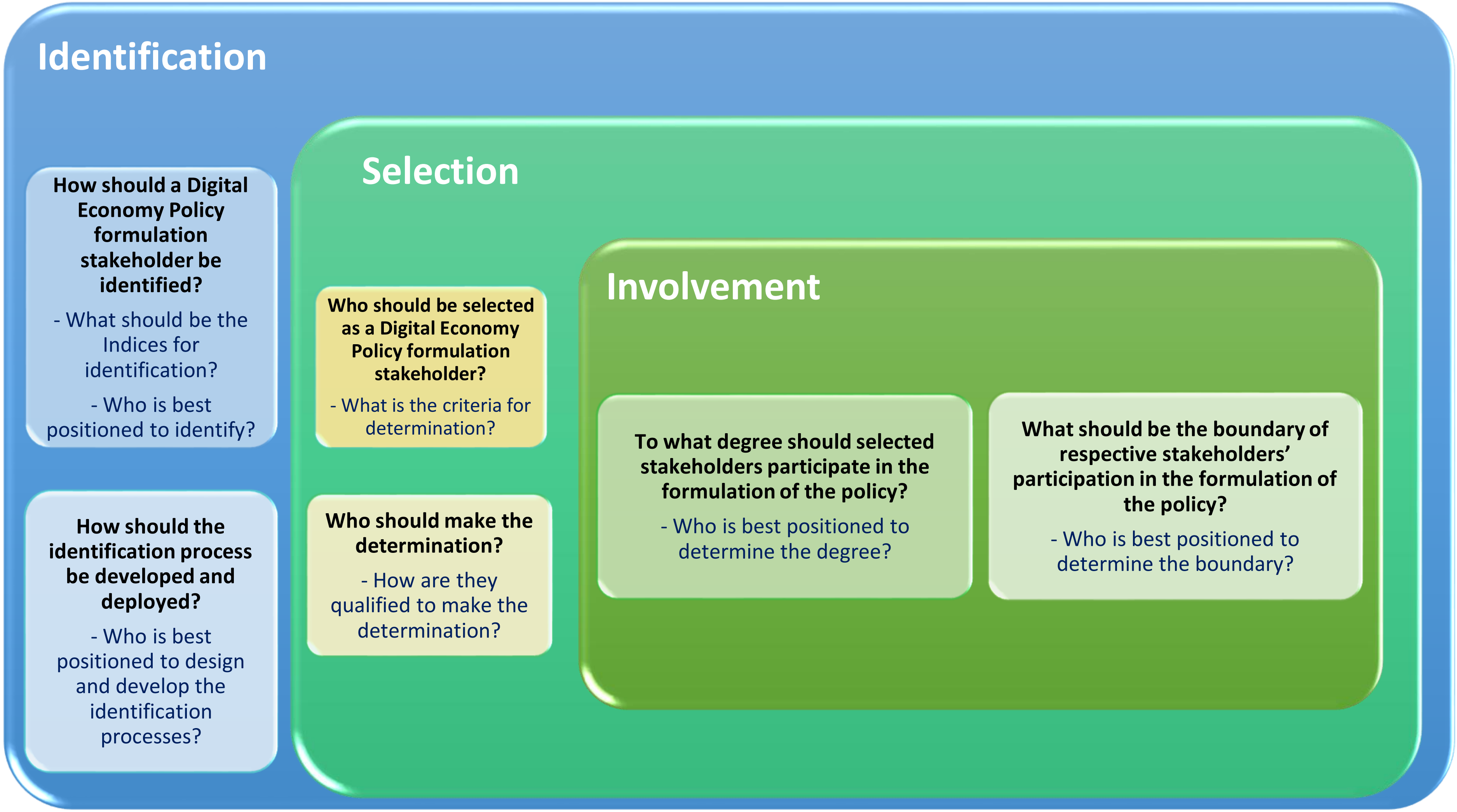
Sources of legitimacy:

10. Who is (ought to be) witness to the interests of those affected but not involved? That is, who is (should be) treated as a legitimate stakeholder, and who argues (should argue) the case of those stakeholders who cannot speak for themselves, including future generations and non-human nature?
11. What secures (ought to secure) the emancipation of those affected from the premises and promises of those involved? That is, where does (should) legitimacy lie?
12. What worldview is (ought to be) determining? That is, what different visions of 'improvement' are (ought to be) considered, and how are they (should they be) reconciled?

Figure 5: A Snapshot of the 4 Dimensions and 12 Questions in the Critical Systems Heuristics Framework

CSH is ideal for this study because it allows for the analysis of the NDEPS policy, examining both the methods used for stakeholder identification, selection, and involvement, and the outcomes achieved. It helps uncover underlying assumptions, decision-making, and power dynamics ingrained in the policy development process. CSH's emphasis on boundary judgments promotes transparency and ethical considerations, ensuring a robust and thorough evaluation of the NDEPS and its impact on the start-up ecosystem. As such, the study employs CSH as the primary evaluative framework. While this approach provides valuable insights, it does not encompass all possible evaluation methodologies.

4 Study End Goal: Work-in-Progress



Identification

How should a Digital Economy Policy formulation stakeholder be identified?

- What should be the Indices for identification?

- Who is best positioned to identify?

How should the identification process be developed and deployed?

- Who is best positioned to design and develop the identification processes?

Selection

Who should be selected as a Digital Economy Policy formulation stakeholder?

- What is the criteria for determination?

Who should make the determination?

- How are they qualified to make the determination?

Involvement

To what degree should selected stakeholders participate in the formulation of the policy?

- Who is best positioned to determine the degree?

What should be the boundary of respective stakeholders' participation in the formulation of the policy?

- Who is best positioned to determine the boundary?

Figure 6: Study End Goal: To Develop a Digital Economy Policy Stakeholder Identification, Selection, and Involvement Framework (DEPSISIF) suitable for formulating beneficiary-stakeholder-focused digital economy policies in developing countries.

5 Conclusion

The literature indicates that there is need for research into the digital economy in developing countries, like Nigeria. Evidence from past scholarly works also shows that public policy outcomes are influenced by the policy development processes, specifically how critical stakeholders are identified, selected, and involved in the policy formulation phases. Consequently, this study endeavours to shed light on the intricate interplay between the NDEPS and its development processes, tech start-up incubation hubs, and digital tech-driven start-ups, using CSH as an evaluative framework. As a final output, this research aims to develop a framework for improving digital economy policy formulation in developing countries by promoting evidence-based, bottom-up approaches that consider stakeholder needs.

6 References

Abubakar-Sadeeq, A., Othman, A. U., Audu, A. S., Ramalan, M. I., & Abdullahi, I. (2021). Impact of Technology Incubation Programme in Promoting Entrepreneurship in Nigeria. *ECONOMICS*, *10*(4), 105. https://doi.org/10.11648/j.eco.20211004.11

Adegbite, O., & Adegbite, O. (2021). Business incubators and small enterprise development in Nigeria. *Perspectives on Industrial Development in Nigeria: Issues, Challenges and Hard Choices*, 285-301.

Adelowo, C. M., Ilori, M. O., Siyanbola, W. O., & Oluwale, B. A. (2015). Technological Learning Mechanisms in Nigeria's Technology Incubation Centre. *African Journal of Economic and Management Studies*, *6*(1), 72-89. https://doi.org/10.1108/ajems-10-2014-0071

Ajala, A. A. (2023). Building the Dynamic Capabilities of SMEs in the Field of Food and Beverages in Nigeria: The Case of CcHUB in the Lagos Entrepreneurial Ecosystem.

Akhuemonkhan, I., Raimi, L., Patel, A., & Fadipe, A. O. (2014). Harnessing the Potentials of Technology Incubation Centres (TICs) as Tools for Fast-Tracking Entrepreneurship Development and Actualisation of the Vision 20:2020 in Nigeria. *Humanomics*, *30*(4), 349-372. https://doi.org/10.1108/h-11-2013-0069

Akpoviroro, K. S., Oba-Adenuga, O. A., & Akanmu, P. M. (2021). The Role of Business Incubation in Promoting Entrepreneurship and SMEs Development. *Management and Entrepreneurship Trends of Development*, *2*(16). https://doi.org/10.26661/2522-1566/2021-1/16-07

Aminullah, E., Fizzanty, T., Nawawi, N., Suryanto, J., Pranata, N., Maulana, I., Ariyani, L., Wicaksono, A., Suardi, I., & Azis, N. L. L. (2024). Interactive components of digital MSMEs ecosystem for inclusive digital economy in Indonesia. *Journal of the Knowledge Economy*, *15*(1), 487-517.

Atiase, V. Y., Kolade, O., & Liedong, T. A. (2020). The emergence and strategy of tech hubs in Africa: Implications for knowledge production and value creation. *Technological Forecasting and Social Change*, *161*, 120307.

Balogun, T., & Adjei, E. (2018). Challenges of Digitization of the National Archives of Nigeria. *Information Development*, *35*(4), 612-623. https://doi.org/10.1177/0266666918778099

Blanchard, J. W., Petherick, J. T., & Basara, H. (2015). Stakeholder Engagement. *American Journal of Preventive Medicine*, *48*(1), S44-S46. https://doi.org/10.1016/j.amepre.2014.09.025

Bubou, G. M., & Okrigwe, F. N. (2011). Fostering Technological Entrepreneurship for Socioeconomic Development: A Case for Technology Incubation in Bayelsa State, Nigeria. *Journal of Sustainable Development*, *4*(6). https://doi.org/10.5539/jsd.v4n6p138

Cao, Z., & Shi, X. (2021). A systematic literature review of entrepreneurial ecosystems in advanced and emerging economies. *Small Business Economics*, *57*, 75-110.

Commission, N. C. (2020). National Digital Economy Policy and Strategy (2020-2030).

Dahlman, C., Mealy, S., & Wermelinger, M. (2016). Harnessing the digital economy for developing countries.

Daraojimba, C., Abioye, K. M., Bakare, A. D., Mhlongo, N. Z., Onunka, O., & Daraojimba, D. O. (2023). Technology and innovation to growth of entrepreneurship and financial boost: a decade in review (2013-2023). *International Journal of Management & Entrepreneurship Research*, *5*(10), 769-792.

Friederici, N. (2018). Hope and hype in Africa's digital economy: The rise of innovation hubs. *Digital economies at global margins*, 193-222.

Hutcheson, M., Morton, A., & Blair, S. (2024). Critical systems heuristics: a systematic review. *Systemic Practice and Action Research*, *37*(4), 499-514.

Ikebuaku, K., & Dinbabo, M. F. (2018). Beyond Entrepreneurship Education: Business Incubation and Entrepreneurial Capabilities. *Journal of Entrepreneurship in Emerging Economies*, *10*(1), 154-174. https://doi.org/10.1108/jeee-03-2017-0022

Kaggwa, S. (2024). Evaluating the Efficacy of Technology Incubation Centres in Fostering Entrepreneurship: Case Studies From the Global Sout. *International Journal of Management & Entrepreneurship Research*, *6*(1), 46-68. https://doi.org/10.51594/ijmer.v6i1.695

Kayser, K. (2023). Digital Start-Up Ecosystems: A Systematic Literature Review and Model Development for South Africa. *Sustainability (Switzerland)*, *15*(16), 12513. https://doi.org/10.3390/su151612513

Landini, D., & Marini, M. (2021). Policy analytics for collaborative networks: the role of stakeholders in the Italian Digital Civilian service.

Madichie, N. O., & Hinson, R. E. (2022). Value Co-creation of Places and Spaces in Africa's Creative Hubs. In *The Creative Industries and International Business Development in Africa* (pp. 91-108). Emerald Publishing Limited.

Ngene, G., Pinet, M., & Maclay, C. (2021). *Strengthening youth livelihoods and enterprise innovation in Africa's digital era*.

Ofili, O. U. (2014). Challenges Facing Entrepreneurship in Nigeria. *International Journal of Business and Management*, *9*(12). https://doi.org/10.5539/ijbm.v9n12p258

Omon, U. K., Fletcher, G., & Albakri, M. (2024). Mapping and Visualising the Digital Economy in The Context ofDeveloping Countries: A Bibliometric AnalysisDeveloping Countries: A Bibliometric Analysis. [https://www.researchgate.net/publication/381215758_Mapping_and_Visualising_the_Digital_Economy_in_The_Context_of_Developing_Countries_A_Bibliometric_Analysis]. UK Academy forInformation Systems Conference Proceedings

Onwujekwe, O., Uguru, N., Russo, G., Etiaba, E., Mbachu, C., Mirzoev, T., & Uzochukwu, B. (2015). Role and Use of Evidence in Policymaking: An Analysis of Case Studies From the Health Sector in Nigeria. *Health Research Policy and Systems*, *13*(1). https://doi.org/10.1186/s12961-015-0049-0

Osabohien, R., Worgwu, H., Adediran, O., & Soomro, J. A. (2022). Social Entrepreneurship and Future Employment in Nigeria. *International Social Science Journal*, *73*(250), 927-937. https://doi.org/10.1111/issj.12360

Osiakwan, E. M. (2017). The KINGS of Africa's digital economy. *Digital Kenya*, 55-92.

Saunders, M., & Tosey, P. (2012). Research Design. In: Academia.

Tikoudi, A. (2023). *ICT-based Social Innovation in Africa: the case of Rwanda* Norwegian University of Life Sciences].

Ulrich, W., & Reynolds, M. (2010). Critical systems heuristics. In *Systems approaches to managing change: A practical guide* (pp. 243-292). Springer.

Wagg, S., & Simeonova, B. (2021). A Policy-Level Perspective to Tackle Rural Digital Inclusion. *Information Technology and People*, *35*(7), 1884-1911. https://doi.org/10.1108/itp-01-2020-0047

Xu, Y. (2023). Can the Establishment of an Innovative City Improve the Level of Technological Entrepreneurship? *Plos One*, *18*(10), e0289806. https://doi.org/10.1371/journal.pone.0289806

Good for self, but what about your team? Job crafting in the context of artificial intelligence

Swati Garg
Keele Business School
Keele University
Staffordshire, UK
s.garg@keele.ac.uk

Akanksha Malik Jamwal
Guildhall School of Business and Law
London Metropolitan University
London, UK
a.jamwal@londonmet.ac.uk

Abstract

Artificial intelligence (AI) has garnered significant attention from both academicians and practitioners belonging to various industries. It has reshaped structure of traditional work and presented promising potential approaches to enhance work systems. Through this paper, we aim to understand the impact of AI-enabled job crafting on performance. We propose that when employees craft their jobs using AI, it will lead to an increased self-efficacy and meaningfulness of work. This state can impact the employee's and their team's performance depending upon AI advocacy by organisation and the level of employee's commitment to the organisation and its values.

Keywords: Artificial intelligence, AI-enabled job crafting. meaningfulness of work, performance, self-efficacy, teams

1.0 Introduction

Defined as "a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation (Kaplan & Haenlein, 2019, p. 17)", artificial intelligence (AI) has rapidly transformed work and workplaces across various industries such as hospitality, healthcare, finance, manufacturing and retail (Dwivedi et al., 2023; Ooi et al., 2023). In layman terms, AI represents making use of machine learning, deep learning and natural language processing algorithms to analyse data, predict patterns and make decisions based on those predicted patterns (Kaplan & Haenlein, 2019). The latest advancement in AI, generative AI, has made AI accessible to all employees, irrespective of their occupations and level of position within their organisations. This is because generative AI makes use of huge datasets and produces human-like outputs (Noy and Zhang, 2023). This helps organisations enhance their performance in terms of efficiency and quality of work (Gambacorta et al., 2024). Advanced tools and resources based on AI and generative AI have transformed workflows (Dell'Acqua et al., 2023), by assisting employees in automating their time-consuming tasks, responding to their queries and doubts, helping them brainstorm ideas and create newer content (Ross et al., 2023). Increased usage of AI-based tools has resulted in changes in job design for employees (Tong et al., 2021). Appreciation and acknowledgement of these changes in jobs and job designs has led to new discussions on the advantages of integrating AI in workplaces and the implications of the same for a variety of stakeholders including leaders, designers and users of AI-based systems and business and clients (Parker and Grote, 2022).

A major advantage of AI that is being repeatedly discussed by practitioners and researchers alike is its ability to automate redundant work and create space and energy for creative thinking (Eapen et al., 2023). To exemplify this, people working in the domain of marketing and advertising are now using AI-based tools like ChatGPT and Midjourney to generate marketing content including social media posts, graphic posters and campaigns. Similarly, people working in the software development domain are now using tools like Co-Pilot to write product development codes. By reducing the time spent of some of the core tasks, employees are now able to dedicate more energy and space to strategic and creative tasks. This is expected to increase not just employees' productivity, but also their work engagement and role innovation.

AI adoption by employees and employers has revealed some additional benefits in terms of skill acquisition (Cazzaniga et al., 2024). As organisations strive to adopt AI in their processes and strategies, employees find themselves working in a new world where they are compelled to become AI literate at first, and then proficient and practitioner gradually. This upskilling and gradual confidence with using AI-based capabilities contributes to organisational agility as the upskilled employees are able to make faster data-based decisions in an increasingly competitive world (Sharma et al., 2022).

In the functional domain of human resource management, research has shown that AI is being used in several core HRM functions such as recruitment, career development, employee engagement and performance appraisal and management (Ardichvili et al., 2024; Chowdhury et al., 2024). AI is now being used to respond to job applicants' questions, onboard new employees, design evidence-based pay and compensation packages, suggest training and development opportunities to employees, etc. (Chowdhury et al., 2024; Yorks and Rotaroti, 2021). There is plethora of research on benefits of AI adoption for the HRM function in organisations (Budhwar et al., 2023), but research on AI adoption by employees and the possible consequences for them is still in its nascent stage.

There is a new research interest on organisational AI integration and its impact on employees. Researchers are making efforts to study the impact on employees' work engagement and trust in AI (Dutta et al., 2023; Glikson and Woolley, 2020). However, these studies are focused on individual level outcome variables. We propose that in addition to individual level outcomes, team-level outcomes should also be given due consideration in early research on AI, as employees are increasingly working in an inter-connected environment which is characteristic of task interdependencies and collaborative work designs.

In our study, we focus on employees' job crafting behaviour in the context of AI adoption. We conceptualise AI-enabled job crafting as employees' efforts to use AI to modify their jobs to increase their fitment with the job and meet their career development needs and desires. We propose that AI-enabled job crafting will enhance employees' appraisals of their capabilities and their work meaningfulness, and this will impact their attitude, behaviour and well-being. In explaining our proposed research model, we will use social cognitive theory and self-determination theory. The study is expected to contribute to HRM practitioners and employees as it will provide evidence on how employees use AI to improve their jobs and how this usage affects their performance from an individual and team perspective.

In the following sections, we lay the foundation of our proposed research model. We explain the meaning of job crafting, the conceptualisation of AI-enabled job crafting and our proposed research model. In doing so, we use established theoretical frameworks of social cognitive theory and self-determination theory.

2.0 Development of Conceptual Model

2.1 What is Job Crafting?

Crant's (2000) seminal review on proactive behaviour encouraged researchers to conduct deeper contextualised research on proactive behaviour constructs. Parker and Collins (2010) further emphasized the importance of employee proactivity for both employees and employers. In the proactive behaviour literature, several constructs have emerged during the past few decades. All these constructs focus on employees' self-started initiatives undertaken by them to improve their situation at work (current or future). Taking charge, feedback seeking, voice, selling issues and job crafting are some examples of these proactive behaviour constructs.

The focus of the current paper is on job crafting. Wrzesniewski and Dutton (2001) defined job crafting as "actions [that] employees take to shape, mould, and redefine their jobs (p. 180)." They further proposed that employees pursue three types of job crafting behaviours: (a) task crafting that represents their efforts to make changes to the core tasks of their jobs, (b) relationship crafting, that represents their efforts to make changes to the number and quality of interactions on their jobs, and (c) cognitive crafting, that represents their efforts to change their perceptions and thoughts about their jobs. They argued that pursuit of these job crafting behaviours would result in enhanced work meaningfulness and work identity. Almost a decade later, Tims, Bakker and Derks (2012) linked job crafting with the job demands-resources theory and defined it as "the changes that employees may make to balance their job demands and job resources with their personal abilities and needs (p. 174)." According to this conceptualisation, job crafting signifies modification of job resources and demands by employees. They further proposed that employees may expand on resources like job autonomy, feedback, challenging work, etc. and decrease demands like time pressure and work overload.

Both the definitions have garnered significant research attention in the last two decades. Several studies have examined the motivations and outcomes of job crafting behaviour using these two conceptualisations. Resultantly, there has been confusion as to which conceptualisation to follow, which has been addressed by the recent meta-analyses and literature reviews on job crafting (Bruning and Campion, 2018; Zhang and Parker, 2019).These reviews have synthesized both the conceptualisations and operationalised job crafting as employees' proactive efforts of changing or modifying their jobs to improve their jobs for themselves. Per se, job crafting is now defined as a self-initiated behaviour representing modifications in jobs made by employees to meet their needs and desires. These changes can be made in tasks, relationships, thoughts, etc. and can represent an incremental change in resources or a reductional change in hindering demands.

Researchers have suggested employees to pursue job crafting behaviour to proactively cater to their needs and thrive in today's complex job environment (Demerouti et al., 2014). Job crafting helps employees enhance their work meaningfulness and person-organisation fit (Wrześniewski & Dutton, 2001; Zhang & Parker, 2019). which enhances their performance, well-being and work engagement (Tims et al., 2013; Tims et al., 2022). Several studies reveal different predictors and outcomes for job crafting behaviour. Individual variables such as proactive personality and regulatory focus have a positive relationship with job crafting (Bakker et al., 2012; Petrou, 2013). Additionally, people who want to satisfy their additional callings (Batova, 2018), and indemnify any missed callings (Berg et al., 2010) may also pursue job crafting behaviour. Research shows that these job crafters accrue positive benefits from crafting when they appraise personal factors like confidence positively (Lazazzara et al, 2020). When the same personal factors are appraised negatively, for example: a lack of confidence is appraised, then crafters may accrue negative consequences like stress (Lazazzara et al, 2020).

2.2 Impact of AI on Job Crafting

The advent of AI has revolutionized how employees engage with their work and craft their roles. The job crafting literature emphasized that employees are not merely passive recipients of environmental changes, but they actively construct or craft their roles (Tims et al., 2012; Wrzesniewski & Dutton, 2001). The use of AI has led employees to improve their fit, particularly in the context of evolving job designs (Bindl & Parker, 2010; Parker & Grote, 2020). Recent research proposes AI crafting, which refers to the process where AI itself autonomously modifies, personalises or tailors job roles and tasks based on the employee data, skills or organisation needs (Jarrahi et al., 2018). At the same time, AI-enabled job crafting refers to how AI acts as a tool to proactively craft employees' tasks, relationships or cognitive perceptions of work (Glikson & Woolley, 2020). Li et al. (2024) defines it as volitional acts to shape and redefine one's job to better adapt to the AI in workplace and effectively integrate AI into the work routines. Huang et al. (2019) also explain how AI can support employees in personalising their jobs by augmenting human capabilities. The current study focusses on AI-enabled job crafting, where job crafting is an employee driven behaviour but can be augmented through AI tools.

The evidence on job crafting in the context of AI gives insights into how AI accentuates crafting by employees and the possible outcomes of the same. AI can shape the way employees redefine and modify their jobs as it gives them the required tools to tailor their jobs to their strengths and interests (Ayinde & Kirkwood, 2020). AI helps them to automate repetitive tasks and focus on more meaningful work (Parker & Grote, 2020). For example, employees could use AI to automate data collection and analysis to focus on strategic thinking, problem solving or creative tasks which require utilization of unique skills (task crafting). By doing so, employees can align their daily responsibilities with their core strengths (Dengler & Matthes, 2018). Similarly, AI can provide employees with insights that reframe their understanding of the importance or impact of their jobs (cognitive crafting) and may contribute to their perception of their role as more purposeful or impactful (Li et al., 2024). It can also make it easier to make their work more presentable, facilitating better drafts and share of ideas (Dhoni, 2023). Tan (2024) show that AI awareness may make employees anxious, but the resultant anxiety may motivate them to craft their jobs, which in turn increases their competitive productivity. Further, AI can facilitate more collaboration and can make employees focus on high-value interpersonal activities, thus fostering better collaboration and knowledge sharing across organisations (relational crafting) (Glikson & Woolley, 2020).

Another study by He et al. (2023) demonstrate that leaders' acceptance of and affinity for AI encourages employees to pursue job crafting by making them feel ready for new changes related to AI adoption by the organisation. Cheng et al. (2023) explicate that employees may pursue promotion-oriented job crafting when they view organisational AI adoption as a challenge, as opposed to a hindrance in which case they may pursue prevention-oriented job crafting. An interesting take on job crafting is by Li et al. (2024) who introduced AI crafting as employees' proactive efforts to modify their jobs in response to AI adoption and demonstrated that leader's AI crafting behaviour encourages employees to pursue the same behaviour, which in turn improves their engagement and helping behaviour at their organisations. Thus, the implementation of AI-driven job crafting has potential for increased job performance and greater work engagement as employees feel greater alignment between their roles and personal goals (Tims et al., 2013). It can also foster a sense of ownership and higher engagement in discretionary behaviour such as organizational citizenship behaviour that benefits the organisation (Wrzesniewski & Dutton, 2001). Despite its benefits, AI-enabled job crafting can be difficult as employees might need sufficient training to effectively use AI tools (Jarrahi, 2018) and while AI can enhance job roles, it can also lead to job displacement or skill redundancy if not managed properly (Dengler & Matthes, 2018).

2.3 Proposed Model

Drawing from Social Cognitive Theory (SCT) (Bandura, 1997) and Self-Determination Theory (SDT) (Deci & Ryan, 1985), this research examines the impact of AI-enabled job crafting on workplace outcomes through the mediating influence of self-efficacy and work meaningfulness. For the workplace outcomes, we plan to include work role innovation and stress under individual performance and organisational citizenship behaviour and collaborative climate under team performance.

According to SCT, self-efficacy serves as a primary determinant of task motivated behaviour and task performance (Bandura, 1986; Saks, 1995). Self-efficacy is defined as "people's judgements of their capabilities to organize and execute courses of action required to attain designated types of performances" (Bandura, 1986, p. 391). It was further clarified that it refers to one’s belief in their capabilities to mobilize resources and take the required action to meet the demands (Wood & Bandura, 1989). Self-efficacy is related to self-regulation (Kanfer, 1991) as well as goal setting (Locke & Latham, 1990) and thus has significant importance as a basic element of attitudes and behaviours in the workspace. In the study, we posit that as employees craft their roles with the help of AI, their self-efficacy increases as they are able to modify their tasks based on personal strengths and thus facilitating learning opportunities. Employees with high-self efficacy are more likely to engage in proactive behaviour and have better performance (Zhang & Bartol, 2010). Hence, we propose that it will lead to increased individual performance and better engagement with the team.

According to SDT (Deci & Ryan, 2012), there are three universal psychological needs, specifically, need for competence, autonomy and relatedness that are essential for optimal development and functioning of an individual. These needs form the basis for intrinsic motivation and facilitate performance (Gagne & Deci, 2005). The theory suggests that these needs could be enhanced or undermined based on whether the social conditions thwarts or supports the psychological needs (Deci et al., 2017). In the context of job crafting using AI, it may help employees meet these needs and can be reflected in increased self-efficacy and improved meaningfulness of work which will further impact individual and team performance. However, undermining or enhancement of these needs depends on the social environment around the integration of AI in the organisation, such that if proper organisational support around AI including training, resources and encouragement for AI adoption is provided, it can improve the impact of AI-enabled job crafting on self-efficacy and work meaningfulness. When AI is not able to increase self-efficacy of employees, they may rely more on AI, but the competence would decrease (for example, when employees avoid learning new skills), which will adversely impact the individual and team performance. Similarly, when job crafting is not able to make the work more meaningful, it might not be able to satisfy the three needs and will limit the individual performance and their contribution to the teamwork. For example, it may reduce relatedness if AI handles more tasks than previously fostered by teamwork and social bonds. It, thus, becomes imperative to explore how AI-enabled job crafting can impact these internal needs and how the interplay of these needs can impact one's own and team's performance. The use of AI may lead to over-reliance or dependence on AI, which can reduce the overall efficacy of the individuals and in the long run make the work less meaningful. This may lead to decreased engagement, as employees may feel that they have less of a ''human'' role in their work (Garcia et al., 2023). It may also lead to skill atrophy in the long run if employees lean too much on AI for certain skill and neglect their development (Ganuthula, 2024). With the ease of working on tasks through AI, employees may over-craft or take on way more that they can handle or blur the boundaries between roles, which can lead to burnout or role ambiguity (Ganuthula, 2024; Garcia et al., 2023). It can also impact the contextual performance on the employee or employee's relation with other team members. For example, Bansal et al (2021) mention that AI has made the teams relations better as AI facilitates timely discussion, improves communication between the group members and as such employees are better able to reach team outcomes. However, team cohesion can suffer if individuals are more focused on their interactions with AI than on collaborating with each other (Zercher et al., 2023). It can lead to team members working in silos and reduce team's shared understanding and purpose (Zercher et al., 2023).

In our study, we thus propose that job crafting through AI would increase individual performance and develop a collaborative environment when it increases employee's self-efficacy and makes the work more meaningful, such that when job crafting develops greater confidence in employees' own abilities and make their work more valuable and significant, they are better able to contribute individually and within team. However, the impact on individual and team performance would also be influenced by how it is integrated in the organisational setting and how committed the individual is towards the organization. Individual performance could be observed through work role innovation or overall stress levels, whereas contextual performance could be seen with respect to team cohesiveness facilitating organizational citizenship behaviour (Morgan and Lassiter, 1992), and an empowered team environment (Jehn and Chatman, 2000).

3.0 Future Work

This is a research-in-progress paper to understand the impact of AI enabled job crafting on individual and team performance. We aim to collect data using a quantitative study. For the conference, we would present a comprehensive understanding of research in the area along with the conceptual model. We hope the feedback from academic community would help us refine the conceptual model before empirical testing.

References

Ardichvili, A., Dirani, K., Jabarkhail, S., El Mansour, W., & Aboulhosn, S. (2024). Using generative AI in human resource development: an applied research study. *Human Resource Development International, 27*(3), 388-409.

Ayinde, L., & Kirkwood, H. (2020). Rethinking the roles and skills of information professionals in the 4th Industrial Revolution. *Business Information Review*, *37*(4), 142-153.

Bandura, A. (1986). Social foundations of thought and action. *Englewood Cliffs, NJ*, *1986*(23-28), 2.

Batova, T. (2018). Work motivation in the rhetoric of component content management. *Journal of Business and Technical Communication*, 32(3), 308-346.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., ... & Weld, D. (2021, May). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-16).

Berg, J. M., Grant, A. M., & Johnson, V. (2010). When callings are calling: Crafting work and leisure in pursuit of unanswered occupational callings. *Organization Science*, 21(5), 973-994.

Bindl, U. K., & Parker, S. K. (2016). New perspectives and directions for understanding proactivity in organizations. *Proactivity at work*, 577-602.

Bruning, P. F., & Campion, M. A. (2018). A role–resource approach–avoidance model of job crafting: A multimethod integration and extension of job crafting theory. *Academy of Management Journal*, *61*(2), 499-522.

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., ... & Varma, A. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal, 33*(3), 606-659.

Cazzaniga, M., Jaumotte, M. F., Li, L., Melina, M. G., Panton, A. J., Pizzinelli, C., ... & Tavares, M. M. M. (2024). *Gen-AI: Artificial intelligence and the future of work*. International Monetary Fund.

Cheng, B., Lin, H., & Kong, Y. (2023). Challenge or hindrance? How and when organizational artificial intelligence adoption influences employee job crafting. *Journal of Business Research*, *164*, 113987.

Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Human Resource Management Review*, 33(1), 100899.

Deci, E. L., Olafsen, A. H., & Ryan, R. M. (2017). Self-determination theory in work organizations: The state of a science. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*(1), 19-43.

Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. *Handbook of theories of social psychology*, *1*(20), 416-436.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Dengler, K., & Matthes, B. (2018). The impacts of digital transformation on the labour market: Substitution potentials of occupations in Germany. *Technological Forecasting and Social Change*, *137*, 304-316.

Dhoni, P. (2023). Unleashing the potential: overcoming hurdles and embracing generative AI in IT workplaces: advantages, guidelines, and policies. Authorea Preprints.

Dutta, D., Mishra, S. K., & Tyagi, D. (2023). Augmented employee voice and employee engagement using artificial intelligence-enabled chatbots: a field study. The *International Journal of Human Resource Management, 34*(12), 2451-2480.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., & Wright, R. (2023). "So what if ChatGPT wrote it?" Multi-disciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71,* 102642.

Eapen, T., Finkenstadt, D. J., Folk, J., & Venkataswamy, L. (2023). How generative AI can augment human creativity. *Harvard Business Review, 101*(4), 56-64.

Gambacorta, L., Qiu, H., Shan, S., & Rees, D. M. (2024). Generative AI and labour productivity: a field experiment on coding (No. 1208). Bank for International Settlements (Working paper, 1208).

Gagné, M., & Deci, E. L. (2005). Self-determination theory and work motivation. *Journal of Organizational Behavior*, *26*(4), 331-362.

Ganuthula, V. R. R. (2024). The Paradox of Augmentation: A Theoretical Model of AI-Induced Skill Atrophy. *Available at SSRN 4974044.*

Garcia, R., Thompson, L., & Smith, A. (2023). Strategies for mitigating AI-induced skill atrophy in professional environments. *Harvard Business Review, 101*(4), 98-107

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627-660.

He, G., Liu, P., Zheng, X., Zheng, L., Hewlin, P. F., & Yuan, L. (2023). Being proactive in the age of AI: exploring the effectiveness of leaders' AI symbolization in stimulating employee job crafting. *Management Decision, 61*(10), 2896-2919.

Huang, M. H., Rust, R. T., & Maksimovic, V. (2019). The feeling economy: Managing in the next generation of AI. *California Management Review, 61*(4), 43-65.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577-586.

Jehn, K. A., & Chatman, J. A. (2000). The influence of proportional and perceptual conflict composition on team performance. *International Journal of Conflict Management*, *11*(1), 56-73.

Kanfer, R. (1991). Motivation theory and industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (pp. 75-170). Palo Alto, CA: Consulting Psychologists Press.

Ooi, K. B., Tan, G. W. H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., ... & Wong, L. W. (2023). The potential of generative artificial intelligence across disciplines: Perspectives and future directions. *Journal of Computer Information Systems*, 1-32.

Lazazzara, A., Tims, M., & De Gennaro, D. (2020). The process of reinventing a job: A meta–synthesis of qualitative job crafting research. *Journal of Vocational Behavior,* 116, 103267.

Li, W., Qin, X., Yam, K. C., Deng, H., Chen, C., Dong, X., ... & Tang, W. (2024). Embracing artificial intelligence (AI) with job crafting: Exploring trickle-down effect and employees' outcomes. *Tourism Management*, *104*, 104935.

Locke, E. A., & Latham, G. P. (1990). Work motivation and satisfaction: Light at the end of the tunnel. *Psychological Science*, *1*(4), 240-246.

Morgan, B. B., Jr., & Lassiter, D. L. (1992). Team composition and staffing. In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 75–100). Ablex Publishing.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187-192.

Parker, S. K., & Collins, C. G. (2010). Taking stock: Integrating and differentiating multiple proactive behaviors. *Journal of Management*, *36*(3), 633-662.

Parker, S. K., & Grote, G. (2022). More than 'more than ever': Revisiting a work design and sociotechnical perspective on digital technologies. *Applied psychology*, *71*(4), 1215-1223.

Parker, S. K., & Grote, G. (2022). Automation, algorithms, and beyond: Why work design matters more than ever in a digital world. *Applied psychology*, *71*(4), 1171-1204.

Pournader, Mehrdokht, Hadi Ghaderi, Amir Hassanzadegan, and Behnam Fahimnia. (2021). Artificial Intelligence Applications in Supply Chain Management. *International Journal of Production Economics, 241*, 108250.

Ross, S. I., Martinez, F., Houde, S., Muller, M., & Weisz, J. D. (2023). The programmer's assistant: Conversational interaction with a large language model for software development. In Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23) (pp. 491–514). Association for Computing Machinery, New York, NY, USA.

Saks, A. M. (1995). Longitudinal field investigation of the moderating and mediating effects of self-efficacy on the relationship between training and newcomer adjustment. *Journal of Applied Psychology*, *80*(2), 211-225.

Sharma, Rohit, Anjali Shishodia, Angappa Gunasekaran, Hokey Min, and Ziaul Haque Munim. (2022). The Role of Artificial Intelligence in Supply Chain Management: Mapping the Territory. *International Journal of Production Research, 60*(24), 7527–7550.

Sooraksa, N. 2021. "A Survey of Using Computational Intelligence (CI) and Artificial Intelligence (AI) in Human Resource (HR) Analytics." 7th International Conference on Engineering, *Applied Sciences and Technology (ICEAST*), April.

Tan, T. K. (2024). Artificial intelligence and basic human needs: the shadow aspects of emerging technology. In *Ethics in Online AI-based Systems* (pp. 259-278). Academic Press.

Tims, M., Bakker, A. B., & Derks, D. (2012). Development and validation of the job crafting scale. *Journal of Vocational Behavior*, *80*(1), 173-186.

Tims, M., Bakker, A. B., & Derks, D. (2013). The impact of job crafting on job demands, job resources, and wellbeing. *Journal of Occupational Health Psychology, 18*, 234–245.

Tims, M., Twemlow, M., & Fong, C. Y. M. (2022). A state-of-the-art overview of job-crafting research: current trends and future research directions. *Career Development International*, 27(1), 54-78.

Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal, 42(*9), 1600-1631.

Wood, R., & Bandura, A. (1989). Social cognitive theory of organizational management. *Academy of management Review*, *14*(3), 361-384.

Wrzesniewski, A., & Dutton, J. E. (2001). Crafting a job: Revisioning employees as active crafters of their work. *Academy of management review*, *26*(2), 179-201.

Yorks, A., and Rotaroti. 2021. "Digitization, Artificial Intelligence, and HRD." In Strategic Human Resource Development in Practice: Leveraging Talent for Sustained Performance in the Digital Age of AI. Springer.

Zercher, D., Jussupow, E., & Heinzl, A. (2023). When AI joins the team: a literature review on intragroup processes and their effect on team performance in team-AI collaboration. *ECIS 2023 Research Papers, 307.*

Zhang, F., & Parker, S. K. (2019). Reorienting job crafting research: A hierarchical structure of job crafting concepts and integrative review. *Journal of organizational behavior*, *40*(2), 126-146.

Exploring the Impact of AI Adoption in Radiology: A Qualitative Study in Saudi Arabia

Abdullatif Alshamrani (University College Cork), Dr Stephen Treacy (University College Cork), Dr Wendy Rowan (University College Cork) and Dr Brian O'Flaherty (University College Cork)

Research In progress

Abstract

Artificial intelligence (AI) intervention in the clinical environment has possessed the capacity to revolutionise radiology by enhancing the accuracy and efficiency of diagnostics. However, the success of implementing such AI-based diagnostic tools relies significantly on understanding the influence of AI on clinical decision-making and the risks involved, which is pivotal in delivering better healthcare quality by supporting clinical decision-making, assessing risk, and enhancing clinical workflows. This study explores the impact of AI on clinical decision-making and evaluates risks in radiology as Saudi Arabia's healthcare system transforms under Vision 2030. The investigation utilises semi-structured interviews to provide in-depth insights and employs thematic analysis to recognise emerging trends. Preliminary results from the literature focus on the challenges and enablers aspects of AI, such as education, training, workflow, and cultural readiness, which have a significant role in the medical AI context. As a work-in-progress, there is a demand to explore the impact of AI on clinical decision-making, particularly semi-structured decisions. This investigation aims to contribute theoretically and practically to the domain of digital health innovation by addressing gaps in comprehending the viewpoints of radiology healthcare professionals. The developmental paper exhibits the research framework, methodological technique, and foreshadowed contributions, seeking constructive criticism to refine the research and augment its significance.

Keywords: Artificial Intelligence, Radiology, Healthcare professionals, Decision making, Risk.

1. Introduction

The growth in capabilities of Artificial Intelligence (AI) applications in recent years has transformed the future visions in many industries, especially healthcare since McCarthy first introduced it in 1956 (McCarthy et al., 2006). The adoption of AI in the healthcare sector has substantial potential in terms of improving diagnostic precision, simplifying workflows, and enhancing patient outcomes (Aldwean & Tenney, 2024). However, the adoption of AI tools in healthcare poses obstacles, such as data privacy concerns, the necessity for considerable data training, and unpredicted algorithm bias (Solanki et al., 2023). Therefore, comprehending the impact of artificial intelligence applications on healthcare professionals' decision-making is essential. Thus, explore the impact of AI on professional clinicians' decision-making, which plays a significant role in the success of AI as a new technology. In Saudi Arabia, digital transformation initiatives in the healthcare sector are experiencing significant investment by US$1.7 billion, including AI technology (PIF,2024). Although these advantages are present, there is a lack of qualitative research concentrating on AI's influence on clinical decisions and risks associated with radiology. There is a necessity to address this gap to provide AI solutions that are tailored and aligned with the healthcare professionals' needs and address workers' concerns in the Kingdom of Saudi Arabia as there is a significant investment to transform the health industry by adopting new technologies as planned in the 2023 vision, , which may result in several risks if there is a lack of studies on the implications of adopting artificial intelligence for clinical decisions and its associated risks. This investigation aims to assess the impact of Artificial Intelligence (AI) on healthcare professionals' clinical decision-making and risk assessment of the adoption of AI technology. It utilises a qualitative approach to assess AI's potential advantages and limitations that impact AI clinical decision-making in medical practice. The focus on understanding the implications of AI involvement in medical decision-making and establishing risk assessment will be a novel perspective in the literature, offering in-depth insight into the impact of Artificial Intelligence (AI) on practitioners' clinical decision-making and exhibiting key factors for assessing related AI risk to support healthcare professionals throughout the AI transition in practice. This has resulted in the formulation of the subsequent research questions:

•RQ1: What enablers and barriers do healthcare professionals identify when using AI to enhance clinical decision-making?

•RQ2: How does AI impact healthcare professionals' decision-making processes in semi-structured scenarios?

•RQ3: How do practitioners assess risk associated with AI in semi structured decision-making in radiology?

2. Literature review

AI has become a transformative strength in healthcare, bringing promising advancements in early diagnosis, personalised treatment, and patient management capabilities. Several studies have shown AI's potential capabilities in optimising diagnostic accuracy, facilitating clinical workflows, and enhancing patient outcomes (Petersson et al., 2022). Nevertheless, the successful integration of AI technology in healthcare systems is dependent on diverse factors such as technological, ethical, and sociocultural obstacles (Callon 1881; Penston, 2007), Technological obstacles majorly hinder the adoption of AI applications in healthcare. A key concern is the compatibility of AI applications with the existing electronic health records (EHRs) in the integration phase. Numerous healthcare organisations employ outdated systems inconsistent with AI tools, leading to an interruption of information flow. The absence of interoperability is crucial as it can affect the implementation of AI solutions and raise fear of data consistency and accuracy (Petersson et al., 2022). In addition, healthcare database lacks interoperability as a result of isolated databases and software ownership raises technical risks (Francisca et al., 2024). Moreover, AI technology in healthcare has ethical issues such as bias, privacy, security, transparency and accountability which must be addressed to provide reliable and fair implementation. For instance, bias is a primary concern caused by insufficient training for AI systems or even training on non-representative datasets, leading to false medical decisions (Choukhi, 2024; Ueda et al., 2024). Furthermore, the utilisation of AI in healthcare can be influenced by sociocultural factors that impact the perceptions and acceptance of patients and healthcare providers. Another substantial challenge is "algorithm aversion," where users show hesitation to trust AI-driven solutions, often due to algorithm error and the lack of human empathy in machine-based decision-making (Filiz et al., 2023).

Current research has mainly concentrated on AI capabilities and applications in healthcare radiology. While some studies have investigated healthcare professionals' attitudes towards AI adoption (Laï et al., 2020), there is a lack of exploring the impact of AI on semi structured decision-making and the risks involved. This research intends to utilise the actor network (ANT)**,** which is creative as it integrates the socio-technical, that can play a role in analysing dynamic relations between users and technologies (Latour, 1996). Hence, ANT acknowledges humans and non-humans as equal actors in a network, building on assumptions that all entities are connected (Ryan et al., 2024).

This approach is unlike traditional models such as TAM or UTAUT that focus primarily on static behavioural predictors like perceived usefulness (Davis, 1989; Venkatesh et al., 2003). ANT is suitable for complicated cultural, organisational, and personal factors (Dankert, 2012). In addition, exploring the impact of AI on clinical decision-making through a qualitative approach can distinguish it from existing observational studies (Tikhomirov et al., 2024), and provide in-depth insight. The study aims to understand how AI impacts semi-structured decision-making and risk assessment by healthcare professionals in the radiology field. By concentrating on how AI affects medical professionals' decision-making and risk assessment specifically AI adoption AI for image recognition, decision support, and workflow automation to offer insights. AI tools are now integrated into several domains, such as X-ray, CT, and MRI. Therefore, examining the AI technology in these modelists provides significant value for both research and practice fields.

2. Research Design/Methodology

This investigation employs a qualitative research approach to analyse the impact of AI on semi structured decision-making and risk assessment. This investigation employs a qualitative approaches are significant for complex social phenomena that require a deep understanding as they are appropriate for the in-depth analysis of personal experiences, beliefs, and perspectives (Creswell, 2014, p. 4). In the healthcare field, a qualitative study can facilitate the investigation of the effects of AI and risks encountered by medical personnel in the AI era by providing in-depth insight into the current status.

2.1 Data Collection Methods

This study intends to employ semi-structured interviews as the principal data collection technique. This approach facilitates adaptability in the investigation perspectives of the interviewees while preserving a uniform structure throughout interviews (Tanwir et al., 2021). The participants will include 10 radiologists, 10 technicians, and 10 nurses from 2 hospitals in Saudi Arabia. The interviews will be performed either online via the University Approved MS Teams platform and in person at healthcare site, contingent upon participants' preferences and cultural factors. Every interview will be recorded in audio format (both online and in-person) and then transcribed and analysed, adopting a thematic approach using NVivo software to recognise patterns for coding, categorising and specifying themes related to obtaining a comprehensive understanding of AI implications.

2.2 Anticipated Challenges

Several difficulties may be foreseen in this study:

- Data access: Gaining access to healthcare professionals for interviews may prove difficult due to clinical practitioner packed schedules and administrative obstacles. To prevent this risk, this study will solicit assistance and support from key gatekeepers such as hospital management and professional associations to enhance recruitment efforts.

- Cultural Subtleties: Cultural distinctions in Saudi Arabia might impact the participants' readiness to engage in open discussions on particular subjects due to religious beliefs, organisational cultural and language. Therefore, the study will follow culturally adapted interview such as online interview to address religious views and institutional cultural and, perform interviews in the participant's language of choice and, leverage AI translation platform in the transcription of interviews.
- Ethical concerns: This study, it will comply with ethical standards by obtaining institutional review board approval and implementing a protocol to protect participants' rights in concealing their identities and data. In addition, the researchers will seek to obtain ethical approval to conduct this research and will ensure confidentiality and obtain informed consent prior to any data collection taking place. Through qualitative methods and addressing obstacles, the study seeks to provide an in-depth understanding of significant impact of AI on semi structured decision-making evaluate risk in radiology.

3. Theoretical Contributions

In a daily medical environment, human agencies evaluate clinical decision-making models based on their cognitive skills. This study uses Actor-Network Theory (ANT) to explore AI's influence on clinical decision-making and assessing AI risk in the healthcare domain, concentrating on the relationship between human and non-human objects that exist, named actants, that share responsibility in network assembly (Dankert, 2012). However, with the advancement in AI technology, ANT implies that AI tools have a considerable effect on clinical decision-making, not simply diagnosing disease. AI systems generate medical predictions, displaying abnormalities in radiological images and suggesting therapy plans (Ghaffar Nia et al., 2023). Therefore ANT illustrates how early relationships lead to establishing new entities that might not follow the same attributes of the foundational entities (Dankert, 2012). ANT emphasises how AI tools act as intermediaries in socio-technical networks, forming and being formed by interactions with healthcare professionals and other entities such as hospital regulations, and medical data. ANT can offer a valuable foundation for comprehending the implications of AI on semi-structured decision-making and associated risks in the healthcare context. ANT acknowledges AI as a co-actor rather than treating AI technology as an inactive tool in other traditional models. It underlines how clinical decision-making is influenced by established networks of non-human and human actors, defines the direction in which AI reassigns authority, identifies potential risks and transforms professional responsibilities. In addition, ANT's networks underscore those risks caused by AI errors is due to the inconsistency between actants (Callon 1881; Callon, 1984). Through this investigation of the impact of AI on semi structured decision-making and risk assessment. This study critically analyses AI implications on semi-structured decision-making and risk assessment that influence AI adoption in radiology. By incorporating ANT, this exploration offers empirical insights about the role of AI as a transforming actor in healthcare, providing a valuable lens to analyse how AI impacts healthcare professionals' decision-making and address the risks of AI adoption, emphasising their ability to mitigate risk, enhance workflows, develop trust in AI systems, and uphold responsibility in healthcare decision-making. These insights enhance theoretical frameworks, providing a more profound comprehension of the socio-technical dynamics that enable AI adoption in healthcare settings.

4. Practical Contributions

This investigation aims to offer practical recommendations to inform healthcare policy and practice. By identifying AI implications on semi-structured decision-making, such as responsibilities concerns, infrastructural gaps, and risk issues, the study provides a valuable understanding of AI's influence on clinical decision-making and the risks involved to effectively utilise AI within healthcare systems. In addition, the research highlights the significance of the sustainability of AI evaluation by addressing any challenges that healthcare professionals encounter during this digital transformation. These valuable insights will assist healthcare institutions and policymakers in developing practical solutions to optimise the advantages of AI while reducing opposition. Eventually, the findings seek to establish fair and efficient policy guidelines for ethical AI-driven healthcare systems in Saudi Arabia, as well as guidance for various international healthcare domains.

4. Conclusions

This study is a work in progress and seeks to highlight the significance of understanding of AI's impact on semi-structured decision-making in radiology. It explores AI's implications on medical decision-making and risk assessment to provide valuable insights into the intricate dynamics of AI integration within healthcare. Employing the ANT will contribute to knowledge of socio-technical relationships and potentially provide pragmatic solutions for successful AI implementation in clinical environments. The following action will include gathering data through semi-structured interviews with healthcare professionals. subsequently, a thematic analysis will be executed to identify critical patterns and themes. The results will be contextualised within the theoretical frameworks to refine insights and provide recommendations based on evidence. This study aims to bridge the gap in the field by providing relevant academic knowledge to the domain and supporting the development of clinical decision-making, particularly in semi-structured decision-making scenarios, addressing risks and safety issues, and ensuring sustainable AI solutions in the medical radiology field.

References

Aldwean, A., & Tenney, D. (2024). Artificial Intelligence in Healthcare Sector: A Literature Review of the Adoption Challenges. *Open Journal of Business and Management*, *12*(01), 129–147. https://doi.org/10.4236/ojbm.2024.121009

Callon, M. (1984). Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. *The Sociological Review*, *32*(1_suppl), 196–233. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x

Choukhi, A. (2024). Artificial Intelligence In Health From Challenges To Impacts Through The Case Study Of The Valenciennes Hospital Center.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed). SAGE Publications.

Dankert, R. (2012). Actor–Network Theory. In *International Encyclopedia of Housing and Home* (pp. 46–50). Elsevier. https://doi.org/10.1016/B978-0-08-047163-1.00606-8

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319. https://doi.org/10.2307/249008

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2023). The extent of algorithm aversion in decision-making situations with varying gravity. *PLOS ONE*, *18*(2), e0278751. https://doi.org/10.1371/journal.pone.0278751

Ghaffar Nia, N., Kaplanoglu, E., & Nasab, A. (2023). Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence*, *3*(1), 5. https://doi.org/10.1007/s44163-023-00049-5

Latour, B. (1996). *On actor-network theory: A few clarifications.*

Penston, J. (2007). Patients' preferences shed light on the murky world of guideline-based medicine. *Journal of Evaluation in Clinical Practice*, *13*(1), 154–159. https://doi.org/10.1111/j.1365-2753.2006.00701.x

Petersson, L., Larsson, I., Nygren, J. M., Nilsen, P., Neher, M., Reed, J. E., Tyskbo, D., & Svedberg, P. (2022). Challenges to implementing artificial intelligence in healthcare: A qualitative interview study with healthcare leaders in Sweden. *BMC Health Services Research*, *22*(1), 850. https://doi.org/10.1186/s12913-022-08215-8

Ryan, T., Hynes, B., Ryan, N., & Finucane, A. (2024). Investigating the use of actor-network theory in healthcare: A protocol for a systematic review. *BMJ Open*, *14*(5), e079951. https://doi.org/10.1136/bmjopen-2023-079951

Solanki, P., Grundy, J., & Hussain, W. (2023). Operationalising ethics in artificial intelligence for healthcare: A framework for AI developers. *AI and Ethics*, *3*(1), 223–240. https://doi.org/10.1007/s43681-022-00195-z

Tanwir, F., Moideen, S., & Habib, R. (2021). Interviews in Healthcare: A Phenomenological Approach A Qualitative Research Methodology. *Journal of Public Health International*, *4*(2), 10–15. https://doi.org/10.14302/issn.2641-4538.jphi-21-3881

Tikhomirov, L., Semmler, C., McCradden, M., Searston, R., Ghassemi, M., & Oakden-Rayner, L. (2024). Medical artificial intelligence for clinicians: The lost cognitive perspective. *The Lancet Digital Health*, *6*(8), e589–e594. https://doi.org/10.1016/S2589-7500(24)00095-5

Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S.

(2024). Fairness of artificial intelligence in healthcare: Review and recommendations. *Japanese Journal of Radiology*, *42*(1), 3–15. https://doi.org/10.1007/s11604-023-01474-3

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Report on Modifications:

First Editor's Comment	**Issue Identified**	Response/Revision
"While the study aims to examine 'the barriers and facilitators that influence health workers' experiences,' it is unclear what is meant by this. Are they looking at aspects of the implementation that influence health workers' perspectives of the system? Are they looking at aspects of the implementation that influence the implementation itself, including health workers' perspectives of the system?"	Lack of clarity in research focus.	The research focus on. The impact of AI on healthcare professionals semi-structured decision making. Healthcare hospitals now are implementing AI in clinical practice practically in radiology imaging their work which reshaping their decision making.
"It is unclear how Actor Network Theory (ANT) and the Digital Agency model will be combined fruitfully. They seem in tension with each other, as ANT offers agency to both the technologies	Theoretical inconsistency between ANT and Digital Agency	This study will focus on ANT to map socio-technical interactions identifying both human and non-human actors and it's been used in healthcare and

involved in socio-technical systems and the agents that use them, while digital agency focuses only on the human agents."		technology context.
"The explanation of their use together is unsatisfactory. *However, the use of these two theoretical approaches in tandem is novel and could add interesting insight if the authors were able to develop this further."*	Combining ANT and Digital Agency	This study will focus on ANT to map socio-technical interactions identifying both human and non-human actors and it's been used in healthcare and technology context.
Second Editor's Comment	**Issue Identified**	Response/Revision
"AI covers a wide range of technologies and methods, so it's important to specify which AI systems are being discussed."	Insufficient clarity regarding the **particular AI tools** under investigation.	AI for image recognition, decision support, and workflow automation
"Giving more details about the specific AI tools used in the study would help better understand the health	Insufficient detail on **how AI impacts health professionals' experiences**.	Professionals' confidence and assessing risk associated with AI implementation.

workers' experiences and make the findings more relevant to the wider conversation about AI in healthcare."		
"A key issue is that AI covers a wide range of technologies and methods, so it's important to specify which AI systems are being discussed."	**Broad discussion** on AI, complicating the contextualization of findings.	X-ray, CT, and MRI.
"As a research-in-progress paper, it has the potential to offer valuable insights into the use of AI in healthcare. With more development, especially in areas like participant details, AI clarification, and methodology, this paper could make an important contribution to the field."	Aspects need enhancement: **participant information, specificity of AI, and methodological approach.**	This issue is solved by Adding more details.

Bridging the AI Divide: Barriers and Challenges to AI Adoption for Nigerian SMEs

Aishatu Mohammed Lawan
University of Salford
Salford Business School
m.albakri@salford.ac.uk

Dr. Maria Kutar
University of Salford
Salford Business School
m.kutar@salford.ac.uk

Dr. Mohammed Albakri
University of Salford
Salford Business School
m.albakri@salford.ac.uk

Abstract

The increasing relevance of Artificial Intelligence (AI) in the Fourth Industrial Revolution has highlighted significant disparities in AI adoption among Small and Medium Enterprises (SMEs), particularly in developing regions such as Nigeria. This research builds on a Systematic Literature Review (SLR) that identified the socio-technical factors contributing to the AI divide, using empirical data from 144 Nigerian SMEs to deepen understanding of these challenges. Findings reveal that technical challenges, such as inadequate infrastructure and outdated technology, intersect with social issues, including resistance to change and low digital literacy, exacerbating the divide. Socio-technical barriers, such as skills shortages, ethical concerns, and regulatory gaps, hinder AI integration and equitable outcomes. The study highlights the need for targeted interventions, including policy support, infrastructure development, and capacity-building programs, to enable SMEs to harness AI's transformative potential. It advocates for broader research to contextualise these findings and develop actionable strategies for addressing the AI divide in similar economies.

Keywords: AI Divide, Artificial Intelligence, Digital Divide, Socio-Technical, Challenges

1.0 Introduction

Artificial Intelligence (AI) is a critical enabler of the Fourth Industrial Revolution, with transformative potential across various sectors. Organisations worldwide are adopting AI to solve various managerial challenges and drive innovation (X. Yu et al., 2023) Although there is no single universally accepted definition of AI, it is universally understood as the capability of machines to perform tasks that typically require human intelligence, such as learning, reasoning, and decision-making (Dwivedi et al., 2021). In the context of this study, AI refers specifically to the use of machine learning, natural language processing, and robotics to enhance decision-making, automate processes, and improve business operations in SMEs. While automation is a component of AI, AI extends beyond repetitive task automation by integrating cognitive, relational, and structural functions (Syam & Sharma, 2018). Unlike previous industrial revolutions driven by mechanization and steam power, AI reshapes industries by enabling data-driven decision-making, predictive analytics, and adaptive learning capabilities (Wright & Schultz, 2018).

At the organisational level, AI is applied to streamline operations and improve outcomes. For example, customer service chatbots handle routine inquiries, freeing employees to focus on more complex, value-added tasks (Boustani, 2022). Scholars argue that when deployed effectively, AI can contribute to socioeconomic and environmental development, enhancing overall quality of life (Boustani, 2022). AI is increasingly recognised as a critical driver of future work (Huang & Rust, 2018; Wilson et al., 2019). However, its adoption presents challenges, particularly for small and medium-sized enterprises (SMEs) that need more resources and expertise to fully leverage AI technologies (Szedlak et al., 2020).

The emerging concept of the "AI divide" highlights the growing disparities in access to AI technologies, mirroring the earlier digital divide (Yu, 2020). The AI divide reflects unequal access to AI tools, resulting in inequalities between those who can harness its potential and those who cannot. While some countries and organisations are advancing rapidly in AI adoption, others, particularly in developing regions such as Nigeria, need more access to infrastructure, skills, and resources. The United Nations' Global Digital Compact, presented at the UN General Assembly in 2024, emphasises the need for international cooperation to ensure that AI's benefits are equitably distributed without exacerbating existing inequalities (United Nations, 2024). The Global Digital Compact outlines principles to foster an open, fair, and inclusive digital environment, enabling micro, small, and medium-sized enterprises (MSMEs) to access and compete in the digital economy, aligning with Sustainable Development Goal 9 (SDG 9).

In Nigeria, AI presents numerous opportunities across sectors such as education (Abayomi et al., 2021), security (Falode et al., 2021), energy (Mobayo, 2021) and healthcare (Anazodo et al., 2022; Muhammad & Algehyne, 2021). However, several challenges hinder AI adoption in the country, including limited awareness, unreliable power supply, and a lack of trust in AI technologies (Mobayo, 2021). These challenges are particularly acute for SMEs in Nigeria, where infrastructural deficits and socio-political instability further complicate efforts to harness AI's potential.

SMEs play a crucial role in economic development, particularly in developing nations like Nigeria. They are central to job creation, poverty alleviation, and regional development (Etuk et al., 2014; Oregwu & Chima, 2013). SMEs account for 96% of businesses and 84% of employment in Nigeria, demonstrating their significant economic contribution (Ogbe et al., 2020). These obstacles hinder their ability to innovate and leverage AI technologies. However, Research indicates that most SMEs have lower levels of digitalisation than larger enterprises, with many still in the early stages of digital transformation (Szedlak et al., 2020). Barriers such as high costs, lack of knowledge, and concerns over data security hinder AI adoption (Oldemeyer et al., 2024).

Integrating AI technologies into SMEs holds immense potential for transforming operations and addressing managerial challenges (Yu et al., 2023). However, the AI divide prevents many SMEs from accessing these benefits, particularly in developing regions like Nigeria. The UN's Global Digital Compact emphasises that bridging this divide is essential to ensure that AI contributes to global sustainable development rather than exacerbating existing inequalities (United Nations, 2024).

In our previous study, Mohammed et al. (2023) introduced a socio-technical framework that identifies access, capability, and outcomes as key factors contributing to the AI divide. The study emphasized the need for further empirical investigation to understand how these factors influence AI adoption in developing economies. While other scholars, such as Carter et al. (2020) and Wu (2020), have explored the AI divide, their studies primarily focus on developed economies, making their findings less applicable to contexts like Nigeria. Building on our earlier work, this study expands the framework by examining the specific barriers Nigerian SMEs face in adopting AI. To address this gap, this study asks: How do socio-technical barriers contribute to the AI divide among SMEs in Nigeria, and what are their implications for AI adoption? Through survey data from 144 SMEs and thematic analysis, we provide new insights into the socio-technical challenges shaping AI adoption, offering a more contextualized perspective on the AI divide in developing economies.

1.1 The Concept of the AI Divide

The Artificial Intelligence (AI) divide highlights the growing disparities in access to, understanding of, and effective use of AI technologies across regions and populations. Since John McCarthy introduced the term in 1956, AI has advanced significantly, encompassing various disciplines and applications designed to replicate human cognition through complex algorithms (Bjola, 2022). As AI becomes integral to healthcare, education, and business, disparities in access and ability to leverage these technologies are becoming more evident (Mohammed et al., 2023). Building on the concept of the digital divide, which highlights disparities in digital technology access, the AI divide focuses on inequalities in AI adoption, skills, and benefits (Wu, 2022). It captures the challenges faced by those with limited AI resources, restricting their ability to harness its potential.

The AI divide extends beyond the mere availability of AI technologies, including the capacity to understand and utilise them effectively and the benefits derived from their application (Mohammed et al., 2023). Yu (2020) notes that the digital divide has historically marginalised certain groups, and the risk of exclusion has deepened as AI evolves. Without intervention, the AI divide may worsen inequalities in education, employment, and economic opportunities (Wu, 2020). Addressing this divide requires inclusive strategies to ensure AI technologies promote equity rather than perpetuate disparities.

1.2 Levels of the AI Divide

The AI divide can be understood through three key levels: access, capability, and outcomes. These levels highlight different dimensions of disparities in AI adoption and provide a framework for designing targeted strategies to reduce these inequalities.

1.2.1 Access to AI Technologies

The first level of the AI divide concerns access to AI technologies and the necessary infrastructure to support their use, such as reliable internet connectivity. Socioeconomic factors heavily influence access, with rural and low-income populations often unable to afford or utilise AI technologies (Sakapaji & Puthenkalam, 2023). Challenges like poor internet service, limited access to advanced hardware, and the high costs of AI tools exacerbate these disparities.

Addressing these barriers is essential for equitable AI adoption. Improving digital infrastructure, lowering the cost of AI tools, and implementing inclusive policies can significantly reduce access inequalities. With access issues resolved, efforts to bridge the AI divide will be completed.

1.2.2 Capability to Utilise AI Technologies

The second level focuses on the capability to use AI effectively. Access alone is insufficient if individuals or organisations lack the skills to integrate AI into their practices. Wu (2022) highlights capability as a critical element of the AI divide, alongside access and outcomes.

Educational background and socioeconomic status often determine one's ability to understand and apply AI (Ball & Huang, 2023). Those with higher education levels are better positioned to leverage AI tools, gaining advantages in employment and innovation. Conversely, limited education or technical skills hinder effective use, deepening inequalities. Addressing this gap requires targeted investments in education and training to equip diverse populations with essential AI skills, fostering inclusivity in AI adoption.

1.2.3 Outcomes from AI Utilisation

The third level addresses disparities in the benefits derived from AI technologies. Even with access and capability, unequal outcomes persist. Wu (2022) notes that while some organisations gain significant advantages like improved productivity, others fail to benefit, widening the gap.

This inequality can deepen economic and social divides. Dwivedi et al., (2021) highlights the economic growth experienced by countries and organisations effectively adopting AI, contrasting it with stagnation in those lagging. Ensuring equitable outcomes requires improving access and capability and enacting policies that promote inclusive growth and fair distribution of AI benefits. These efforts are critical to mitigating disparities and fostering balanced opportunities for all.

1.3 AI Transforming the SME Landscape

Small and Medium Enterprises (SMEs) have historically been cautious about adopting digital technologies due to concerns over cost, complexity, and relevance to their operations (Bradač & Huđek 2023). However, the rapid advancement of Artificial Intelligence (AI) is prompting even the most hesitant SMEs to rethink their approach. Once reserved for large corporations, AI has become an essential tool for businesses of all sizes. A 2023 survey by SBE (2023) revealed that 75% of small businesses are using AI tools across various functions, underscoring its growing role in SME operations. This adoption is driven by AI's ability to enhance efficiency, improve decision-making, and provide a competitive edge.

Globally, AI is recognised as a transformative technology for SMEs, particularly in optimising production processes and boosting export performance and global competitiveness (Denicolai et al., 2021; OECD, 2023). Smaller enterprises are increasingly adopting AI solutions tailored to their unique needs, integrating them into production and strategic business operations.

Generative AI, in particular, experienced explosive growth in 2023, with profound implications across industries (Jafari-Sadeghi et al., 2022). Alongside this trend are the democratisation of AI, evolving regulations, and increased collaboration between humans and AI (Chui et al., 2023). These shifts highlight AI's transformative potential, reshaping SME strategies and empowering businesses of various scales. Table 1 below outlines the various business applications of AI in the context of SMEs illustrating the breadth of functions for which solutions are available.

SME functions	Business applications of AI
Marketing and Sales	The integration of generative AI automates repetitive tasks, liberating employees to focus on more creative and higher-level responsibilities (McKinsey Global Report, 2023). This not only

	streamlines operations but also significantly boosts productivity (Neeli, 2020).
Supply Chain Management	It addresses a spectrum of challenges, ranging from demand forecasting and supplier selection to inventory management, production planning, scheduling, quality control, and logistics and transportation (Toorajipour et al., 2021) . The application of AI in SCM extends to managing demand variability, selecting suppliers, and determining facility locations, collectively optimising the entire supply chain.
Customer Service	AI technology is adept at creating personalised messages tailored to individual customer interests, preferences, and behaviours. It also performs tasks such as generating first drafts of brand advertising, headlines, slogans, social media posts, and product descriptions (Jahanbakhsh Javid & Amini, 2023). Moreover, AI provides faster and more accurate responses to customer queries and requests, leveraging techniques like machine learning, deep learning, and expert systems (Borah et al., 2022).
Decision Making	Empowers businesses to make informed decisions, particularly in areas like customer data analysis for product development (Mannuru et al., 2023).

Table 1. Business Applications of AI in SMEs

1.4 Impact and Benefits of AI on Small Businesses

Integrating Artificial Intelligence (AI) into small and medium-sized enterprises (SMEs) holds transformative potential, reshaping the competitive landscape and highlighting the risk of diminished competitiveness, reduced market share, and weakened economic influence for SMEs that fail to adopt it (Baabdullah et al., 2021). AI's role has evolved from a mere technological upgrade to a strategic necessity for SMEs' survival and growth. Despite challenges like privacy concerns and skill requirements (Schönberger, 2023), AI is increasingly vital for SMEs to thrive in a competitive, digitalised business landscape (Schönberger, 2023).

1.5 Challenges and Barriers to AI Adoption in SMEs

AI adoption is critical for SMEs to stay competitive in the digital economy, but significant barriers such as financial constraints, technical complexity, limited digital maturity, and skill gaps impede progress. High costs and complex AI systems are incredibly challenging for SMEs with limited budgets and technical capabilities (Hansen& Bøgh,2021; Schönberger, 2023). Despite awareness of AI's potential, only a fraction of SMEs have adopted these technologies due to low digital maturity, lack of tailored solutions, and insufficient data quality (Bettoni et al., 2021; Szedlak et al., 2020; Ulrich et al., 2021). Furthermore, organisational culture and public trust play crucial roles; SMEs often resist change and fear losing control to automation, especially where cultural resistance and data privacy concerns are high (Schoeman, 2024). Security and privacy issues and weak regulatory frameworks create further obstacles, especially in regions like sub-Saharan Africa, where digital literacy is lower (Borah et al., 2022; Madden & Kanos, (2020). These factors contribute to a growing digital divide between larger firms and SMEs, necessitating concerted efforts to improve infrastructure, digital literacy, and regulatory support. Further details on these barriers are outlined in the table below.

Barrier	Description	References
High Costs and Financial Constraints	AI technologies can be prohibitively expensive for SMEs, which often operate with limited budgets and access to capital.	(Radziwon et al., 2022; Ulrich et al., 2021)
Complexity of AI Implementation	The technical complexity of AI systems can be overwhelming for SMEs, which may lack the necessary expertise to implement AI.	(Bettoni et al., 2021; Hansen & Bøgh, 2021)
Digital Maturity	Many SMEs are in the early stages of digitalisation, lacking the foundational infrastructure needed for AI adoption.	(Mittal et al., 2018; Szedlak et al., 2020)
Lack of Digital Skills	A shortage of digital literacy and AI-specific skills among employees hinders AI adoption in SMEs, particularly in Africa.	(Borah et al., 2022; Madden & Kanos, 2020)
Data Availability and Quality	SMEs often struggle with collecting and maintaining high-quality data required for effective AI applications.	(Bettoni et al., 2021; Szedlak et al., 2020)
Cultural and Organisational Barriers	Resistance to change, lack of executive leadership, and poor integration with existing business models impede AI adoption.	(Iftikhar & Nordbjerg, 2022; Paul et al., 2022; Uwagaba et al., 2023)
Lack of Customised AI Solutions	Most AI solutions are designed for large enterprises and are not tailored to the specific needs of SMEs.	(Kaiseret al., 2023; Velmurugan et al., 2024)
Security and Privacy Concerns	Ethical concerns around data security and privacy present significant challenges, particularly in regions with weak governance.	(Borah et al., 2022; Partadiredja et al., 2020)
Regulatory Challenges	The absence of robust AI regulatory frameworks in many regions, especially in Africa, creates additional hurdles for SMEs.	(Madden & Kanos, 2020; Partadiredja et al., 2020)
Perceived Risks and Fear of AI	SMEs may fear losing control over business processes to AI, leading to hesitancy in adoption.	(Bettoni et al., 2021; Schoeman, 2024)

Table 2. Summary of Barriers to AI adoption in SMEs

1.6 Socio-Technical Factors Influencing the AI Divide

Integrating AI into various sectors brings a dual challenge: advancing technology while avoiding increasing social inequalities. Closing the AI divide requires a balanced look at social, technical, and socio-technical factors that affect adoption, with each area presenting specific challenges.

1.6.1 Social Factors

The social effects of AI adoption are substantial, particularly regarding bias, inequality, and ethical considerations. Lauterbach (2019) argues that combining technical skills with clear policy strategies is essential, but more than policy alone may be needed. Schwartz et al. (2022) emphasise the need for transparency to counter biases affecting marginalised groups, yet this is only achievable with solid regulations (Stypinska, 2022; Ulnicane et al., 2022)) view AI as both a technical and social system, where both sides need to be considered to avoid social harm. However, the current focus on technical efficiency often overshadows these social concerns. Hickok (2020) argues that existing regulations lack the depth needed to handle the broader socio-technical issues AI presents. Another barrier is cultural resistance; Asif et al. (2024) argued that organisations with a culture that supports innovation are more likely to adopt AI, while others who resist change tend to fall behind. Public trust is also crucial. Sartori and Theodorou (2022) point out that concerns over job security and data privacy lead to resistance.

1.6.2 Technical Factors

Technical challenges are also significant in AI adoption, often due to limited infrastructure, poor data quality, and weak regulations. Data quality is crucial; Cubric (2020) argued that poor data leads to inaccuracies and biases in AI systems. Ulrichet al. (2021) highlight the need for scalable infrastructure to make AI viable, but such investments are complex for many organisations, especially in lower-resourced sectors such as SMEs. Dwivedi et al. (2021) point out that access to digital infrastructure, like cloud services, is vital for AI readiness but can be too expensive for some sectors, leading to unequal access. In agriculture, for instance, (Abioye et al., 2021) shows how more infrastructure is needed to ensure AI reaches its full potential.

Regulation is also a factor; Du et al. (2024) criticise the gap between ethical AI discussions and actionable policies. (Jacobs and Simon, 2022) argue that the European Digital Services Act does not address AI's unique risks. Ridzuan et al. (2024)call for stronger regulations to ensure responsible AI use in finance. Still, the rapid development of AI makes it challenging for rules to keep up, creating inconsistencies that widen the divide.

1.6.3 Socio-Technical Factors

Socio-technical factors, which cover the interaction between social and technical systems, are crucial but often overlooked in efforts to bridge the AI divide. Weger and Yeazitzis (2023) suggest that adopting AI requires technical tools and social understanding. Digital literacy and specific skills are essential to engaging with AI, yet Yu et al. (2022) stated that many communities still need to prepare with targeted education. Dai & Liang (2022) argue that people need specific skills to interact effectively with AI systems, but disparities in access to education mean that some groups need to be included. Motivation and trust also play a significant role; (Upadhyay & Shukla, 2024) found that people adopt AI when they believe it will solve problems. Gudigantala et al. (2023) stress that organisations must communicate AI's benefits. Kelly(2023) highlights that trust in AI is critical, but many AI systems are opaque, making it hard for users to trust the technology entirely.

The model below, adapted from our previous work (Mohammed et al., 2023), highlights the socio-technical factors contributing to the AI divide. While the full discussion on the model's development is detailed in our earlier study, its inclusion here demonstrates how this research builds upon that foundation by applying the framework to Nigerian SMEs and expanding the analysis with new empirical findings.

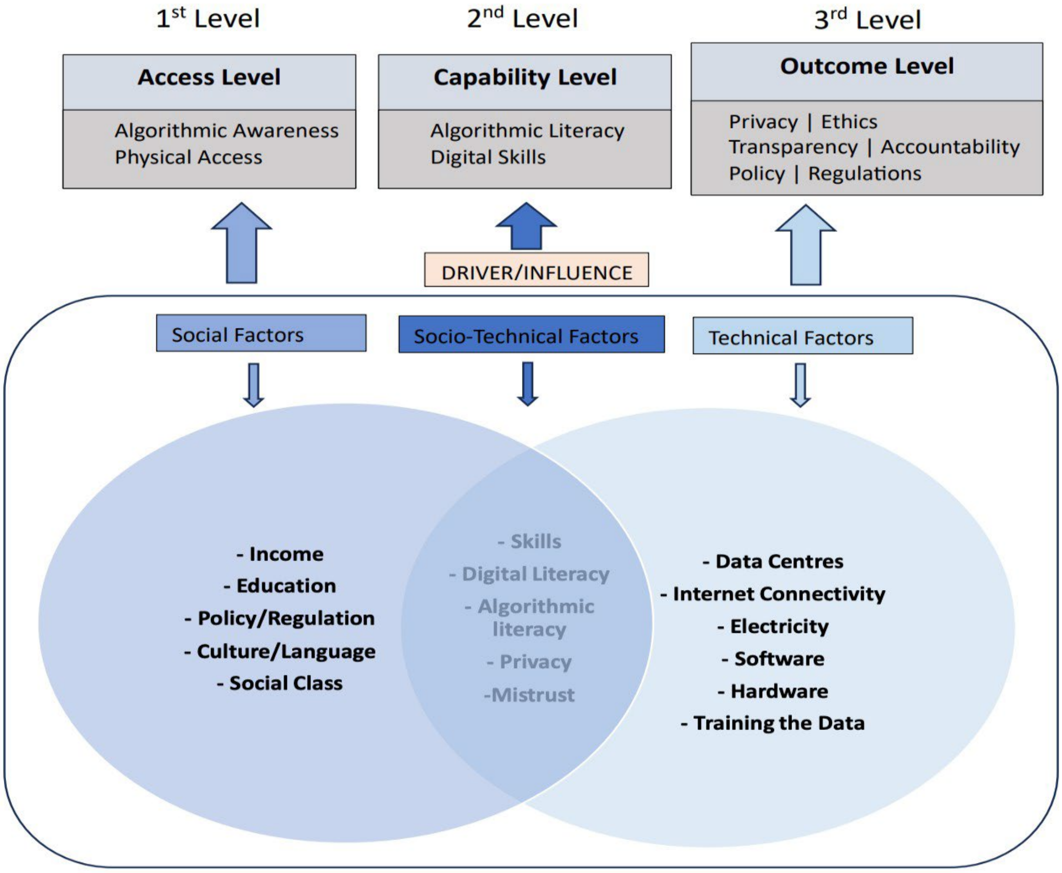


Fig 1. Factors contributing to the AI divide (Mohammed et al., 2023)

2.0 Methodology

Our methodology takes a case study approach, focusing on SMEs in Nigeria to understand the extent of AI adoption in different sectors and their perceptions of AI adoption barriers and challenges. This research is part of a broader study that will later include an in-depth investigation of a specific SME sector. The initial survey serves as a foundational step in the study, providing an overview of AI adoption trends and identifying key themes for further in-depth qualitative exploration via interviews. This approach is justified due to the high diversity of the SME population, which includes various industries and business types with owners and employees from various backgrounds (Curran & Blackburn, 2001). While interviews offer deeper insights into SME experiences, a survey was chosen at this stage to capture diverse perspectives from a larger sample, helping to map sectoral differences in AI adoption. By collecting responses from a diverse range of SMEs, the survey helps map out the different levels of the digital divide across various sectors, offering valuable insights into the socio-cultural and technological factors influencing AI adoption and enabling a sample to be drawn more widely from the SME population.

For this research, SMEs are classified based on their digital maturity, as it directly influences their ability to adopt AI technologies. AI implementation requires a foundational level of digital literacy, infrastructure, and integration capacity (Davenport et al., 2018). SMEs lacking any digital capability would struggle to engage with AI effectively. To ensure a structured classification, the study adapts existing Digital Maturity Models (DMM) (Kane et al., 2017) and aligns with frameworks used in SME digital transformation research (OECD, 2021; Omol, 2024).

Consequently, SMEs in Northern Nigeria are categorised into three levels of digital maturity:

Category	**Definitions and Characteristics**	**AI Readiness Level**	**Inclusion in Study**
Category A: Minimal Digital Infrastructure	SMEs primarily reliant on basic mobile devices, minimal internet usage, and lack of structured digital operations. No formal IT systems, ERP, or automation. Limited digital literacy among employees.	Not AI-ready – These SMEs do not meet the foundational technological requirements for AI adoption.	Excluded: (Omol, 2024)
Category B: Basic Digital Infrastructure	SMEs using basic digital tools (e.g., social media for marketing, basic cloud storage). Some employees have basic digital skills. No AI use yet, but willingness to adopt exists due to market and customer demands.	Emerging AI readiness – These SMEs are starting to explore AI-driven tools for efficiency, but face scalability challenges	Included: (Davenport et al., 2018)
Category C: Advanced Digital Transformation	SMEs actively using integrated digital solutions, such as CRM systems, cloud computing, and data analytics. Demonstrate a strategic approach to digital adoption and automation.	High AI readiness – positioned for AI integration and automation to optimize operations.	Included: (OECD, 2021)

Table 3. Categories of AI Readiness in SMEs

Only Category B and C SMEs are included, as their digital readiness aligns with the study’s focus on AI adoption potential. Furthermore, the study targets SMEs in NN specifically, given the region’s unique socio-economic conditions, such as limited digital infrastructure and cultural resistance to technology (Lawal, 2022; Ojeme, 2018). This approach aims to address the digital divide and foster AI adoption in NN SMEs

2.2 Data collection

The study's data collection instrument was a survey questionnaire. This instrument is designed to capture a broad view of SMEs' different sectors and their perceptions of AI adoption. Given the diverse and heterogeneous nature of SMEs in Northern Nigeria, the survey allows respondents to express their views, challenges, and readiness for AI adoption in their own words.

This approach incorporates both closed- and open-ended questions, enabling the researcher to gather rich qualitative data reflecting SMEs' varied digital maturity levels across different sectors.

Microsoft forms were used to generate the questions. Microsoft Forms was chosen as the survey instrument for this study due to its user-friendly interface, accessibility, and flexibility in designing and distributing open-ended surveys. It also provides real-time data collection and analytics features, making monitoring responses and gaining immediate insights easier. Given the widespread familiarity with Microsoft products among SMEs, using Microsoft Forms reduces potential technological barriers, ensuring a higher response rate and enhancing the overall quality and reliability of the data collected.

The sampling strategy for this research combined purposive, criterion-based, and snowball sampling techniques, with support from government agencies (Small and Medium Enterprises Development Agency of Nigeria and National Information Technology Development Agency) to ensure a comprehensive understanding of AI adoption challenges and the socio-technical factors contributing to the AI divide among SMEs in Northern Nigeria. This approach was designed to capture diverse experiences and perspectives while ensuring the inclusion of SMEs most relevant to the study's objectives. 144 SMEs across various sectors filled out the survey within three months. The qualitative data for this study was analysed using thematic analysis, following the approach by Braun & Clarke (2006). The process involved organising the data into meaningful themes, going beyond simple description to interpret fundamental aspects of the research topic. As Boyatzis (1998) noted, thematic analysis allows for deeper insights within the data. This interpretative approach helped uncover both explicit and implicit meanings in the data.

2.3 Sample Size & Population

This study surveyed 144 SMEs in Northern Nigeria to examine AI adoption across different sectors. The survey was distributed to over 200 SMEs, but 144 responded, providing a 72% response rate. The sample includes businesses from IT, Health, Construction, Agriculture, Education, Mining, Retail, Fashion, Finance, Manufacturing, and Food-related industries, ensuring sectoral diversity.

Data collection took place from August 2, 2024, to August 27, 2024, using a structured survey approach. The survey included both closed and open-ended questions, allowing for quantitative insights on AI adoption trends and qualitative responses to capture SMEs' perceptions, challenges, and expectations regarding AI technologies. SMEs were selected based on their digital maturity, with only those classified as Basic (Category B) or Advanced (Category C) Digital Infrastructure included, as they demonstrate AI adoption potential.

Focusing on SMEs in Northern Nigeria provides insights into regional challenges such as limited digital infrastructure, skill shortages, and cultural barriers to technology adoption. This targeted approach helps identify sector-specific AI adoption trends and informs strategies for bridging the AI divide in developing economies. While statistical calculations were not used to determine the sample size, the diverse sectoral distribution and response rate enhance the validity and generalizability of the findings within the context of digitally capable SMEs. To ensure research integrity and participant protection, ethical guidelines were strictly followed. Informed consent was obtained from all participants before data collection, and they were informed about the purpose of the study, voluntary participation, and the right to withdraw at any time. Additionally, data anonymity and confidentiality were maintained by removing identifiable information and securely storing survey responses. Ethical approval for this research was obtained from the Salford Business School Ethics Committee. This approval demonstrates that the study approach adhered to institutional and international ethical standards, reinforcing the credibility and ethical integrity of the research process.

3.0 Data Analysis and Presentation

This study employed a mixed-methods approach, combining quantitative and qualitative analyses to gain a comprehensive understanding of AI adoption among SMEs in Northern Nigeria. The quantitative analysis involved analysing closed-ended survey responses using descriptive statistics. This provided insights into AI adoption levels, digital maturity classifications, and sectoral variations among SMEs. Frequency distributions and percentages were used to identify key trends, such as the proportion of SMEs adopting AI, major barriers to adoption, and differences across industries. These statistical findings offered a structured assessment of digital readiness and the extent of the AI divide among SMEs.

For the qualitative analysis, open-ended responses were analysed using thematic analysis, guided by Braun & Clarke's (2006) six-step framework. This process included:

1. Familiarization with the Data: Responses were reviewed multiple times to identify key patterns.
2. Generating Initial Codes – Recurring concepts related to AI adoption were extracted.
3. Searching for Themes –Similar codes were grouped into broader themes such as technical, social, and socio-technical barriers.
4. Reviewing Themes – Themes were refined to ensure consistency with the data.
5. Defining and Naming Themes – Themes were clearly labelled to improve clarity and interpretation.
6. Producing the Report – Direct SME quotes were included to support findings and provide deeper insight.

By integrating statistical insights with thematic findings, this study offers a holistic perspective on AI adoption challenges. The quantitative data highlights the extent of adoption and key obstacles, while qualitative insights reveal SMEs' perceptions, motivations, and sector-specific barriers. This combined approach enhances the study's ability to formulate targeted recommendations for addressing the AI divide and improving SME digital readiness.

3.1 Level of AI skill and AI adoption by various sectors

The closed-ended questions asked participants for information about their sector, their familiarity and confidence with AI, and the level of AI skill within the organisation. The survey highlights significant disparities in AI adoption across SME sectors in Northern Nigeria. 40.8% of the respondents responded they are adopting AI from the IT sector, reflecting its digital orientation and access to better infrastructure. Retail follows with 30.3%, showing growing interest in AI for customer engagement and inventory management despite limitations like poor internet and resource constraints.

Conversely, sectors such as food (14.8%) and agriculture (7.7%) recorded moderate engagement, constrained by traditional practices, resource limitations, and lower awareness of AI's applications. Sectors like fashion (2.1%), construction (1.4%), banking (1.4%), and mining (0.7%) saw minimal participation, likely due to infrastructural deficiencies, high costs, and limited knowledge about AI.

However, the disparity in survey participation across sectors highlights the region's socio-economic and infrastructural challenges SMEs face. This emphasises the need for tailored interventions, capacity-building programs, and improved access to AI technologies to bridge the divide and foster inclusive adoption.

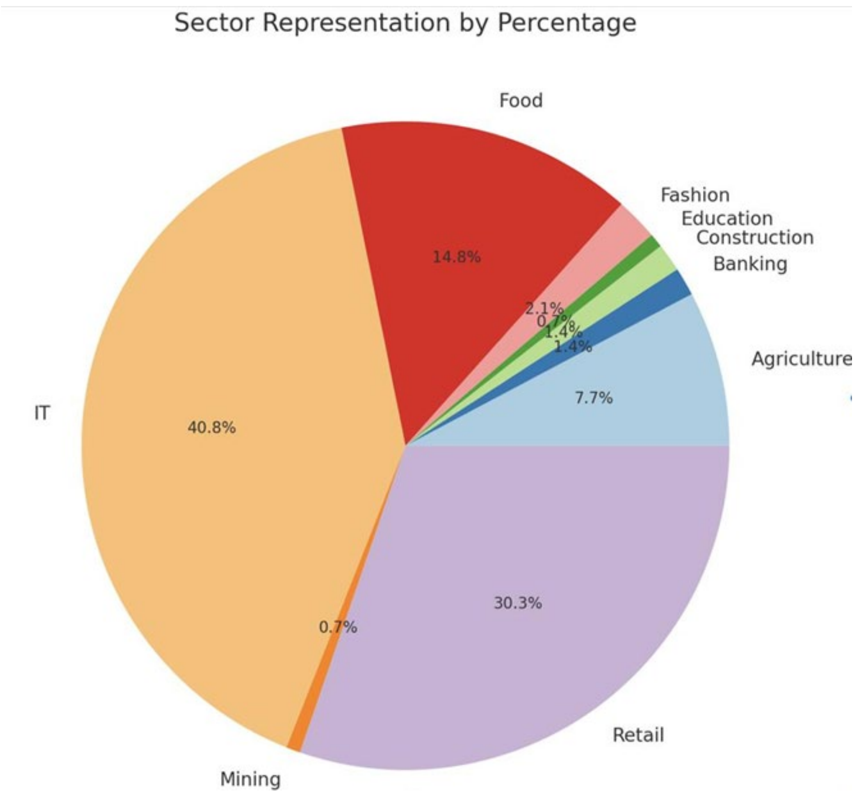


Fig 2. **Sectors of SME Participation in the Survey**

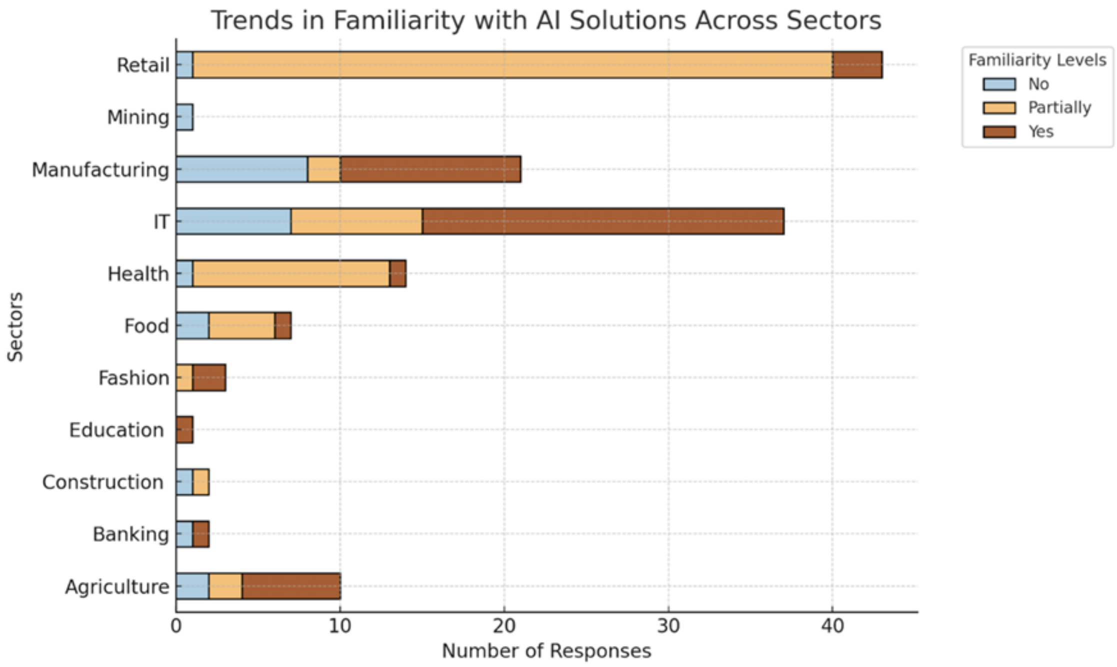


Fig 3. **SMEs’ familiarity with AI**

Sector	**Estimated AI Adoption**	**Confidence**	**Skills**	**Familiarity**
IT	80-90%	High	High	High
Banking	60-70%	Moderate to High	Moderate	Moderate
Agriculture	40-50%	Low to Moderate	Moderate	Low to Moderate
Construction	30-40%	Moderate	Moderate	Low
Education	50-60%	Moderate	Moderate	Moderate
Fashion	50-60%	High	Moderate	Moderate

Table 4. Summary of AI Confidence, skills and familiarity in SMEs

3.2 Confidence in AI Technologies

Confidence in AI adoption varied across sectors. The IT sector showed the highest confidence due to engagement with digital technologies. In Agriculture, confidence was mixed, reflecting uncertainty about AI's relevance to traditional practices. Banking showed strong confidence, tempered by concerns over regulation and job displacement. Construction expressed moderate confidence, recognising AI's potential but not seeing it as essential. Education was positive, but there was a need for more exposure. Fashion displayed growing confidence, especially in AI for inventory and design. Overall, the IT sector led in readiness, driven by its technological expertise.

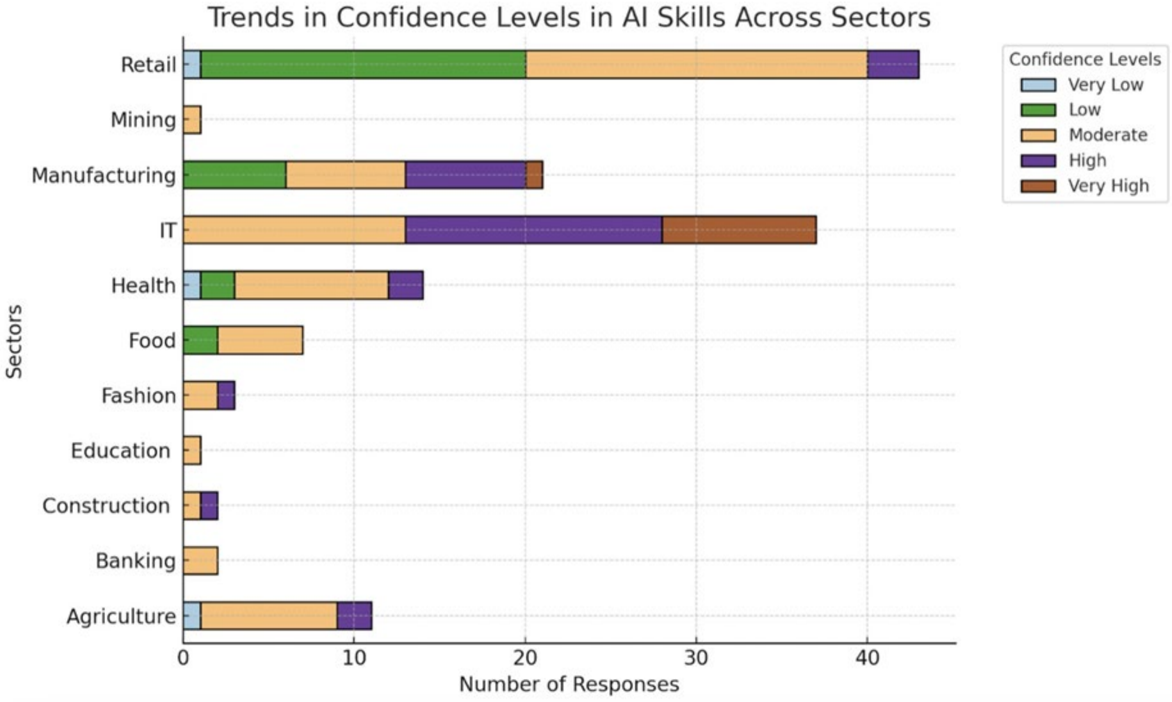


Figure 4. Trends in confidence levels in AI skills across sectors

3.3 Challenges of AI Adoption in SMEs in Nigeria

The following analysis classifies the challenges mentioned by the different sectors into technical, social, and socio-technical themes. Each theme includes relevant sub-themes and examples from the responses provided by SMEs.

3.3.1 Technical Challenges

Technical challenges are a significant barrier to AI adoption across sectors, mainly due to infrastructure limitations, outdated technologies, and data quality issues. In agriculture, poor internet connectivity and high infrastructure costs prevent effective adoption, frequently mentioning “poor internet service” and a “lack of system.” Similarly, the banking sector faces challenges related to inadequate power supply and technological infrastructure, with references to “electricity and computer” limitations, hindering its ability to integrate AI effectively.

Other sectors face unique challenges. In food, outdated technologies and unskilled personnel are common issues, as participants stated, “My device is obsolete” and “Some features are not available.” Health sectors suffer from network and power issues alongside poor data quality, with observations such as “errors with AI use” and “data quality and availability.” Similarly, the IT sector struggles with “outdated hardware,” “inadequate network infrastructure,” and the cost of the internet, complicating system integration and performance. Manufacturing and retail also report similar infrastructural shortcomings, such as “insufficient power supply” and difficulties accessing “data equality and availability.” These examples highlight the need to address these technical challenges to unlock AI’s potential across industries.

3.3.2. Social Challenges

The thematic analysis of social challenges highlights vital barriers to AI adoption that stem from socio-cultural and behavioural factors, including resistance to change, awareness gaps, concerns over job displacement, and issues of trust and accessibility. Low literacy levels and a lack of awareness are prominent barriers in sectors like agriculture and food. For example, one participant expressed difficulty "meeting up with the current technology like layers cage," while another cited "low literacy" and the challenge of affording AI tools. Similarly, in fashion and IT, resistance to change is evident, with participants noting challenges in "getting staff to understand" and stating that "training staff on the usage of the new technologies" is time-consuming.

Concerns about trust and job displacement also emerged as critical challenges, particularly in health and retail. Participants highlighted "public trust and acceptance" issues and expressed concerns about AI's potential to "substitute human beings with machines." Accessibility to AI technologies further compounds these challenges, with several respondents pointing out affordability and infrastructure barriers. For instance, one participant in the manufacturing sector shared that "access to the new technology is the major problem we are facing here." At the same time, another stated, "AI needs to be easily accessible to the extent that everyone can use it." These insights underscore the need for targeted strategies to address socio-cultural resistance, build awareness, and improve accessibility and trust in AI solutions.

3.3.3 Socio-Technical Challenges

The data analysis reveals the socio-technical challenges different sectors face when adopting AI. It highlights issues related to skills and training, ethical concerns, regulatory gaps, and cybersecurity risks, which collectively impact the successful implementation of AI technologies. Socio-technical challenges in AI adoption arise from social and technical factors, including skills shortages, training gaps, cybersecurity risks, bias, and regulatory issues. Across sectors, these challenges manifest differently. Accessing the “right skills to implement the technology” is a significant agricultural barrier. Similarly, the construction sector struggles with skill shortages, while the fashion industry highlights difficulties in staff adaptation, such as “getting staff to understand” new AI tools. The food sector also lacks understanding, with participants citing “unskilled personnel” and an inability to grasp “new technological languages.” These challenges underline the need for tailored training and capacity-building programs.

Other sectors face additional socio-technical barriers. Ethical concerns, bias, and cybersecurity risks are pressing issues in healthcare, with mentions of “bias and discrimination” and “cybersecurity risks.” The IT sector struggles with the complexity of AI, compounded by the “unavailability of experts” and a “knowledge gap,” making AI integration challenging. Manufacturing and retail experience similar hurdles, such as the “lack of skilled personnel” and ethical concerns like AI's “black box nature.” Regulatory challenges, such as a “lack of transparency and accountability” and inadequate cybersecurity measures, further complicate AI adoption. Addressing these socio-technical challenges is essential for fostering trust and maximising the benefits of AI in these industries.

3.4 Mapping Socio-Technical Factors to Various Levels of the AI Divide

As discussed earlier, the AI divide can be understood through three interconnected levels: Access, Capability, and Outcomes. Each level represents a different aspect of how businesses interact with AI technologies, and the factors identified in the study contribute to the divide at each level.

3.4.1 Access Level: Barriers to Accessing AI Technology

Access to AI is determined by the availability of essential infrastructure, tools, and technologies required to implement AI. This level of the AI divide is primarily influenced by technical factors such as infrastructure and outdated technologies, but social factors like resistance to change and awareness of AI also contribute to the AI divide at the level.

At the access level, technical factors such as poor infrastructure, including unreliable internet connectivity and inconsistent power supply, are significant barriers to AI adoption, particularly in sectors like Agriculture and Manufacturing, where respondents cited issues like "Poor internet connectivity" and "Insufficient power supply." Additionally, using outdated technology in sectors like IT and Food further hinders access to advanced AI tools, with respondents mentioning "Outdated hardware" and "Device is obsolete." On the social side, resistance to change within organisations discourages the exploration of AI technologies, as evidenced in sectors like Agriculture and Food, where responses included "Yes, adopting the technology is difficult" and "We feel it is not important."

These factors contribute to the AI divide at the access level by preventing SMEs from acquiring or implementing AI tools. Businesses with proper infrastructure or those unwilling to adopt AI due to cultural resistance are included in AI opportunities, creating a gap between those who can access AI and those who cannot.

3.4.2 Capability Level: Skills and Knowledge Gaps in Using AI

At the capability level, socio-technical factors such as the lack of skills and training to operate AI technologies were prevalent, especially in sectors like IT, Agriculture, and Manufacturing.

Respondents frequently mentioned "We lack proper knowledge of using AI" and "The ability to access the right skills to implement the technology isa major challenge." In addition, social factors like low digital literacy and lack of awareness about AI’s benefits were barriers in sectors such as Agriculture, Fashion, and Food, where comments like "Low literacy" and "The staff are not familiar with most technologies" were common. Even when AI tools are available, technical factors such as outdated hardware prevent businesses from effectively building AI capability. This challenge was particularly noted in Manufacturing and IT, with respondents citing "Outdated hardware" and "Difficulty integrating new tools with existing systems." These factors collectively hinder many SMEs from using AI effectively, contributing to the AI divide at the capability level. Businesses with access to AI but lacking the ability to use it risk falling behind in AI-driven solutions.

3.4.3 Outcome Level: Impact of AI on Business Performance and Growth

At the outcome level, the AI divide is shaped by the benefits or consequences businesses experience after implementing AI technologies. This stage is heavily influenced by both technical and socio-technical factors, particularly in terms of how well businesses can integrate AI into their operations and address ethical and regulatory challenges. Socio-technical factors, such as integration challenges, prevent businesses from fully realising AI's potential in sectors like IT and Manufacturing, where respondents mentioned "Difficulty integrating new tools with existing systems" and "Integration issues." Additionally, ethical concerns and regulatory challenges were prominent in sectors like Health and Retail, where worries about "Bias and discrimination" and "Cybersecurity risks" limited businesses' willingness to deploy AI fully. On the technical side, data quality and availability issues also constrained the effectiveness of AI, particularly in IT and Retail, with respondents reporting "Insufficient or poor-quality data to train or implement AI models" and "Data equality and availability." The impact of these challenges varies: businesses that can overcome integration, ethical, and regulatory barriers experience significant gains in productivity and innovation, while those facing issues are left behind, contributing to the AI divide at the outcome level.

3.5 Summary of Contributions to the AI Divide Across Levels

Table 4 highlights the various levels of the AI divide and their contributing factors across different sectors. These factors collectively exacerbate the AI divide, making it difficult for SMEs to harness the potential benefits of AI.

Level of AI Divide	**Contributing Factors**	**Example Sectors**	**Impact**
Access	– Poor infrastructure – Outdated technologies Resistance to change – Awareness	– Agriculture – IT – Manufacturing	Limited access to AI tools and infrastructure prevents SMEs from adopting AI. Similarly, awareness and resistance to change increases AI divide
Compatibility	– Lack of skills and training – Low digital literacy Outdated systems	– IT – Agriculture – Food	Even with access to AI, SMEs struggle to use it due to lack of knowledge and outdated systems.
Outcome	– Integration challenges – Ethical and regulatory concerns – Poor data quality – Privacy – Cyber security – Bias and Discrimination	– Health – Retail – Manufacturing	AI adoption does not translate into better business outcomes due to integration issues and ethical concerns.

Table 5. Summary of Contributions to the AI divide across levels

4.0 Discussion and Conclusions

This study leverages empirical data to address critical gaps identified in the systematic literature review (SLR), providing a deeper understanding of the socio-technical challenges contributing to the AI divide among SMEs in Northern Nigeria. While the SLR highlighted general barriers related to access, skills, and outcomes, it lacked context-specific insights into how these challenges manifest across different sectors.

The present study addresses this limitation by using sector-specific data to explore the socio-technical factors influencing AI adoption.

At the access level, the findings validate the SLR's observation that access to AI technologies is limited to larger organisations, leaving SMEs disadvantaged. Empirical evidence from this study highlights infrastructural challenges such as "poor internet service" and "insufficient power supply" in agriculture and manufacturing, aligning with the SLR's emphasis on physical access barriers. Furthermore, the study builds on the SLR by revealing that affordability, evidenced by references to the "high cost of data subscription" and "cost of AI solutions," compounds these access challenges. Addressing these gaps requires integrating technical solutions with policies that democratise access to AI tools and provide financial support for SMEs.

At the capability level, the study underscores the socio-technical challenges related to algorithmic literacy and skills shortages. While the SLR identified the lack of algorithmic expertise as a critical barrier, this study contextualises these findings within SMEs, highlighting specific challenges such as "knowledge gaps," "unavailability of experts," and "difficulty integrating new tools." Sectors like IT and health report significant struggles with training and retraining, further exacerbating the divide.

At the outcome level, this study extends the SLR's discussion of biases and inequalities in AI systems by providing empirical evidence of their practical implications. Sectors such as health and retail reported issues like "bias and discrimination," "cybersecurity risks," and "inequality access," which undermine trust and adoption of AI technologies. Moreover, the lack of "transparency" and "accountability" identified in the study aligns with the SLR's call for algorithmic auditing and ethical frameworks, which are essential for fostering trust and ensuring equitable outcomes.

AI adoption among SMEs in Northern Nigeria can be enhanced through community-driven initiatives and public-private partnerships (PPPs). Government-backed digital literacy programs, industry collaboration, and innovation hubs can provide training, funding, and technical support. Partnerships with tech firms can offer affordable AI solutions and mentorship, while business incubators can facilitate AI experimentation. Case studies of successful AI adoption strategies could guide SMEs in low-cost AI integration. These initiatives promote inclusive and sustainable AI adoption, benefiting small businesses and the local economy.

This study provides important insights into AI adoption among SMEs in Northern Nigeria; however, certain limitations must be acknowledged. The reliance on self-reported survey responses may introduce biases in perception and interpretation, as respondents may overestimate or underestimate their digital readiness. Additionally, the study focuses on SMEs with some level of digital maturity (Category B and C), excluding those with minimal technological adoption (Category A), which limits the understanding of barriers faced by SMEs at the lowest end of digital adoption. The sample size of 144 SMEs, while diverse, may not fully capture the complexities of AI adoption across different SME sizes and business models. Lastly, thematic analysis of open-ended responses, while useful, could have been complemented by in-depth interviews or case studies to provide richer qualitative insights.

In conclusion, the socio-technical challenges identified in this study illustrate the complex factors contributing to the AI divide in SMEs. Addressing these challenges will require technical solutions, policy interventions, and capacity building. This research provides valuable insights into SMEs' unique challenges in Northern Nigeria and demonstrates the importance of socio-technical theory in understanding and addressing the AI divide. Future research should explore these issues in broader contexts and identify practical strategies to help SMEs fully benefit from AI technologies.

5.0 References

Abayomi, O. K., Adenekan, F. N., Abayomi, A. O., Ajayi, T. A., & Aderonke, A. O. (2021). Awareness and Perception of the Artificial Intelligence in the Management of University Libraries in Nigeria. Journal of Interlibrary Loan, Document Delivery and Electronic Reserve, 29(1–2), 13–28. https://doi.org/10.1080/1072303X.2021.1918602

Abioye, S. O., Oyedele, L. O., Akanbi, L., Ajayi, A., Davila Delgado, J. M., Bilal, M., Akinade, O. O., & Ahmed, A. (2021). Artificial intelligence in the construction industry: A review of present

status, opportunities and future challenges. Journal of Building Engineering, 44, 103299. https://doi.org/10.1016/J.JOBE.2021.103299

Anazodo, U. C., Adewole, M., & Dako, F. (2022). AI for Population and Global Health in Radiology. In Radiology: Artificial Intelligence (Vol. 4, Issue 4). Radiological Society of North America Inc. https://doi.org/10.1148/ryai.220107

Asif, M., Yang, L., & Hashim, M. (2024). The Role of Digital Transformation, Corporate Culture, and Leadership in Enhancing Corporate Sustainable Performance in the Manufacturing Sector of China. Sustainability 2024, Vol. 16, Page 2651, 16(7), 2651. https://doi.org/10.3390/SU16072651

Baabdullah, A. M., Alalwan, A. A., Slade, E. L., Raman, R., & Khatatneh, K. F. (2021). SMEs and artificial intelligence (AI): Antecedents and consequences of AI-based B2B practices. Industrial Marketing Management, 98, 255–270. https://doi.org/10.1016/J.INDMARMAN.2021.09.003

Ball, C., & Huang, K.-T. (2023). Generative Artificial Intelligence (GAI) Divide: An Empirical Examination of the Micro–Macro Factors that Predict GAI Knowledge and Use. Proceedings of the Association for Information Science and Technology, 60(1), 878–880. https://doi.org/10.1002/PRA2.884

Bettoni, A., Matteri, D., Montini, E., Gladysz, B., & Carpanzano, E. (2021). An AI adoption model for SMEs: a conceptual framework. IFAC-PapersOnLine, 54(1), 702–708. https://doi.org/10.1016/J.IFACOL.2021.08.082

Bjola, C. (2022). AI for development: Implications for theory and practice. Oxford Development Studies, 50(1), 78–90.

Borah, P. S., Iqbal, S., & Akhtar, S. (2022). Linking social media usage and SME's sustainable performance: The role of digital leadership and innovation capabilities. Technology in Society, 68, 101900. https://doi.org/10.1016/J.TECHSOC.2022.101900

Boustani, N. M. (2022). Artificial intelligence impact on banks clients and employees in an Asian developing country. Journal of Asia Business Studies, 16(2), 267–278. https://doi.org/10.1108/JABS-09-2020-0376

Boyatzis, R. (1998). Transforming qualitative information: Thematic analysis and code development. https://books.google.com/books?hl=en&lr=&id=_rfClWRhIKAC&oi=fnd&pg=PR6&dq=Boya tzis+(1998)+thematic+analysis&ots=ECpGBgap7n&sig=KpcrAiT_jwn2w2TNyWSxZAR622g

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706QP063OA

Chui, M., Hazan, E., Roberts, R., Singla, A., & Smaje, K. (2023). The economic potential of generative AI.

Cubric, M. (2020). Drivers, barriers and social considerations for AI adoption in business and management: A tertiary study. Technology in Society, 62. https://doi.org/10.1016/j.techsoc.2020.101257

Dai, B., & Liang, W. (2022). The Impact of Big Data Technical Skills on Novel Business Model Innovation Based on the Role of Resource Integration and Environmental Uncertainty. Sustainability 2022, Vol. 14, Page 2670, 14(5), 2670. https://doi.org/10.3390/SU14052670

Davenport, T. H., Ronanki, R., Wheaton, J., & Nguyen, A. 2018. Feature Artificial Intelligence For The Real World 108 Harvard Business Review.

Denicolai, S., Zucchella, A., and, G. M.-T. F., & 2021, undefined. (n.d.). Internationalization, digitalization, and sustainability: Are SMEs ready? A survey on synergies and substituting effects among growth paths. Elsevier. Retrieved November 14, 2024, from https://www.sciencedirect.com/science/article/pii/S0040162521000822

Du, H., Niyato, D., Kang, J., Xiong, Z., Zhang, P., Cui, S., & Kim, D. I. (2024). The age of generative AI and AI-generated everything. IEEE Network.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., … Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary Perspectives On Emerging Challenges, Opportunities, And Agenda For Research, Practice And Policy. International Journal of Information Management, 57. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Etuk, R. U., Etuk, G. R., & Michael, B. (2014). Small And Medium Scale Enterprises (Smes) And Nigeria's Economic Development. Mediterranean Journal of Social Sciences, 5(7), 656–662. https://doi.org/10.5901/mjss.2014.v5n7p656

Falode, Faseke, A., Ikeanyichukwu, B. O. &, & Chukwuma. (n.d.). Artificial Intelligence: The Missing Critical Component in Nigeria's Security Architecture Falode / Faseke / Ikeanyichukwu (2021) Artificial Intelligence: The Missing Critical Component in Nigeria's.

Gudigantala, N., Madhavaram, S., & Bicen, P. (2023). An AI Decision-Making Framework For Business Value Maximization. AI Magazine, 44(1), 67–84. https://doi.org/10.1002/AAAI.12076

Haenlein, M., & Kaplan, A. (2019). A Brief History Of Artificial Intelligence: On The Past, Present, And Future Of Artificial Intelligence. California Management Review, 61(4), 5–14.

Hansen, E. B., & Bøgh, S. (2021). Artificial Intelligence and Internet Of Things In Small And Medium-Sized Enterprises: A survey. Journal of Manufacturing Systems, 58, 362–372. https://doi.org/10.1016/J.JMSY.2020.08.009

Hickok, M. (2020). Lessons learned from AI ethics principles for future actions. AI and Ethics 2020 1:1, 1(1), 41–47. https://doi.org/10.1007/S43681-020-00008-1

Huang, M. H. , & Rust, R. T. (2018). Artificial intelligence in service. Journal of Service Research, 21(2), 155–172.

Iftikhar, N., & Nordbjerg, F. E. (2022). Implementing Machine Learning in Small and Medium-Sized Manufacturing Enterprises. Lecture Notes in Mechanical Engineering, 448–456. https://doi.org/10.1007/978-3-030-90700-6_51/FIGURES/2

Jacobs, M. , & Simon, J. (2022). Assigning obligations in AI regulation: A discussion of two frameworks proposed by the European Commission. Digital Society, 1(1), 6.

Jafari-Sadeghi, V., Amoozad Mahdiraji, H., Busso, D., & Yahiaoui, D. (2022). Towards agility in international high-tech SMEs: Exploring key drivers and main outcomes of dynamic capabilities. Technological Forecasting and Social Change, 174. https://doi.org/10.1016/J.TECHFORE.2021.121272

Jahanbakhsh Javid, N. , & Amini, M. (2023). Evaluating the effect of supply chain management practice on implementation of halal agroindustry and competitive advantage for small and medium enterprises. International Journal of Computer Science and Information Technology, 15, 8997–9008.

Kaiser, J., Terrazas, G., McFarlane, D., & de Silva, L. (2023). Towards low-cost machine learning solutions for manufacturing SMEs. AI and Society, 38(6), 2659–2665. https://doi.org/10.1007/S00146-021-01332-8/TABLES/2

Lauterbach, A. (2019). Artificial intelligence and policy: quo vadis? Digital Policy, Regulation and Governance , 21(3), 238–263. https://doi.org/10.1108/DPRG-09-2018-0054

Lawal, L. O. , N. M. , & S. A. A. (2022). An Assessment of the Impact of Boko Haram Insurgency on Small and Medium Enterprises (SMEs) in Gombe State of Nigeria: Challenges and Prospects. The International Journal of Business & Management, 10(3).

Madden P, & Kanos D. (2020). Figures of the week: digital skills and the future of work in Africa. Brookings Institution. https://www.brookings.edu/articles/figures-of-the-week-digital-skills-and-the-future-of-work-in-africa/

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. Journal of Business Research, 120, 262–273. https://doi.org/10.1016/j.jbusres.2020.07.045

Mannuru, N. R., Shahriar, S., Teel, Z. A., Wang, T., Lund, B. D., Tijani, S., & Vaidya, P. (2023). Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development. Information Development.

McKinsey Global Report. (2023). The economic potential of generative AI: The next productivity frontier.

Mittal, S., Khan, M. A., Romero, D., & Wuest, T. (2018). A critical review of smart manufacturing & Industry 4.0 maturity models: Implications for small and medium-sized enterprises (SMEs). Journal of Manufacturing Systems, 49, 194–214. https://doi.org/10.1016/J.JMSY.2018.10.005

Mobayo, J. O. , A. A. F. , Y. S. O. , & B. U. (2021). Artificial intelligence: Awareness and adoption for effective facilities management in the energy sector.

Mohammed, A. L. , K. M. , & A. M. (2023). Conceptualising the Artificial Intelligence Divide: A Systematic Literature Review and Research AgendaLiterature Review and Research Agenda. AIS Electronic Library (AISeL)AIS Electronic Library (AISeL).

Muhammad, L. J., & Algehyne, E. A. (2021). Fuzzy based expert system for diagnosis of coronary artery disease in nigeria. Health and Technology, 11(2), 319–329. https://doi.org/10.1007/s12553-021-00531-z

Neeli, A. K. (2020). Impact and Role of Artificial Intelligence in Sales and Marketing. I-Manager's Journal on Management, 15(1), 1.

OECD. (2023). Artificial intelligence: Changing landscape for SMEs. https://www-oecd-ilibrary-org.salford.idm.oclc.org/sites/01a4ae9d-en/index.html?itemId=/content/component/01a4ae9d-en#:~:text=AI%20can%20substantially%20affect%20SME,the%20costs%20of%20experime ntation%20and

Ogbe, S. E., Osayi, C. P., & Amadi. (n.d.). Determinants of Venture Capital Financing among Micro-Agro Enterprises in Abia State, Nigeria.

Ojeme, M. , R. A., & C. N. (2018). Investigating the Nigerian small and medium enterprises (SMEs)-banking long-term relationship building. Nternational Journal of Bank Marketing.

Oldemeyer, L., Jede, A., & Teuteberg, F. (2024). Investigation of artificial intelligence in SMEs: a systematic review of the state of the art and the main implementation challenges. Management Review Quarterly. https://doi.org/10.1007/s11301-024-00405-4

Omol, E. J. (2024). Organizational digital transformation: from evolution to future trends. Digital Transformation and Society, 3(3), 240–256. https://doi.org/10.1108/DTS-08-2023-0061

Oregwu, U., & Chima, B. (2013). The Role of Small and Medium Scale Enterprises in Poverty Reduction in Nigeria: 2001-2011. An International Multidisciplinary Journal, Ethiopia, 7(4), 1–25. https://doi.org/10.4314/afrrev.7i4.1

Partadiredja, R. A., Serrano, C. E., & Ljubenkov, D. (2020, November 26). AI or human: The socio-ethical implications of ai-generated media content. 13th CMI Conference on Cybersecurity and Privacy - Digital Transformation -

Potentials and Challenges, CMI 2020. https://doi.org/10.1109/CMI51275.2020.9322673

Paul, S., Yuan, L., Jain, H., Sporher, J., & Lifshitz-Assaf, H. (2022). Intelligence Augmentation: Human Factors in AI and Future of Work. AIS Transactions on Human-Computer Interaction, 426–445. https://doi.org/10.17705/1thci.00174

Radziwon, A., Bogers, M. L. A. M., Chesbrough, H., & Minssen, T. (2022). Ecosystem effectuation: creating new value through open innovation during a pandemic. R&D Management, 52(2), 376–390. https://doi.org/10.1111/RADM.12512

Ridzuan, A. R., Jamri, M. H., Mohd Zain, Khairuddin, K. , Ibrahim, N., Abdul Rani, N. S., & Ab Hadi, S. N. I. (2024). The usage of ChatGPT as an alternative way of learning. E-Journal of Media and Society, 7(2), 95–101.

Sakapaji, S. C., & Puthenkalam, J. J. (2023). Harnessing AI for Climate-Resilient Agriculture: Opportunities and Challenges. European Journal of Theoretical and Applied Sciences, 1(6), 1144–1158.

Sartori, L., & Theodorou, A. (2022). A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. Ethics and Information Technology, 24(1). https://doi.org/10.1007/s10676-022-09624-3

SBE. (2023). Small Business AI Adoption Survey October 2023.

Schoeman, A. (2024). An exploratory study on supplier acceptance of and engagement in a national tax lottery designed to improve tax compliance. South African Journal of Accounting Research. https://doi.org/10.1080/10291954.2024.2393913

Schönberger, M. (2023). ARTIFICIAL INTELLIGENCE FOR SMALL AND MEDIUM-SIZED ENTERPRISES: IDENTIFYING KEY APPLICATIONS AND CHALLENGES. JOURNAL OF BUSINESS MANAGEMENT, 21, 89–112. https://journals.riseba.eu/index.php/jbm/article/view/336

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. https://doi.org/10.6028/NIST.SP.1270

Stypinska, J. (2022). AI ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies. AI and Society. https://doi.org/10.1007/s00146-022-01553-5

Syam, N., & Sharma, A. (2018). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. Industrial Marketing Management, 69, 135–146.

Szedlak, C., Poetters, P. , & Leyendecker, B. (2020). Application of artificial intelligence in small and medium-sized enterprises. Proceedings of the International Conference on Industrial Engineering and Operations Management.

Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., & Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. Journal of Business Research, 122, 502–517. https://doi.org/10.1016/J.JBUSRES.2020.09.009

Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W. G. (2022). Chapter 2 Governance of Artificial Intelligence. The Global Politics of Artificial Intelligence, 29–55. https://doi.org/10.1201/9780429446726-2

Ulrich, P., Frank, V., & Kratt, M. (2021). Adoption of Articial Intelligence in German SMEs– Results from an Empirical Study. https://aisel.aisnet.org/treos_amcis2021/23/

United Nations. (2024). UN Global Digital Compact 2024.

Upadhyay, A., & Shukla, A. C. (2025). Development of circular supply chain implementation model for MSMEs using extended theory of planned behaviour and DEMATEL approach. Management Science Letters, 15(3), 113–122. https://doi.org/10.5267/J.MSL.2024.6.003

Uwagaba, J., Omotosho, T. D., & George, G. O. (2023). Exploring the Barriers to Artificial Intelligence Adoption in Sub-Saharan Africa's Small and Medium Enterprises and the Potential for Increased Productivity. World Wide Journal of Multidisciplinary Research and Development.

Velmurugan, R., Thirumalaisamy, R., Paquibut, R., & Abouraia, M. (2024). An integrated framework for the financial sustainability and entrepreneurial success of micro, small, and

medium enterprises. Studies in Systems, Decision and Control, 525, 233–246. https://doi.org/10.1007/978-3-031-54383-8_19

Weger, K., & Yeazitzis, T. (2023). Conceptualizing a Socio-technical Model for Evaluating AI-driven Technology. N Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 1639–1644.

Wilson, H. J., Daugherty, P. R., & Davenport, C. (2019). The future of AI will be about less data, not more. Harvard Business Review, 14(1).

Wright, S., & Schultz, A. (2018). The rising tide of artificial intelligence and business automation: developing an ethical framework. Business Horizons, 61, 823–832.

Wu, Y. (2022). An Overview Analysis of AI Divide: Applications and Prospects of AI Divide in China's Society. 3rd International Conference on Mental Health, Education and Human Development, 3–9.

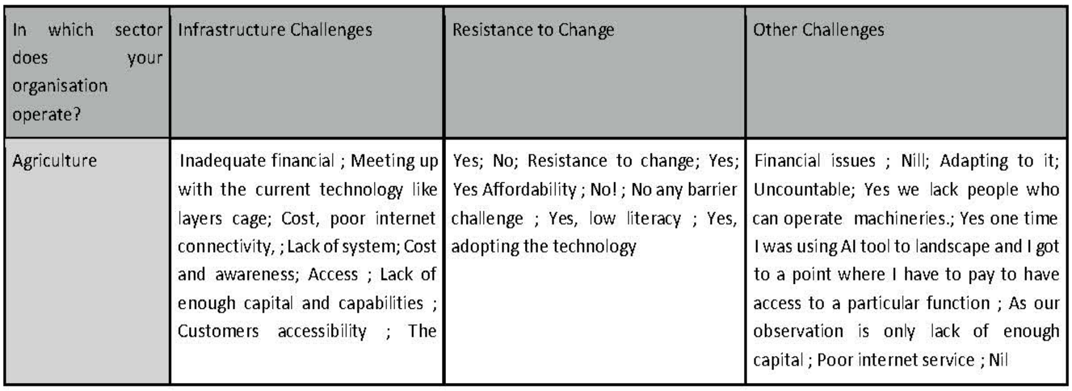
Yu, J., Peng, F., Shi, X., & Yang, L. (2022). Impact of credit guarantee on firm performance: Evidence from China's SMEs. Economic Analysis and Policy, 75, 624–636. https://doi.org/10.1016/J.EAP.2022.06.017

Yu, P. K. (n.d.). THE ALGORITHMIC DIVIDE AND EQUALITY IN THE AGE OF ARTIFICIAL INTELLIGENCE. https://www.nitrd.

Yu, X., Xu, S., & Ashton, M. (2023). Antecedents and outcomes of artificial intelligence adoption and application in the workplace: the socio-technical system theory perspective. Information Technology and People, 36(1), 454–474. https://doi.org/10.1108/ITP-04-2021-0254

Appendix

Survey Data (Challenges)



In which sector does your organisation operate?	Infrastructure Challenges	Resistance to Change	Other Challenges
Agriculture	Inadequate financial ; Meeting up with the current technology like layers cage; Cost, poor internet connectivity, ; Lack of system; Cost and awareness; Access ; Lack of enough capital and capabilities ; Customers accessibility ; The	Yes; No; Resistance to change; Yes; Yes Affordability ; No! ; No any barrier challenge ; Yes, low literacy ; Yes, adopting the technology	Financial issues ; Nill; Adapting to it; Uncountable; Yes we lack people who can operate machineries.; Yes one time I was using AI tool to landscape and I got to a point where I have to pay to have access to a particular function ; As our observation is only lack of enough capital ; Poor internet service ; Nil

	ability to access the right skills to implement the technology		
Banking	Electricity & computer	No	Manpower & energy
Construction	Prices ; Nil	No; Nil	No; Nil
Fashion	None; None	None; Not really	None; They were a bit challenging as it is a totally different face for us
Food	The staff are not familiar with most technologies; We are not using it so we feel it's not important ; Have Neva implement it Unskilled personnel ; My device. Is one of the most challenge as the device I use can be termed obsolete hence some features are not available	No; Yes; No Computer grammar ; No	Getting the staff to understand how it works is our major challenge ; I'm not the reading type ; Have Neva tried it I don't understand most of the new technological languages ; Sometimes the features provided looks a bit away from reality thereby rising suspicion among those who are supposed to buy from me
Health	Network problems; Power supply ; Structures are needed. Building ; Power supply and other structures ; Power supply ; Power supply ; Power supply ; Power supply ; Power supply ; Power supply ; Power supply	No; Unintended consequences ; Inequality and access ; Bias and discrimination ; Lack of human touch ; Data quality and availability ; Inequality access ; Public trust and acceptance ; Unintended consequences ; Cyber security risks ; Lack of human touch ; Unintended consequences	Lot of errors and inability to adress them; Regulatory challenges ; Regulatory challenges ; Job displacement ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Ethical concerns ; Regulatory challenges ; Regulatory challenges
IT	Funds; Cost : they are usually somewhat expensive ; It is hard teaching the staff with less tech exposure; Learning curve of the technology; Nil; Cost of internet access; 1. Outdated hardware or software: Incompatible or obsolete systems can hinder new technology integration. 2. Insufficient bandwidth or network capacity: Inadequate network infrastructure can lead to slow performance, downtime, or in-house failed implementations. manage new 3. Limited storage capacity: Inadequate data storage can impede data-intensive technologies or lead to data loss.; Network ; Finance ; Internet connectivity ; Not applicable; Some of few challenges faced are; 1. At time you can't get accurate data s expected because is not everything one can rely on in terms of Ai	Sure; Not really ; No; Non; Nil; No; Yes technology know-how ; Mostly new development to my staff; The AI need be easily accessible to the extent that everyone can use; No; Not applicable; Yes, users should provide feedback and important update and large input to improve and enhance the use of Ai in our businesses and daily life; Installation barriers and trainers ; Yes; No; Yes especially is its migration to a new system my staff are not familiar with. ; No; Training & retraining ; It wil bring more customers ; Uncover hidden pattern ; Time Knowledge gap and need for sensitisation ; No; No Costs and deciding the write tools and features	Adaptability ; Not really ; Getting the staff who are less familiar with tech to adopt new tech; Acceptability.; Nil; None; 1. Integration issues: Difficulty integrating new tools with existing systems or infrastructure. 2. Data quality and availability: Insufficient or poor-quality data to train or implement AI models. 3. Lack of expertise: Limited knowledge or skill to effectively consuming; Yes; implement and technologies. ; Mostly New an difficult to adopt; Not at all ; Cost; Not applicable; Yes, there are challenges in using and applying new technology, it however, challenge the means and use new ideas to achieve good aims and objectives, people find it difficult to migrate from old technology to new as many are not ready to adopt changes fast.; Adequate use of resources ; Training staff on the usage of the new technologies and change; It was not easy for the staffs to learn ; The use of outlook instead of gmail was difficult

	2. People misused Ai to cheat with creating fake data, animation, graphics and designs to achieve a common goal .; Availability of resources ; Usage; Nothing much ; User resistance by some staff; Lack of proper knowledge of operation ; Internet failure ; Power supply ; Power supply ; High cost of data subscription; Having access to network ; Bad network/commectivity Flexible of the app User friendly ; Expertise issues; None Knowing what to automate and what not to.		for the staff to adjust but they eventually accepted. ; Yes we lack proper knowledge of using it ; Internet accessibility ; Lack of skill and training ; Lack of skill ; No; Inadequate training and access to network; Cost of the AI solution, user acceptability ; The unavailability of experts ; Inconsistent results and inaccurate answers We need to always keep improving and scaling the solution. And even automated features need monitoring by a human.
Manufacturing	Electricity ; None really ; Lack of good network; Network ; Unreliable electricity distribution, lack of skilled personnel that can operate the machine ; Illiteracy ; Electric power supply, skill and accessibility ; Insufficient power supply and lack of accessibility ; Insufficient power supply ; Insufficient power supply, availability of the new technology and fund; Insufficient power supply ; Power supply ; Power supply ; Power supply and Capital ; Power supply ; Electric power supply ; ChatGPT; Power supply ; Power supply ; Power supply and access roads to neighboring villages	No; Yes, some employees frown at technological change. ; Commitment; We are Operating most of the technological things in an overloaded network ; Lack of experience personnel ; Provide a machine that can double the work of the current equipment ; Lack of skill personnel ; Insufficient power supply ; If we have vast knowledge of the new technology, then we can adopt it; Yes, illiteracy ; There is no skill personnel in this organization that can operate any AI technology ; Yes, work fastened, output increased ; Lack of skill personnel ; Substitute human being ; Substitute human being with machine ; Yes; Yes; No; No; Delayance	In adequate knowledge of operate tools to the workers ; None; Understanding; We cannot properly operate new machines and most of the equivalent used in my organization are outdated ; Access to the new technology is the major problem we are facing here; Skill; Is not accessible and even if you can access the new technology, but it won't be affordable to a small business owners; We cannot access the technology, meaning is not accessible ; I cannot access the new technology ; The new technology is inaccessible ; The AI is not accessible at all; Yes, knowledge ; Operation problem ; Don't know the impact of the new technology ; No, I did not see any challenge; It can't use with electric power and is not reliable ; Lack of knowledge ; My employees lack skills ; Lack of skill and training ; Spoilage and damages
Retail	Access to Internet, power supply, cyber crime and fraud.; Reliable internet network ; Lack of capital ; No; Job displacement ; Lack public trust and acceptance; Bias and discrimination ; Power supply ; Lack of human touch ; Job displacement ; Lack of human touch ; Bias and discrimination ; Bias and discrimination ; Dependence on technology ; Cyber security risks; Power supply ; Protect us against the harm of cyber security ; Power supply ; Power supply ; Power supply ; Power supply ; Power supply ; Power supply ; Power supply ; Stable electric power	No; None; Educating the staffs ; No; Depending on technology ; Ethical concerns ; Job displacement ; Lack of human touch ; Inequality and access ; Ethical concerns ; Unintended consequences ; Lack of transparency and accountability ; Public trust and acceptance ; Privacy concerns ; Explainability and interpretability ; Public trust and acceptance ; Existential risks ; Unintended consequences ; Public trust and acceptance ; Inequality access ; Regulatory challenges ; Public trust and acceptance ; Inequality and access ; Lack of transparency and accountability ; Dependence on technology ; Bias and discrimination ;	Training ; Hacking ; Difficult to understand ; No; Regulatory challenges ; Cyber security risk; Lack of transparency and accountability ; Inequality and access; Existence of risk ; Dependence on technology ; Existence of risk ; Dependence on technology ; Regulatory challenges ; Data quality and availability ; Regulatory challenges ; Regulatory challenges ; Inequality access ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Cyber security risks ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Cyber security risks ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory

	supply ; Stable power supply ; Power supply ; Security ; Security ; Security ; Safety ; Safety and power supply ; Power supply ; Power supply ; Power supply ; Power supply ; Power supply ; Power supply ; Power supply ; Power supply ; Power supply	Cyber security risks ; Existential risks ; Lack of human touch ; Job placement ; Data equality and availability ; Public trust and acceptance ; Bias and discrimination ; Job displacement ; It can perpetuate and amplify existing biases if train on bias data; Explainability and interpretability ; Job displacement ; Dependence on technology ; Cyber security risks ; Explainability and interpretability ; Job displacement	challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges ; Regulatory challenges

Familiarity with AI

In which sector does your organisation operate?	Sector	Confidence Level	Skill Level	Familiarity with AI	Positive Confidence (%)	Positive Skills (%)	Positive Familiarity (%)
Agriculture	Agriculture	Agree; Disagree; Agree; Neutral; Strongly Disagree; Agree; Neutral; Neutral; Neutral; Agree	Moderate; Moderate; Moderate; Moderate; Moderate; Moderate; Moderate; Very Low; Moderate; High	Yes; No; Yes; Partially; Yes; Yes; Yes; No; Partially; Yes	40%	10%	60%
Banking	Banking	Strongly Agree; Agree	Moderate; Moderate	Yes; No	100%	0%	50%
Construction	Construction	Agree; Neutral	High; Moderate	Partially; No	50%	50%	0%
Education	Education	Agree	Moderate	Yes	50%	0%	50%
Fashion	Fashion	Agree; Agree; Strongly Agree	Moderate; High; Moderate	Partially; Yes; Yes	100%	33%	67%
Food	Food	Neutral; Neutral; Strongly Agree; Neutral	Low; Low; Moderate; Moderate	Partially; Partially; Yes; No	25%	0%	25%
Food	Food	Neutral; Neutral; Strongly Agree	Moderate; Moderate; Moderate	Partially; Partially; No	33%	0%	0%
Health	Health	Agree; Neutral; Neutral; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree	High; Moderate; Very Low; Low; High; Moderate; Moderate; Moderate; Moderate; Moderate; Moderate; Low; Moderate; Moderate	Yes; No; Partially; Partially; Partially; Partially; Partially; Partially; Partially; Partially; Partially; Partially; Partially; Partially	86%	14%	7%
IT	IT	Strongly Disagree; Agree; Strongly Agree; Agree; Neutral; Agree; Agree; Agree; Strongly Agree; Agree; Neutral; Agree; Strongly Agree; Agree; Agree; Agree; Strongly Agree; Neutral; Neutral;	High; Very High; Very High; Very High; Moderate; High; High; Very High; Moderate; High; Moderate; Moderate; High; High; High; Moderate; High; Moderate; High; Moderate; High; High; High; Moderate; Very High; Moderate;	Yes; Yes; Yes; Yes; Partially; Yes; Yes; Partially; No; Yes; No; Yes; Yes; Yes; No; No; No; Yes; Yes; Partially; Yes; Partially; Yes; Partially; Partially; Yes; No; Yes;	69%	67%	61%

		Neutral; Agree; Neutral; Agree; Neutral; Strongly Agree; Agree; Neutral; Strongly Agree; Agree; Neutral; Agree; Neutral; Agree; Agree; Strongly Agree; Agree	Moderate; High; High; High; Very High; Moderate; Moderate; Very High; Very High; Very High	Yes; Yes; Yes; Partially; No; Yes; Yes; Yes			
Manufacturing	Manufacturing	Agree; Agree; Strongly Agree; Strongly Agree; Agree; Neutral; Neutral; Neutral; Neutral; Neutral; Disagree; Agree; Strongly Agree; Agree; Neutral; Agree; Agree; Neutral; Agree; Neutral; Agree	Low; High; Very High; Moderate; Moderate; Low; Low; Low; Low; Moderate; Low; High; High; High; Moderate; High; High; Moderate; Moderate; Moderate; High	Yes; Partially; Yes; No; Yes; No; No; No; No; No; No; Yes; Yes; Yes; Yes; Yes; Yes; Partially; Yes; No; Yes	57%	38%	52%
Mining	Mining	Agree	Moderate	No	100%	0%	0%
Retail	Retail	Agree; Agree; Agree; Agree; Strongly Disagree; Neutral; Neutral; Neutral; Agree; Agree; Neutral; Neutral; Neutral; Neutral; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Neutral; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Agree; Neutral; Neutral; Agree; Agree	High; Moderate; Moderate; High; Moderate; Moderate; Moderate; Moderate; High; Moderate; Moderate; Moderate; Moderate; Low; Moderate; Moderate; Low; Low; Low; Low; Low; Moderate; Low; Moderate; Low; Moderate; Low; Moderate; Low; Very Low; Moderate; Moderate; Low; Low; Low; Low; Moderate; Moderate; Low; Low; Low; Low; Low	Yes; Partially; No; Partially; Yes; Partially; Partially; Partially; Yes; Partially	74%	7%	7%

Public and private actors in the surveillant assemblage – a dangerous relation

Marie Eneman (University of Gothenburg) and Jan Ljungberg (University of Gothenburg)

Completed Research

Abstract (around 150 words)

Police authorities have high expectations on surveillance technology to increase efficiency and security. Since the technology often is provided by private actors, a set of tensions and dilemmas arise. Public actors are driven by their official authority assignment, while private actors are driven by commercial interests. Drawing on a current example where police officers used facial recognition technology as part of their official work, this paper explores tensions that arise at the intersection between public and private actors in an surveillant assemblage and what consequences they entail. We introduce assemblage theorizing as an approach to explore the increasingly blurred boundaries between public and private actors in complex, socio-technical arrangements. By identifying the tensions, the paper contributes with insights into various dilemmas linked to accountability, control, legitimacy, privacy, power and transparency that can arise in the regulatory context in connection with the availability and use of facial recognition.

Keywords: Surveillance, facial recognition, assemblage theory, public, private

1.0 Introduction

> *"There are thus at least three significant actors in this drama, government agencies, private corporations and, albeit unwittingly, ordinary users."*
> (Lyon, 2014, p.3)

Recent technological advancements extend the scope of ance, becoming increasingly powerful, ubiquitous and embedded in our daily lives (Lyon and Murakami Wood, 2021). In addition, the most recent developments in algorithms and artificial intelligence advances the analytical capabilities in surveillance further (Kitchin, 2022; Kosta, 2022). These new digital potentials are on the one hand associated with expectations of increased security in society, but also causing concern of serious threats to civil liberties such as individuals' privacy since large volumes of personal and sensitive data easily can be generated, processed and combined both within and between different public and private systems (Ball and Snider, 2019; Richards, 2022).

Balancing the opportunities of digital surveillance for security purposes while at the same time protecting privacy generates a set of tensions and dilemmas.

The focus of this paper is state based surveillance practices where a central actor is the police authority, using a range of surveillance technologies such as CCTV, body-worn cameras, secret data interception, drones, facial, motion and object recognition, digital fingerprints, automated number plate recognition, internet of things and a variety of sensors etc in different contexts (reference will be inserted next round). These surveillance technologies are not to be understood as individual applications, but rather as a system - a convergence of discrete surveillance technologies molded by an intricate interplay between institutional forces that condition how they are materialized in certain surveillance practices (Haggerty and Ericson, 2000). Here, private actors play a crucial role as they often develop and provide a major part of digital devices and software applications used in the emerging government digital surveillance practices (Flyverbom, 2019; Richards, 2022). The relation between public and private actors becomes problematic since the police are governed by their authority assignment to increase security in society while private actors adhere to a market logic governed by commercial interests (Ball and Snider, 2019.

Balancing the possibilities of digital surveillance for security purposes while protecting individuals' privacy creates a set of tensions and dilemmas since the government digital surveillance practices mobilize values and logics from both the public and private sectors. Hence, a situation where government practices are increasingly entangled with the commercial private sector, raises important questions (Dencik et al., 2022). For example, in relation to the design of technologies and algorithms - considering the unintended negative consequences they may bring in relation to the collected data (Glaser et al., 2021). This refers to an intricate, but not arbitrary, arrangement of discursive and non-discursive entities. In this paper, we call such an arrangement an 'assemblage'. The assemblage can be thought of as a particular constellation, a sociotechnical system as it were, where “[t]here is always a material, bodily formation and a discursive, semiotic formation operating together” (Thompson et al., 2022, p. 692). Assemblage theorizing can prove beneficial in explaining a complex, messy, sociotechnical reality. A reality where ‘things’ are uncertain and where boundaries between discrete categories (e.g., public and private) are fuzzy and blurred.

Drawing on a current example, highlighting the complexities that can arise when police authorities use powerful surveillance technology, here facial recognition, provided by a private commercial actor, we introduce assemblage theorizing as an approach to explore the increasingly blurring boundaries between public and private actors in complex, sociotechnical arrangements. The following question will guide the paper: *What tensions arise at the intersection between public and private actors in an emerging surveillant assemblage and what consequences does this entail?*

The paper continues as follows, we begin by briefly presenting the research context of sance and privacy in relation to the public and private actors becoming increasingly entangled with each other. Thereafter, we introduce the concept of assemblage, followed by an example of facial recognition. Finally, we discuss various tensions and conclude the paper.

2.0 Surveillance

The notion of surveillance is broad and difficult to theorize, including everything from practices, processes, and different technologies (Lyon, 2010; 2018). However, a common denominator is that the term refers to the particular interest in watching human behavior through systematic and routine attention to personal information (Lyon, 2007; Solove, 2021). Studies drawing on the concept of panopticon (Foucault, 2020), provide perhaps the most well-known examples on how researchers today investigate this type of deliberate surveillance. Research based on the idea of the panopticon has also shown how the state has for a long time had an interest in monitoring its population linked to the development of the bureaucratic state with records of citizens (Boersma et al, 2014). More recent examples refer to crises and extraordinary events in society that have led to rapid and far-reaching changes in law that have enabled the introduction of powerful surveillance technologies e.g. 9/11 (Lyon, 2015) during the pandemic (Kitchin, 2020) and in relation to serious organised crimes (reference will be inserted in the next round). Lyon (2007) argues that surveillance should be understood as any collection and processing of personal data, whether identifiable or not, for the purposes of influencing or managing those whose data has been collected.

Although surveillance is not a new phenomenon, recent studies (Kitchin, 2022; Lyon and Murakami Wood, 2021) often point to how the increased digitalization of today's society has profoundly altered the surveillance capabilities. In particular, developments in artificial intelligence (AI) and machine learning are said to lay the ground for surveillance capabilities of a magnitude we have not seen before (Kosta, 2022), exposing individuals to data-driven surveillance raising the costs related to privacy (Nissenbaum, 2010, 2015; Véliz, 2020). These new technologies are more concealed, pervasive and automated, enabling many processes and tasks to be performed at the same time, but also facilitating large-scale collection and storage of data, as well as making it possible for data to rapidly flow within and between different systems (Ball and Snider, 2019; Richards, 2022). By leaving individuals in situations where they may not be aware of when they are being exposed to surveillance, this development is associated with serious threats to democratic rights such as privacy. Hence, these new digital potentials are associated with expectations of increased security in society, but research (Lyon and Murakami Wood, 2021; Richards, 2022) highlights that the they also are associated with concerns about risks to individuals' privacy since large volumes of personal data are easily collected without individuals knowing the extent of what is collected and how it can be used and shared between different systems and/or commodified, known as surveillance capitalism (Zuboff, 2019; 2022).

Researchers (Ball et al, 2015; Richards, 2022) have underscored that today's surveillance practice involves private actors as part of public security, urging us to investigate the complex relationship between public and private actors to understand the emergence of today's different modes of surveillance and its implications for individuals' privacy. The term privacy has often been defined as the right to control information about oneself (Hildebrandt, 2020; Solove, 2021). Nissenbaum (2010; 2015) argues however that privacy concerns should not be limited solely to concern about control over personal information since what people react to when they complain and protest that privacy has been violated is not the act of sharing information itself, but the inappropriate, improper sharing of information.

Furthermore, she points out that privacy is constantly renegotiated, context dependent and takes on different meanings in different contexts, which was illustrated in a recent study (reference will be inserted in the next round) where they investigated how individuals perceive privacy in relation to different types of surveillance technologies used by various public and private actors. In addition, studies have pointed to the increased legal requirements for handling such large amounts of information about individuals' behaviour and personal data (Black and Murray, 2019; Murray, 2021). To ensure individuals' privacy and to protect individual's right to control personal information about themselves a number of legislations and statutes have also been introduced, both on national level and on EU-level (see e.g. European Commission, 2021).

3.0 Assemblage theory

The relationship between the social and the material is a ceaseless area of concern within information systems research (Cecez-Kecmanovic et al., 2014, Orlikowski and Scott, 2008; 2023). One could even argue that the concept 'sociotechnical' is one of the central tenets of the discipline as such (Sarker et al., 2019). But how to grasp and make sense of it? In this paper, we turn to the concept of 'assemblage' (Deleuze and Guattari, 1988; Nail, 2017; Buchanan 2021) as a promising candidate. In the field of information systems (IS), the term assemblage has occasionally appeared and become quite popular as a way of attacking and explaining the sociomaterial conundrum. Often, however, the use of the term assemblage does not denote a theoretical concept per se, but refers more to an everyday understanding and explanation of how things are gathered. But more recently, IS research has begun to use and explore the concept's more established theoretical and philosophical perspectives (e.g., Patel, Baiyere and Johnsen, 2022; Hanseth and Modol, 2021).

The concept of assemblage does not refer to a stable entity with fixed boundaries but as something emergent, thus depicting an open system prone to change rather than a static and closed one (Buchanan, 2015; Nail, 2017). But what exactly then is an assemblage? If you start from a more everyday definition, you might understand it as a number of (material) things coming together to form a larger whole. However, although the assemblage consists of a variety of heterogeneous objects, what really makes it an ‘assemblage’ is the fact that they function as a unit – that their unity derives from them “working” together (Haggerty and Ericson, 2000). Consequently, assemblages can be seen as relational multiplicities arranged to fulfil certain purposes and needs. This suggests that there is something more to the concept than simply explaining how things converge. In their development of the concept of surveillant assemblage, Haggerty and Ericson draw on Deleuze and Guattari (1988) and their notion of 'assemblage'. And for Deleuze and Guattari, the notion of ‘desire’ is central to the whole project (Buchanan, 2021). As Haggerty and Ericson (2000, p. 609) puts it: “Deleuze and Guattari approach desire as an active, positive force that exists only in determinate systems.” Thus, notions such as ‘purpose’ and ‘function’ also becomes important: “Assemblages have in common the fact that they are all arrangements of desire, but this does not mean that desire is the same in every assemblage, nor does it mean that all assemblages arrange desire in the same way or that they all have the same components” (Buchanan, 2021, p. 79).

The concept can thus be used to explain and understand a complex, messy, material reality, but one would not utilise its full potential if not also account for the discursive dimension – that there is a purpose (desire) for why an assemblage is arranged in a specific way (Deleuze and Guattari, 1988; Nail, 2107). Hence, "[e]ven if it is the material dimension of the assemblage which piques our interest, as it frequently is, analysis should not begin there because the material does not disclose its meaning, function - its mattering - by itself" (Buchanan, 2021, p. 73). Starting from this perspective on what an 'assemblage' is, it is therefore not enough to simply claim that an assemblage is, for example, a collection of databases, servers, algorithms, cables, switches, routers, engineers, managers, and users. Of course, such an enumeration would give us a feeling, and help building an intuition, what each component could contribute with - but it does not really say anything about the overall purpose of the arrangement. In order to be able to account for 'function' and 'desire', the need for a discursive dimension of the assemblage - an orchestrating force, as it were - becomes obvious (Buchanan, 2021).

3.1. Theorising surveillance as assemblage

Although AI is largely claimed to be one of the most disruptive technologies in a foreseeable future, not the least in the field of surveillance (Fenwick et al., 2017:

Kitchin, 2022), its development is molded by different institutional forces conditioning the enactment of the potentials that the technology affords (Gustavsson, 2023). For instance, if the technology affords potentials enabling new businesses to prosper, the maintenance of law, regulations and policy making may still end up having problems to keep up with the use of the technology (Black and Murray, 2019; Murray, 2021). This situation is well suited to be interpreted from an assemblage perspective as "[i]t combines the two basic elements of the assemblage: the machinic (technological) and the expressive (socio)" (Buchanan, 2021, p. 140). In a sense, a surveillance assemblage can be thought of as simultaneously a machine (relations of material artefacts) that connects, collects, identifies, categorizes, and an autonomous discourse (a discursive system) that specifies norms, rules, purposes, and goals which have implications for how the assemblage is (re)arranged and utilized. In the following sections, we describe the material and the discursive perspectives of a surveillant assemblage.

3.2. The materiality of surveillance

Technologies such as AI and machine learning have made surveillance systems more powerful, subtler, and large-scale (Lyon and Murakami Wood, 2021: Zuboff, 2019; 2022). Some researchers argue that the new approaches of surveillance require more effective and formal measures to protect civil liberties such as privacy (Richards, 2022; Véliz, 2020). Contemporary surveillance consists of a multiplicity of data sources, algorithms, and hardware. Notably, a vast amount of data is generated from digital services and devices initially designed for other purposes than surveillance, but which could be fed into surveillance processes nevertheless. Additionally, data collection could be performed by specific devices such as cameras, biometrics, sensors, GPS etc., either to be analyzed in batches, in real time or used to train machine learning algorithms for later automatic monitoring (Smyth, 2019). Machine learning (ML) algorithms excel in pattern recognition by sifting through huge amounts of data. They thus provide a range of capabilities of interest to surveillance, such as prioritization, recommendation, categorization, prediction, and translation (Brayne, 2021). But these algorithms do not work in isolation (Kitchin, 2022), they operate in a wider ecology of platforms, databases, applications and physical devices such as cameras and sensors (mounted on, for example, buildings, vehicles or people), emphasizing that algorithms need to be understood in its specific context with the conditions shaping how they are developed and used (Kitchin, 2017).

The materialist perspective on the surveillance assemblage thus concerns artifacts and their relations. Artifacts of both physical (e.g., CCTV, body-worn cameras, drones) and digital (e.g., facial, motion and object recognition, a variety of sensors) origins. One aspect of digital artifacts worth noticing is their ambivalent ontology, pointing to the incompleteness of digital technology and the constant evolution of algorithms as social agents Airoldi, 2022; Glaser et al., 2021; Kallinikos et al., 2013.

3.3. The institutional conditions for surveillance

When public authorities and private companies interplay in surveillance practices, their actions reflect different norms and practices (Ball and Snider, 2019; Czarniawska, 2014). Consequently, various actors can imagine different outcomes of technology use. The utilization of the technology may thus be driven by different desires. In that sense, surveillance can be seen as a technology's potential, a means, but there may be different ends. For instance, companies may apply surveillance as a means for monitoring and deepening their understanding of the market and different consumer behavior (Zuboff, 2019; 2022), while the police authority introduces and uses surveillance to prevent and investigate crime and maintain security in society governed by their official assignment. In both cases, the use of surveillance technologies raises privacy concerns, forcing actors to consider the importance of society recognizing them as trustworthy actors.

The police as an authority with a unique sanctioned monopoly of violence must thus continuously prove their legitimacy (Königs, 2022; Rolandsson, 2020; Tankebe, 2013; Tyler, 2011) as a prerequisite for public trust for their exercise of authority which include the use of surveillance technology as part of their work methods. Legitimacy refers to how individuals, despite potential concerns, maintain a willingness to trust in - or be vulnerable to - corporations or public authorities and how they monitor behaviors (Suddaby et al., 2017; Thornton et al., 2012). The notion of legitimacy here underlines the crucial connection between organizations and the societal recognition of them as trustworthy actors serving or supporting various public domains and thus common goods (Tyler, 2011), which is of particular importance for the police authorities as legitimate holder of the monopoly of violence in democratic societies (Tankebe, 2013).

The institutional perspective on the surveillance assemblage thus concerns the discursive work conducted around norms, rules and regulations. A work that in turn privileges certain actors, technologies, purposes, values and actions over others. The material (re)arrangement of the assemblage, that is, which technologies are deemed proper and how they are to be used, is thus stipulated through the discursive work.

4.0 Clearview AI

We have used Clearview AI's facial recognition application as an illustrative example of a surveillant assemblage consisting of a complex arrangement of actors, technologies and other elements. In January 2020, the New York Times published an internationally acclaimed article, "The secretive company that might end privacy as we know it", which revealed Clearview's controversial business model (Hill, 2020; Devany, 2022). Clearview's facial recognition application goes far beyond traditional facial recognition technologies (McSorley, 2021; Rezende, 2020). The company uses an automated image scraper to scrape facial images from the open Internet (Devany, 2022), not least from social media platforms such as Facebook, Instagram and Twitter. The images are used to build a giant biometric database that currently contains more than three billion images. The business model is based on marketing and selling access to its database to law enforcement agencies and private security companies worldwide. offering law enforcement officers, a 30-day free trial.

When Clearview's customer list was leaked (BuzzFeed News, 2020), it was revealed that law enforcement agencies in the United States, Canada and several European countries had used the application. Typically, this use was performed by individual officers without permission from any authorities and without the knowledge or control from the management level. Police officers expressed that despite their limited knowledge of how the company behind the application and how it was created and operates, they have chosen to use it as part of crime investigation work that include shoplifting, identity theft, credit card fraud, murder, serious organized crime and child sexual exploitation. The leaked list also showed that several authoritarian regimes were customers of Clearview, which has further contributed to the debate.

When it was revealed in 2020 that (name of the country will be inserted in the next round) police officers were using Clearview's facial recognition application, a formal investigation was initiated (referenced will be inserted in the next round), that deemed the use of Clearview AI as illegal.

We will insert a description here - of the specific case from a country, where the police used Clearview AI as part of their authority work – in the next round if accepted. We removed the section now for the review.

5.0 Discussion

> *"Since the concrete elements are always changing with their conditioning relations the assemblage is always becoming capable of different things. This requires a constantly renewed analysis of assemblages"*
> (Nail, 2017, p. 26-27).

The traditional view of technology as an exogenous factor is challenged as digital technology becomes intertwined with organizations in fundamentally new ways (Faraj and Pachidi, 2021; Faraj and Leonardi, 2022). Now digital technology is not just a means to streamline and support processes of various kinds, but increasingly constitutes the very fabric of organizations. However, as digital capabilities are increasingly leased from external actors (e.g., cloud services, Software-as-a-Service) (Narayan, 2022), the idea of a siloed computing experiences that occur solely within an organization's boundary can lead to an overly simplistic explanation of the digital realm. Instead, these experiences are probably better understood as assembled ones. As information systems become increasingly difficult to think of as discrete and well-bounded artefacts (Amoore, 2020; Kallinikos et al., 2013), a relational approach may prove beneficial (Faraj and Leonardi, 2022). An approach that stimulates processual thinking and how relationships between entities occur and develop.

Following that line of thinking, one could, for example, regard the boundary of an organization as provisional and therefore perpetually in the making. Here, assemblage thinking may open new vistas as it challenges static positions that assume boundaries and relations as fixed, to instead perceive them as tentative and thus continuously (re)created over time. For example (and as illustrated in our example), as artificial intelligence emerges as a fundamental and pervasive phenomenon, studies of surveillance systems may require a focus on how various types of technologies become intertwined with organizations, societies, and individuals over time.

In the information systems literature, digital technologies are often depicted as (re)configurable and (re)programmable – properties that in turn open for combinatorial innovation to take place (e.g., Henfridsson et al., 2018; Yoo, 2013; Yoo et al., 2010; 2012). These characteristics lend themselves particularly well to be analysed through an assemblage lens as they speak of a large, complex space of possibilities (Benbya et al., 2020; Pentland et al., 2020) where digital, socio-technical assemblages could be arranged. Although these assemblages often are materially realized, they are discursively orchestrated and maintained where the discursive work concerns how a sociotechnical assemblage is imagined, pieced together, and further developed. Hence, assemblage theorizing may offer a way to navigate the conundrum between technological determinism and social constructivism "by grounding its analysis in an ontology that suspends these categories in favor of an understanding of the dynamic evolutionary systems that cut across them" (Bousquet, 2014, p. 91).

Assemblage theorizing also challenges us to think about time in various ways. On the one hand, one can think of an assemblage as something that occurs at a specific point in time and thus can be analyzed as a momentary 'event'. But in a wider analytical perspective, it can also be thought of as a continuously ongoing, socio-technical arrangement (Nail, 2017). Depending on analytical perspective and ontological stance, assemblage theorizing can thus offer interesting and complementary alternatives to the information

5.1. Tensions in the interplay between public and private actors

As surveillance technology is rapidly introduced into police authorities around the world – technology often provided by private actors, and motivated by expectations of increased societal security (Brayne, 2021; Kosta, 2022) – there are voices pointing to far-reaching risks to democratic rights such as privacy (Véliz, 2020). In a broader institutional perspective, this raises concerns about the legitimacy of police authorities when they entangle with private commercial actors. In this paper, we presented an example that illustrates the complex interplay between public and private actors in government surveillance practices. More specifically, a police authority's unauthorized use of a readily available powerful surveillance technology - a facial recognition application, provided by a private actor - with the potential result that data could flow between the state and the private actor's systems. Thus, how the capacity of potent technology, paired with the desire for more effective policing, can trigger use even when not sanctioned by the authority.

Our example indicates that government surveillance practices are increasingly entangled with the commercial private sector. This situation gives rise to various dilemmas as these state practices now mobilize values and logics from both the public and private sectors (Ball and Snider, 2019), where one of the more pressing issue being the potential risk that data collected on citizens flows between public and private sectors (Richards, 2022). We can recognize the root of this predicament in the different desires of the actors. While the government is guided by an aim to increase security in society, private companies are guided by a market logic; consequently, private companies drive the development of new sophisticated technologies that in turn are marketed to potential customers, such as governments (Lyon and Murakami Wood, 2021).

The dilemmas that surfaced in our example, and where assemblage theorizing may prove beneficial as an analytical perspective, are as follows: information sharing where data, which may be personal and also sensitive, flows between public and private systems (Richards, 2022). Lack of control, transparency and the possibility of accountability (Dencik et al., 2022), as a result of private commercial actors developing and selling surveillance technologies being guided by commercial interests and thus often prioritizing profit interests over ethical considerations (Stahl et al., 2023).

The case also spurs reflecting on further issues, such as: (1) How a market logic can potentially lead to a lack of responsibility (Stahl, 2012) and transparency around the design, implementation and use of these technologies with the consequence that authorities do not get full insight into how technology/algorithms work (Glaser et al., 2021) or the potential biases and limitations that may exist (Airoldi, 2022). (2) The rapid pace of technological advances often outpaces (Feenwick, 2017) the development of legal and regulatory frameworks to adequately govern their use (Black and Murray, 2019) which can create a legal vacuum where the use of such technologies by law enforcement authorities lacks clear legal support, guidelines and control (Murray, 2021). (3) The risk of purpose slippage once the technology is implemented. (4) Surveillance technologies can be powerful tools in the 'right' hands but there is also a tangible risk for misuse. When private companies provide these technologies, there is a risk that they may be used beyond their intended purpose or used in a manner that infringes upon individuals' rights to privacy and freedom of expression (Ball and Webster, 2019). The boundaries between legitimate law enforcement activities and unwarranted intrusion can become blurred further eroding police legitimacy (Königs, 2022; Tyler, 2011).

Thus, to fully grasp the role emerging digital technologies play in a contemporary era, they should be understood not only as discrete tools for practicing certain methods, but also as components of assemblages; assemblages governed by norms and logics, and which may span discrete categories such as private/public, connecting a growing number of devices and services that provide and exchange data (Flyverbom, 2019; Ball and Webster, 2019). This paper's example highlights the potential institutional complexity of a surveillant assemblage (Haggerty and Ericson, 2000; Buchanan, 2021), as state and private actors (Ball et al., 2015), through a digital infrastructure, converge into an arrangement - an assemblage. An assemblage characterized by an intricate interplay between different institutionalized missions, logics, and practices that shape and determine the affordances of technology (Airoldi, 2022). This perspective acknowledges and problematizes the tension around the legitimate boundaries between private and public control of personal data and the algorithms that use it (Kitchin, 2022; Kosta, 2022). A tension that may place further demands on transparency, legitimacy and accountability (Königs, 2022).

5.2. Future intelligent assemblages

> *"Investigating machine learning systems as social agents culturally entangled with humans in the context of platformised fields may appear to be no more than a niche yet fashionable research direction. However, it has increasingly become a necessity, since very few realms of the social world remain untouched by the ubiquitous application of these information technologies"* (Airoldi, 2022, p.149).

Artificial intelligence increasingly affects societies, organisations, and individuals. As these technologies are distributed by external actors (Narayan, 2022; Ferrari, 2023) this may have implications for well-established concepts such as 'organization', 'strategy' and 'the boundary of the firm' (Faraj and Leonardi, 2022) – what they mean and, fundamentally, how we understand them. If we imagine a situation where organizations rent 'intelligent' capabilities from external actors, we can approach it in at least two ways with regards to assemblage theorizing:

(1) On the one hand, we can analyze the situation from the perspective of one organization (as in the example in this paper). Now we can imagine how an external technology is added as an element to an already existing 'organization assemblage'. In a sense, the territory of the assemblage has not expanded, the desire is still the same, but the ability to deliver on that desire may have increased due to technological sophistication. This perspective would focus on how the technology could/would help the organization to fulfil its desire. But also, and which is becoming more and more relevant in the age of AI, how issues around data protection and privacy emerge and are discussed. Here the discursive dimension of the assemblage becomes visible as the technology may show great promise and potential in its way of functioning but can nevertheless be questioned on other grounds. That is to say, even if the assemblage is driven by a desire to fulfil certain tasks, there are norms and rules that prescribe how the assemblage is to be arranged – that is, what is legally feasible and not.

(2) But we could also analyze the situation from the private actor perspective. Here, the focus would be on how the 'private actor assemblage' expands its territory as organizations (or maybe more correctly, 'customers') rents/uses its technology. That is, we can imagine how the territory of the private actor assemblage is widening as

more and more 'customers' make use of its services. This analysis could, for example, focus on how the technology is developed and maintained, but also how the private actor discursively conveys itself as a legitimate and trustworthy actor. If the assemblage’s territory expands to become large and widespread, and where customers depend on the private actor for their functioning, it can ultimately become a question of assembled power where the concepts of ‘public’ and ‘private’ blur and are up for negotiation (Liebetrau and Monsees, 2023).

These two approaches to assemblage theorizing showcase that while the discussion of how capabilities of a specific technology (such as AI) affect organizations’ possibility to fulfil their desires is critical to investigate (perspective 1), it is increasingly important to also study from where these capabilities are delivered (perspective 2). As more and more of the computing power used by organizations is delivered by external actors (Narayan, 2022), it becomes important to understand and be able to analyze the intricate socio-technical web of relationships between public and private organizations as it may ultimately be about relations of power (Van Dijck, 2021). In this potential power play, one can sense the essential nature of digital platforms as central resources for computation of various kinds. They thus provide computational, infrastructural opportunities for organizations to build on. Opportunities that the organizations themselves often may not be able to develop and maintain due to lack of knowledge and/or capital. This emerging computational, almost symbiotic-like, relationship between private and public actors may become more relevant than ever when considering the use and distribution of artificial intelligence as a technology (e.g., Ferrari, 2023). Just think of the epitome of today's artificial intelligence, the large language model, and the potential impact it could have. Here we can imagine a situation where platforms, through the 'intelligent' building blocks they provide, become fundamental elements of countless assemblages (with various, yet countless desires).

5..0 Conclusions

In this paper, we explored how various tensions arise when state surveillance practices are increasingly entangled in the market logic of the commercial private sector. To be able to get to and consequently reflect on these tensions we turned to the concept of assemblage. Three things lead us to believe that assemblage theorizing is well suited to analyze this scenario: (1) It emphasizes the sociotechnical aspect of a phenomenon and thus reflects both the material and the discursive dimension. (2) It focuses on purposeful relations - how ‘something’ is arranged by someone to fulfil a need, a desire - and can thus help reveal how a phenomenon is assembled utilizing artefacts from various domains (e.g., public/private). (3) It can be a valuable approach when researching complex, fuzzy phenomena where it is not totally clear who governs what and where boundaries are blurred.

Emerging surveillant assemblages are shaped by an intricate interplay of institutional forces that condition how technologies' affordances are materialized, which also involve private actors as the private sector develops and provides the main part of digital devices and software applications to be used as part of law enforcement's surveillance practices. As this paper clearly shows, emerging digital surveillance technologies should no longer be understood as individual tools related to certain methods, but as a convergence of previously discrete surveillance technologies into a digital surveillance assemblage that mobilizes values and logics from both the public and private sectors and where data flows between the sectors.

The results show that the public/private interplay in state surveillance practices gives rise to various tensions which shows the importance that authorities (such as the police here), legislators and policy-makers develop organizational and technological measures to ensure that surveillance practices are organized in line with government logics where important democratic values and rights such as privacy and freedom of expression are protected. By identifying these tensions, the paper also contributes with insights into various legitimacy issues that may arise in connection with public and private actors’ interplay in emerging state surveillance practices. Furthermore, the assemblages with public/private partnership also pose major challenges in terms of governance and control, since the boundaries between the public and the private, the technical and the political are unclear. Which means that they are not under the full control of any single state actor, which in turn raises important questions linked to legal certainty. Finally, sociotechnical phenomena such as surveillant assemblages are likely to evolve and become even more complex, ubiquitous and powerful in the future and we argue that assemblage theory is a useful perspective to unpack and better understand its complex arrangement of actors, technologies and institutional conditions.

References

Airoldi, M. (2022). *Machine Habitus: Toward a Sociology of Algorithms*. Polity Press.

Amoore, L. (2014). "Security and the claim to privacy". *International Political Sociology*, 8(1), 108-112.

Amoore, L. (2020). *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham: Duke University Press.

Ball and Snider (2019). *The Surveillance-Industrial Complex: A Political economy of surveillance*, Routledge.

Ball, K., and Webster, W. (2019). *Surveillance and Democracy in Europe*. Routledge.

Ball, K., Canhoto, A., Daniel, E., Dibb, S., Meadows, M and Spiller, K. (2015). *The Private Security State? Surveillance, Consumer Data and the War on Terror*. CBS Press.

Benbya, H., Nan, N., Tanriverdi, H. and Yoo, Y. (2020). "Complexity and Information Systems Research in the Emerging Digital World." *MIS Quarterly, 44*(1), 1-17.

Black, J and Murray, A (2019). "Regulating AI and Machine Learning: Setting the Regulatory Agenda." *European Journal of Law and Technology*, 10(3).

Boersma, K., van Brakel, R., Fonio, C and Wagenaar, P (2014). *Histories of State Surveillance in Europe and Beyond*, Routledge.

Bousquet, A. (2014). "Welcome to the Machine: Rethinking Technology and Society through Assemblage Theory." In Acuto, M., and Curtis, S. (Eds.), *Reassembling International Theory*. London: Palgrave Pivot.

Brayne, Sarah (2021). *Predict and Surveil, Data, Discretion and the Future of Policing*. Oxford: Oxford University Press.

Buchanan, I. (2015). "Assemblage Theory and Its Discontents." *Deleuze Studies, 9*(3), 382-392.

Buchanan, I. (2021). *Assemblage theory and method*. Bloomsbury Academic.

Buzzfeed News (2020). *Police In At Least 24 Countries Have Used Clearview AI. Find Out Which Ones Here*, Retrieved on October 2, 2023.

Cecez-Kecmanovic, D., Gailiers, R. D., Henfridsson, O., Newell, S., and Vidgren, R. (2014). "The Sociomateriality of Information Systems: Current Status, Future Directions." *MIS Quarterly, 38*(3), 809-830.

Czarniawska, B. (2014) *A theory of organizing*, Edward Elgar Publishing Ltd.

Deleuze, G., and Guattari, F. (1988). *A thousand plateaus: capitalism and schizophrenia*. London: The Athlone Press.

Dencik, L., Hintz, A., Redden, J., and Treré, E. (2022). *Data Justice*. London: SAGE Publications Ltd.

Devany, B. E. (2022). Clearview AI's First Amendment: A Dangerous Reality? *Texas Law Review* 101.2. 473-507.

European Commission (2021). *Proposal for a regulation of the European Parliament and of the council, Laying down harmonised rules on artificial intelligence and amending certain union legislative acts*, COM (2021) 206 final.

Faraj, S., and Leonardi, P. M. (2022). "Strategic organization in the digital age: rethinking the concept of technology." *Strategic Organization, 20*(4), 771-785.

Faraj, S., and Pachidi, S. (2021). "Beyond Uberization: The co-constitution of technology and organizing." *Organization Theory, 2*(1), 1-14.

Fenwick, M. D., Kaal, W. A., and Vermeulen, E. P.M. (2017). "Regulation Tomorrow: What Happens When Technology Is Faster than the Law?", *American University Business Law Review*, 6(3).

Ferrari, F. (2023). "Neural production networks: AI's infrastructural geographies." *Environment and Planning F: Philosophy, Theory, Models, Methods and Practice*.

Flyverbom, M. (2019). *The Digital Prism*, Cambridge University Press.

Foucault, M. (2020). *Discipline and Punish*. Penguin Classics.

Glaser, V. L., Pollock, N., & D'Adderio, L. (2021). "The Biography of an Algorithm: Performing algorithmic technologies in organizations." *Organization Theory*, *2*(2).

Gustavsson, M. (2023). *Platformization: Digital Materiality at the Limits of the Discource*. Doctoral Thesis, University of Gothenburg.

Haggerty, K. D., and Ericson, R. V. (2000) "The surveillant assemblage." *British Journal of Sociology*, 51(4).

Hanseth, O., and Modol, J. R. (2021). "The Dynamics of Architecture-governance Configurations: As Assemblage Theory Approach." *Journal of the Association for Information Systems, 22*(1), 130-155.

Henfridsson, O., Nandhakumar, J., Scarbrough, H., and Panourgias, N. S. (2018). "Recombination in the Open-Ended Value Landscape of Digital Innovation." *Information and Organization, 28*(2), 89-100.

Hildebrandt, M. (2020). *Law for computer scientists and other folk.* Oxford University Press.

Hill, K. (2020). *Your Face Belongs to Us*, SimonSchuster Ltd.

Kitchin, R. (2017). "Thinking critically about and researching algorithms". *Information, Communication & Society*, 20: (1).

Kitchin, R. (2020). "Civil Liberties or Public Health, or Civil Liberties and Public Health? Using Surveillance Technologies to Tackle the Spread of COVID-19". *Space & Polity* 24.3. 362-81.

Kitchin, R (2022) *The Data Revolution: A Critical Analysis of Big Data, Open Data & Data Infrastructure*s. SAGE Publications Ltd.

Kosta, E. (2022). "Algorithmic state surveillance: Challenging the notion of agency in human rights." *Regulation & Governance*, 16.

Königs, P. (2022). "Government Surveillance, Privacy, and Legitimacy." *Philos. Technol. 35*(8).

Liebetrau, T., and Monsees, L. (2023). "Assembling Publics: Microsoft, Cybersecurity, and Public-Private Relations." *Politics and Governance, 11*(3), 157-167.

Lyon, D. (2007). *Surveillance Studies: An Overview*. Polity Press.

Lyon, D. (2014). "Surveillance, Snowden, and Big Data: Capacities, consequences, critique." *Big Data & Society, 1*(2).

Lyon, D. (2018). *The culture of surveillance*, Polity Press.

Lyon, D., and Murakami Wood, D. (2021). *Big Data Surveillance and Security Intelligence, The Canadian case*. Vancouver: UBC Press.

McSorley, T. ((2021). "The Case for a Ban on Facial Recognition in Canada". *Surveillance & Society*, 19: (2), 250–254.

Murray, A. (2021). *Almost Human: Law and Human Agency in the Time of Artificial Intelligence.* T.M.C. Asser Press.

Nail, T. (2017). “What is an assemblage?” *SubStance, 46*(1), 21-37.

Narayan, D. (2022). “Platform capitalism and cloud infrastructure: Theorizing a hyper-scalable computing regime.” *Environment and Planning A: Economy and Space, 54*(5), 911-929.

Nissenbaum, H. (2010). *Privacy in context: Technology, Policy, and the Integrity of Social Life*, Stanford University Press.

Nissenbaum, H. (2015). “Respecting context to protect privacy: Why meaning matters.” *Science and Engineering Ethics,* 1–22.

Orlikowski, W., & Scott, S. (2008). “Sociomateriality: Challenging the Separation of Technology, Work and Organization”. *The Academy of Management Annals*, 1(2), 433–474.

Orlikowski, Wanda J., and Susan V. Scott. (2023) "The Digital Undertow and Institutional Displacement: A Sociomaterial Approach." *Organization Theory* 4(2).

Patel, S., Baiyere, A and Johnsen, C. (2022) "Beyond Practice: Assemblage thinking in sociomateriality research", *Proceedings of the European Conference on Information Systems*.

Pentland, B.. T., Mahringer, C. A., Dittrich, K., Feldman, M. S., and Wolf, J. R. (2020). “Process Multiplicity and Process Dynamics: Weaving the Space of Possible Paths.” *Organization Theory, 1*(3), 1-21.

Plesner, U. (2019). *Digital Organizing*, Red Globe Press.

Privacy International. (2022). *Challenging Public Private Surveillance Partnerships: A Handbook for Civil Society.*

Raviola, E. and Norbäck, M. (2013). “Bringing technology and meaning into institutional work: Making news at an italian business newspaper”. *Organization Studies*, 34. 1171–1194.

Raviola, E., and Dubini, P. (2016). “The logic of practice in the practice of logics: practicing journalism and its relationship with business in times of technological changes.” *Journal of Cultural Economy*, *9*(2), 197-213.

Rezende, I. (2020). Facial Recognition in police hands: Assessing the ‘Clearview case’ from a European perspective. in *New Journal of European Criminal Law*, 11: (3), 375–389.

Richards, N. (2022). *Why Privacy Matters*, Oxford University Press Inc.

Rolandsson, B. ((2020). The emergence of connected discretion: Social media and discretionary awareness in the Swedish police. Qualitative Research in Organizations and Management, 15: (3), 370–387.

Sarker, S., Chatterjee, S., Xiao, X, and Elbanna, A. (2019). “The Sociotechnical Axis of Cohesion for the IS Discipline: Its Historical Legacy and its Continued Relevance.” *MIS Quarterly, 43*(3), 695-719.

Solove, D. (2021). *The Myth of the Privacy Paradox*, GWU Law School Public Law Research Paper No. 2020-10.

Smyth, S (2019). *Biometrics, Surveillance and the Law: Societies of Restricted Access, Disciplines and Control*. Oxon, UK/ New York: Routledge.

Stahl, B. C. (2012). “Responsible research and innovation in information systems.” *European Journal of Information Systems*, 21(3), 207–211.

Stahl, B. C., Schroeder, D., and Rowena, R. (2023). *Ethics of Artificial Intelligence*. Cham: Springer Nature.

Suddaby, R., Bitektine, A., and Haack, P. (2017). “Legitimacy.” *ANNALS,* 11, 451–478.

Thompson, G., Sellar, S., and Buchanan, I. (2022). "1996: the OECD policy-making assemblage." *Journal of Education Policy, 37*(5), 685-704.

Thornton, P. H., Ocasio, W., and Lounsbury, M. (2012). *The institutional logics perspective: A new approach to culture, structure, and process*. Oxford: Oxford University Press.

Tyler, T. (2011). "Trust and legitimacy: Policing in the USA and Europe." *European Journal of Criminology*, 8(4).
Van Dijck, J. (2021). "Seeing the forest for the trees: Visualizing platformization and its governance." *New Media & Society, 23*(9), 2801-2819.

Véliz, C. (2020). *Privacy is power*, Bantam Press.

Yoo, Y. (2013). "The tables have turned: How can the information systems field contribute to technology and innovation management research?" *Journal of the Association for Information Systems, 14*(5). 227-236.

Yoo, Y., Henfridsson, O., and Lyytinen, K. (2010). "Research commentary – The new organizing logic of digital innovation: An agenda for information systems research." *Information Systems Research, 21*(4), 724-735.

Yoo, Y., Boland Jr, R. J., Lyytinen, K., and Majchrzak, A. (2012). "Organizing for innovation in the digitized world." *Organization Science, 23*(5), 1398-1408.

Zuboff, S. (2022). "Surveillance Capitalism or Democracy? The Death Match of Institutional Orders and the Politics of Knowledge in Our Information Civilization." *Organization Theory*, 3(3).

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power.* Profile Books Ltd.

Usability and Acceptance of mHealth Self-Testing Tools in a University Environment

Colm Kingdon
University College Cork, 123324776@umail.ucc.ie

Clo O'Riordan
University College Cork, CORiordan@ucc.ie

Maura Smiddy
University College Cork, m.smiddy@ucc.ie

Michael Byrne
University College Cork, M.Byrne@ucc.ie

Ciara Heavin
University College Cork, c.heavin@ucc.ie

John MacSharry
University College Cork, J.MacSharry@ucc.ie

Completed Research

Abstract

Mobile Health (mHealth) has the potential to transform healthcare allowing for rapid diagnosis, care, and public health surveillance. Our research investigates the usability and acceptability of an mHealth application, called UniHealth, used in parallel with a study-supplied multiplex antigen test for self-testing at home for symptomatic or asymptomatic respiratory infections (SARS-CoV-2, Influenza A and B, and RSV). We conducted an anonymous online survey informed by validated questions using a 5-point-Likert derived from Unified Theory of Acceptance and Use of Technology (UTAUT)). The survey conducted with university students and staff who participated in the UniHealth study. After data cleaning, we analysed 62 completed records. Most participants surveyed strongly agreed/agreed that the UniHealth app and the supplied multiplex self-tests were easy to use. Our analysis confirmed the applicability of UTAUT in the context of mHealth and antigen self-testing in the home. The findings of this study will benefit software designers and developers, antigen test designers and manufacturers, policy makers, and healthcare providers.

Keywords: mHealth, antigen test, self-testing, point of care testing, COVID-19, usability

1.0 Introduction

Since the first wave of the COVID-19 pandemic, the potential role of mobile wireless technology for public health, commonly referred to as mHealth (Je, 2020), has gained widespread attention. During the pandemic, many countries worldwide turned to digital solutions for information management, monitoring, and access to healthcare

(Asadzadeh and Kalankesh, 2021). These digital tools played a crucial role in reducing the impact of outbreaks and enhancing patient outcomes (Getachew et al., 2023). Post pandemic, the spread of respiratory viruses continues to be a challenge in terms of absenteeism in proximity settings, such as universities (Cue et al., 2022). This is important because the absence of students and staff in person negatively affects the student experience, from academic understanding and social skills development to all other aspects of university life. By promoting safe, reliable self-testing and reporting, universities are enabling a better learning experience for students. They are also creating a safer working environment for staff, particularly those who are vulnerable or live with vulnerable people.

Here, we explore a specific system implemented by UniHealth—an interdisciplinary research team consisting of academics, clinicians, and researchers—to manage and monitor COVID-19, Influenza, and RSV outbreaks on university campuses in Ireland. Our goal is to evaluate the effectiveness of the UniHealth app, suggest improvements, and consider future implications. By tapping into the expertise of these professionals, we can refine and adapt systems to better align with the needs of those who use them. Hence, we pose the question:

> *"How usable and acceptable are mHealth self-testing tools in a university setting in the post-COVID-19 context?"*

This paper is organised as follows; the next section considers the use of mHealth to support self-testing and the use of these innovations in higher education settings. Subsequently, the usability and acceptability of innovative technology is presented. From this, the research approach and background to the study is outlined. The results of our anonymous survey are presented. The limitations of the study are considered and the implications for research and practice are outlined. Finally, opportunities for future research are discussed.

2.0 Theoretical Background

The COVID-19 pandemic has accelerated the adoption of digital health tools, making understanding technology acceptance and usability more critical than ever. The pandemic highlighted the importance of contactless, remote, and accessible health solutions, leading to an increased use of mHealth platforms for self-testing and symptom monitoring. Post-COVID, technology acceptance models have attracted renewed attention in other disciplines, as patients and healthcare systems transition back to normalcy. Studies suggest that factors like perceived usefulness and trust in technology have become more prominent, as individuals are now more accustomed to using digital health platforms (Andrews et al., 2021).

2.1 mHealth for Self-Testing

Self-testing has gained traction due to its convenience and ability to empower individuals in managing their health (Pittman et al., 2023) while also keeping others safe. Several types of self-testing solutions have emerged, including COVID-19 self-tests (Budd et al., 2023; Undelikwo et al., 2023), sexual health tests (Pearson et al., 2018), and glucose monitoring systems (Cervantes-Torres and Romero-Blanco, 2023). These tools have become increasingly integrated with mHealth platforms, which provide users with guidance, data analysis, and health information. COVID-19 self-tests are a prime example of mHealth's capacity to facilitate widespread testing, with apps guiding users through test procedures and reporting results (Smith et al., 2023). In chronic disease management, devices like continuous glucose monitors link with mobile apps, offering real-time feedback and historical data tracking (Cervantes-Torres and Romero-Blanco, 2023).

mHealth platforms have become central to self-testing by offering tools that simplify testing procedures and data interpretation (Ko et al., 2020). Applications often include features like reminders, alerts, and education modules to enhance user engagement and adherence. One major advantage of mHealth in self-testing is the convenience it offers—users can access testing materials and support at home without the need to visit a healthcare facility (Aicken et al., 2016). However, challenges remain, including ensuring the accuracy of data input by users and managing data privacy concerns (Hall et al., 2024).

2.2 Innovations in mHealth for Higher Education

Higher education institutions have increasingly adopted mHealth tools for self-testing, particularly in response to COVID-19. Some universities implemented app-based platforms that facilitated daily health check-ins, symptom tracking, and COVID-19 test result submissions (Simhan et al., 2020). These platforms aimed to create safer campus environments by encouraging responsible behaviour and monitoring health trends. Studies have assessed the usability of mHealth tools among university students and staff (Nicolaidou et al., 2022; Peprah et al., 2019). These studies often focus on aspects like the ease of navigation, clarity of instructions, and user satisfaction. Findings suggest that younger, tech-savvy students are more likely to embrace mHealth tools (Blebil et al., 2023), while staff members may experience greater challenges due to variations in digital literacy. Improving user interface design and providing training can enhance usability for all demographics within the higher education setting (Aburas and Ayran, 2013).

Acceptance of mHealth tools for self-testing in universities is influenced by a range of factors, including privacy concerns, trust in technology, cost, and accessibility. For example, concerns about data privacy can deter students from using self-testing apps while trust in the app's accuracy can drive higher usage rates (Aicken et al., 2016). Cost is another consideration, especially if the burden falls on students or staff to purchase necessary devices or pay for app subscriptions (Aicken et al., 2016). Demographics such as age, digital literacy, and socioeconomic background significantly impact the usability and acceptability of mHealth tools in a university context (Kim & Lee, 2022). Younger users often find digital tools easier to navigate, while older or less tech-savvy individuals may require additional support (Aranha et al., 2024). Addressing these disparities is crucial for equitable technology adoption.

2.3 Acceptance and Usability of mHealth

User acceptance of new Information Systems and Technologies (IS/IT) is a widely explored topic in current research, leading to the development of several theoretical models that explain individuals' intentions to adopt these innovations. The Technology Acceptance Model (TAM), introduced by Davis (1989), is one of the most influential frameworks. TAM posits that two primary factors influence technology adoption: Perceived Usefulness, the extent to which a person believes that using a technology will enhance their performance and Perceived Ease of Use, the degree to which a person believes that using a technology will be free of effort (Davis, 1989). These factors directly impact the intention to use technology, which predicts actual use. TAM has been modified and expanded to include other factors, such as user trust and perceived enjoyment, to better predict technology acceptance in health settings. Another widely used model is the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al., 2003). This model integrates elements from multiple theories, including TAM, and identifies four core determinants of technology acceptance: Performance Expectancy, Effort Expectancy, Social Influence, and Facilitating Conditions (Venkatesh et al., 2003). Performance expectancy is an individual's belief that technology will facilitate daily activities (Venkatesh et al., 2003). Existing research in health technology suggests that the main factor that affects a person's willingness to use the technology in the long term is performance expectancy (Alam et al., 2020).

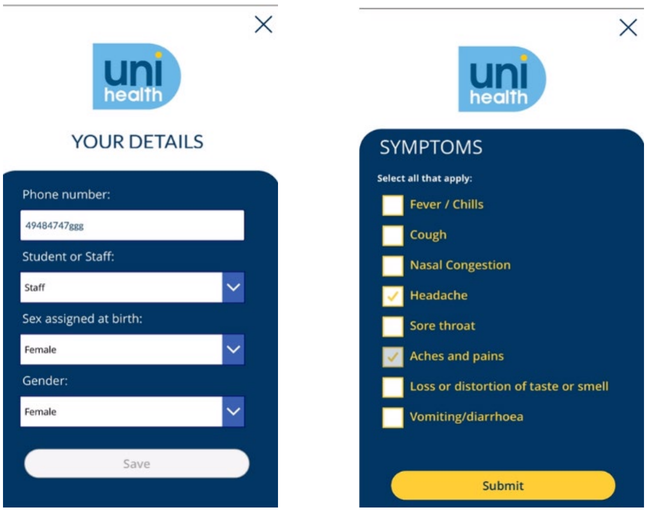
Venkatesh et al. (2003) defined effort expectancy as the ease of using technology. Usability plays a crucial role in the success of mHealth applications (Zapata et al., 2015). In the context of mHealth, usability encompasses factors such as ease of use, user interface design, and perceived ease of use (Birkmeyer et al., 2021). Studies have shown that intuitive and straightforward interfaces contribute to higher adoption rates, especially in populations with varying degrees of digital literacy (Zhou et al., 2019). Ease of use is often linked to how accessible the information and features are, with minimal training or technical skills needed (Karahanna and Straub, 1999) A well-designed interface that simplifies complex health data can make self-testing more manageable and engaging. Usability often focuses on identifying user pain points and making adjustments that enhance the overall user experience. The rapid shift to digital health has altered user expectations and comfort levels, necessitating new research to capture these trends in the context of self-testing.

Social Influence is defined as a person's perception that most people who are important to them think that they should perform the behaviour in question, i.e. use the technology (Venkatesh et al., 2003). Further, Chan et al. (2010) conceptualised social influence as "the interpersonal considerations" of technology adoption and use. Facilitating conditioning is an individual's belief that their infrastructure it possesses supports technology adoption (Venkatesh, 2003). The adoption of health technology such as mHealth applications would be readily accepted by its users if users have adequate support facilities such as smartphones, including internet support (Alam et al., 2020). Behavioural intention is defined as a measure of the strength of one's intention to perform a specific behaviour (Fishbein & Ajzen, 1975). It captures a user's intention to adopt or continue using a technology typically based on their expectations of its benefits, ease of use, social influence, and other factors (Wu and Chen, 2017).

While these theories date back several years and were not originally formulated to be used in the context of remote healthcare technologies, UTAUT has been applied in various health technology studies (Rouidi et al., 2022). Thus, highlighting the importance of social and contextual factors in technology adoption decisions (Van der Waal., 2022). It provides a comprehensive view of how expectations of technology performance and ease of use, alongside social support and infrastructure, shape user behaviour in health settings.

3.0 Research Approach

The UniHealth interdisciplinary project adopted a holistic approach to identify, diagnose, and treat infectious diseases, particularly respiratory infections. By leveraging innovative digital technology, the project aimed to build resilience, reduce the risk of outbreaks, and decrease absenteeism among university staff and students. With the aim of promoting safe, reliable self-testing and reporting in close proximity settings such as universities, we refined a mHealth app (which was developed and rolled out during COVID-19), namely the UniHealth app (Figure 1), used in parallel with a study-supplied multiplex antigen test for self-testing at home for symptomatic or asymptomatic respiratory infections (SARS-CoV-2, Influenza A and B, Adenovirus, and RSV).



YOUR DETAILS

Phone number: 49484747ggg

Student or Staff: Staff

Sex assigned at birth: Female

Gender: Female

SYMPTOMS

Select all that apply:

□ Fever / Chills
□ Cough
□ Nasal Congestion
☑ Headache
□ Sore throat
☑ Aches and pains
□ Loss or distortion of taste or smell
□ Vomiting/diarrhoea

Figure 1. UniHealth App (Screenshots)

The wider study ran from February – May 2024 involving university staff members and students. During this period, we received 181 sign-ups including 107 staff, 36 students, 38 did not disclose. UniHealth study packs were distributed to staff and students via several easily accessible on-campus locations and delivered directly by internal post to staff offices. In terms of the UniHealth app (MS Power App) registrations, we recorded 143 completed records on the "My Details" section of the app. UniHealth app users generated 406 unique app entries. In total we recorded 116 unique users, 77% staff and 23% students.

Following approval from the Social Research Ethics Committee at University College Cork, we conducted an anonymous online survey using Qualtrics with university students and staff from March to April 2024. Our aim was to investigate the usability and acceptability of the UniHealth app and the study-supplied multiplex antigen test. The survey questions were informed by validated questions and adapted from UTAUT (Venkatesh et al., 2003). MS Excel was primarily used for data cleaning and analysis.

4.0 Results

4.1. Study Characteristics

An overview of the study characteristics is presented in Table 1. Out of a total of 78 survey participants, there were 61 complete responses with 1 partially complete and participants ranged from ages 18 to 65+.

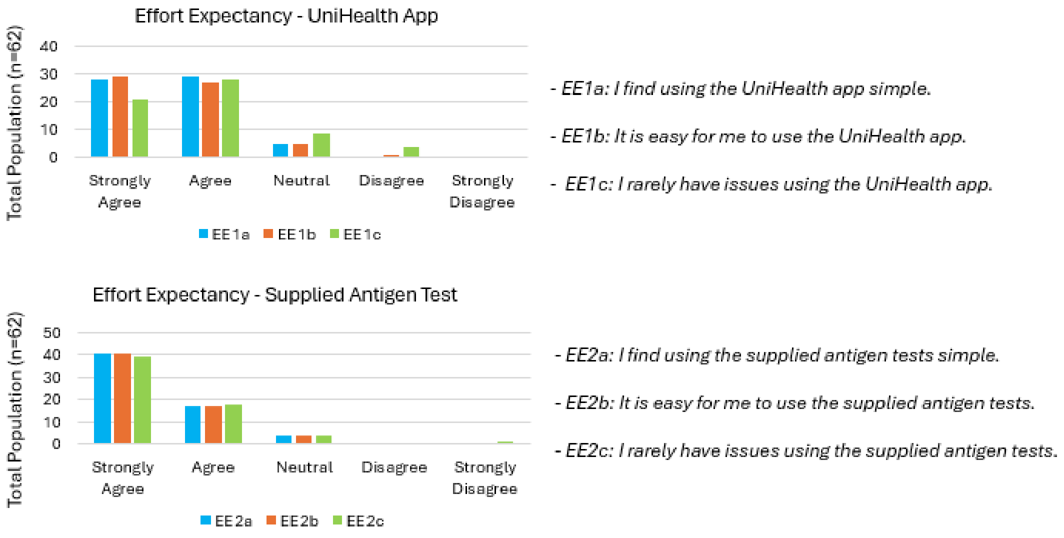
VARIABLE (POPULATION/ CHARACTERISTICS)	GROUPS/ CATEGORIES	NUMBER	PERCENTAGE (%)
Age	18 - 24	4	6.5
	25 - 34	5	8
	35 - 44	14	22.6
	45 - 54	20	32.3
	55 - 64	18	29
	65+	1	1.6
		62	
Role	Staff Member	50	82
	Postgraduate Students	8	13
	Undergraduate Students	3	5
		61	

Table 1. Study characteristics

Participants were comprised of three distinct categories of university members: staff members, undergraduate and postgraduate students. The survey participants were mostly between the ages of 35 - 64 years old (35 – 44 = 22.6%, 45 – 54 = 32.3% and 55 – 64 = 29% on total participants) with staff members making up 82% of total participants. In this study, our initial data analysis involved the use of descriptive statistics.

4.2 Results

Participants were asked about Effort Expectancy of the UniHealth app and the study supplied antigen test. Participants strongly agree that antigen tests were simple, easy to use and mostly without issues, see Figure 2.



- EE1a: I find using the UniHealth app simple.

- EE1b: It is easy for me to use the UniHealth app.

- EE1c: I rarely have issues using the UniHealth app.

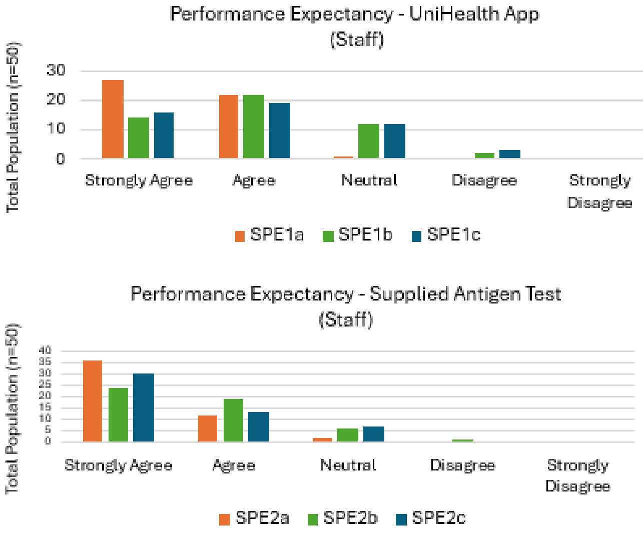
- EE2a: I find using the supplied antigen tests simple.

- EE2b: It is easy for me to use the supplied antigen tests.

- EE2c: I rarely have issues using the supplied antigen tests.

Figure 2. **Effort Expectancy of UniHealth App and Study Supplied Multiplex Test.**

Investigating further, descriptive analysis revealed that the Effort Expectancy of the UniHealth app and the study supplied antigen test among the university staff cohort (n=50) was positive with participants indicating that multiplex antigen tests were simple and easy to use (Figure 3).



-SPE1a: Using the UniHealth app enables me to submit my symptoms and self-test results in a meaningful and accessible way.

- SPE1b: Using the UniHealth app would positively impact my health and wellbeing.

- SPE1c: Using the UniHealth app would give me greater control over my understanding of my symptoms.

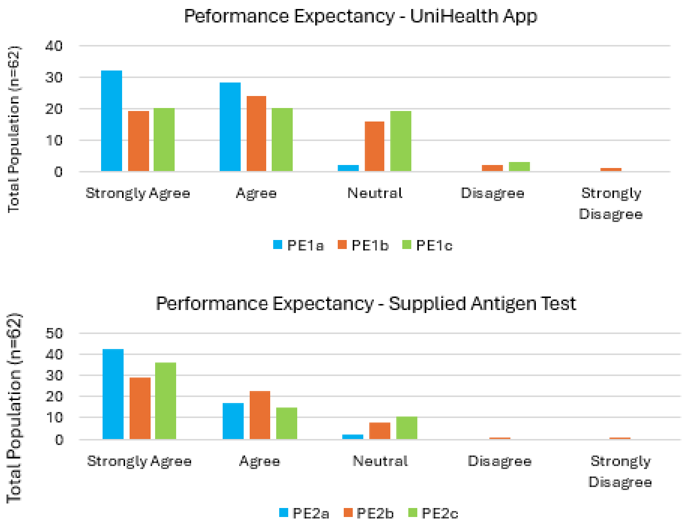
- SPE2a: Using the supplied antigen test enables me to self-test in a meaningful and accessible way.

- SPE2b: Using the supplied antigen test would positively impact my health and wellbeing.

- SPE2c: Using the UniHealth antigen test would give me greater control over my understanding of my symptoms.

Figure 3. Effort Expectancy of UniHealth App and supplied multiplex tests (Staff only).

In Figure 4, we reveal that the Performance Expectancy of both the app and the supplied antigen tests was deemed postive by respondents. Notably, some participants strongly disagreed/disagreed or were neutral about the statement *"Using the supplied antigen test would positively impact my health and wellbeing."*



- PE1a: Using the UniHealth app enables me to submit my symptoms and self-test results in a meaningful and accessible way.

- PE1b: Using the UniHealth app would positively impact my health and wellbeing.

- PE1c: Using the UniHealth app would give me greater control over my understanding of my symptoms.

-PE2a: Using the supplied antigen test enables me to self-test in a meaningful and accessible way.

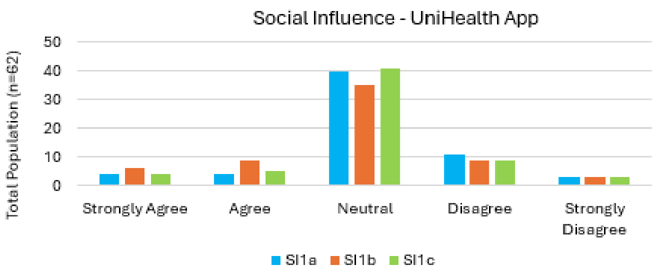
-PE2b: Using the supplied antigen test would positively impact my health and wellbeing.

-PE2c: Using the UniHealth antigen test would give me greater control over my understanding of my symptoms.

Figure 4. Performance Expectancy of UniHealth App and Study Supplied Antigen Test.

In terms of Faciliating Conditions to support particpants use of the UniHealth app and study supplied antigen test, the feedback was postive. The majority of survey participants strongly agreed/agreed with the statements referring to faciliating conditions. Results reveal that 98% strongly agreed/agreed with the statement *"The* *university has been helpful in encouraging the use of the UniHealth app and supplied antigen test"*, 97% strongly agreed/agreed with the statements *"I have the knowledge necessary to use the UniHealth app"*, *"I have the knowledge necessary to use the supplied antigen test"*, and *"Instructions about using the UniHealth app and supplied antigen test are available to me"*. While 89% and 88% of participants respectively strongly agreed/agreed with the statements *"The university has been helpful in encouraging the use of the UniHealth app and supplied antigen test"* and *"A specific person (or group) is available for assistance with difficulties associated with accessing and using the UniHealth app and supplied antigen test"*.

Results indicated that study participants were broadly positive about the UniHealth app and antigen tests. However, post pandemic it seems that Social Influence plays a lesser role in influencing behaviour as illustrated in Figure 5. Our survey revealed some neutrality and disagreement amongst participants regarding the social perceptions of an individual's use of mHealth and antigen self-tests, for example 19% strongly disagreeing/disagreeing and 66% neutral in their response to the statement *"People who are in my social circle think that I should use the UniHealth app"*.



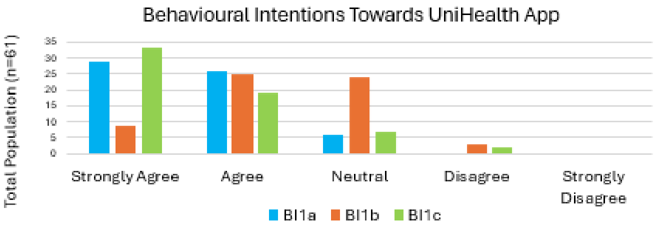
- SI1a: People who influence my behaviour would think that I should use the UniHealth app.

- SI2b: People who are important to me would think that I should use the UniHealth app.

- SI2c: People who are in my social circle would think that I should use the UniHealth app.

Figure 5. Social Influence and use of the UniHealth App.

Behavioural intention towards the use of the UniHealth App in the future highlighted mixed opinion amongst participants (Figure 6). With uncertainty emerging around the future use of the UniHealth app in universities.



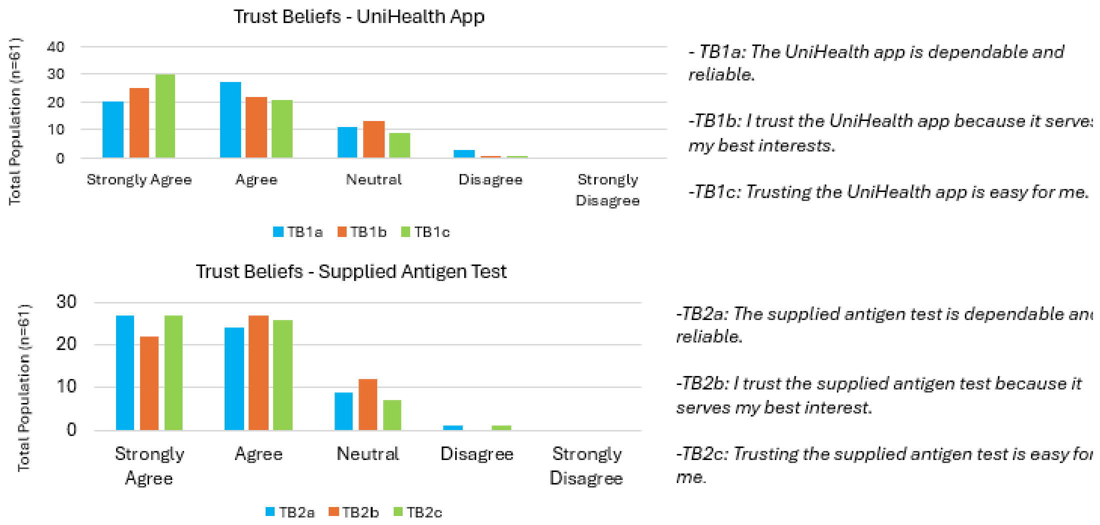
-BI1a: I will use the UniHealth app for my health and wellbeing needs.

-BI1b: I predict many universities will use the UniHealth app in the future.

-BI1c: I will recommend the UniHealth app to others.

Figure 6. Behavioural Intentions towards the UniHealth App.

In Figure 7, results showed that participants broadly trusted the UniHealth App and study supplied antigen test.



- TB1a: The UniHealth app is dependable and reliable.

-TB1b: I trust the UniHealth app because it serve my best interests.

-TB1c: Trusting the UniHealth app is easy for me.

-TB2a: The supplied antigen test is dependable an reliable.

-TB2b: I trust the supplied antigen test because it serves my best interest.

-TB2c: Trusting the supplied antigen test is easy fo me.

Figure 7. Trusting Beliefs towards the UniHealth App and Supplied antigen test.

Overall, participants expressed confidence in their ability to self-report their symptoms through the UniHealth app, supporting them to take responsibility for their own health. Our findings indicate that users' perceptions of ease of use and reliability of the study-supplied antigen tests contributed to a positive overall experience.

5.0 Discussion

As part of the wider UniHealth study, our aim was to better understand the usability and acceptance of a mHealth app and study supplied multiplex antigen tests to support self-testing at home for respiratory illnesses including COVID-19, Influenza A and B, and RSV. Our descriptive analysis supports the usability and acceptability of both UniHealth app and study supplied antigen tests among a cohort of university staff and students. However, evidence supports the need for additional work on the design of the UniHealth app. During this study we refined a MS Power app which was originally designed and evaluated during the COVID-19 pandemic. We deployed version 3.0 of the UniHealth app as part of this study. While the results are mostly positive, it seems that some participants had issues using the app. Further investigation and additional participant feedback revealed that Power Apps do not perform consistently across various devices due to differences in hardware specifications, screen sizes, and operating systems (iOS vs. Android). Microsoft (2024) explicitly acknowledge the limitations of Power Apps. In this study, some users experience slower performance, lagging, or app crashes, especially on older or lower-specification devices. In addition, users on slower networks can experience delays or incomplete data loads, impacting the user experience. Performance can vary depending on the mobile network quality, particularly when large datasets or media files are involved (Nasralla et al., 2023).

The widespread social consciousness and willingness to use mHealth apps for contact tracing and symptom submission was perceived as a priority during COVID-19. Notably, our findings reveal that social factors play less of a role in influencing user intention to use this technology in the absence of the threat of a pandemic. With fewer COVID-19 cases and a perceived reduction in risk, the incentive to use mHealth apps appears to have weakened. Social and institutional pressure to comply with contact tracing and symptom reporting has significantly diminished.

Considering behavioural intentions, this research shows that participants had varied views on whether they intend to use the UniHealth App in the future. While some participants showed a positive attitude and indicated they were likely to continue using the app, others were uncertain or did not plan to use it moving forward. The mixed opinions could reflect a range of factors influencing user behaviour.

Little is known about the usability of antigen self-tests (Maurer, 2024), this study leveraged innovative technology in the form of multiplex tests. Surprisingly, respondents indicated their satisfaction with using these tests for self-testing at home. While it is not the focus of this study, this research uncovered issues with the reliability of the multiplex tests in terms of false positives and in a malfunction in a small number of cases. To the best of our knowledge, our findings uncover new possibilities for research and practice beyond the emergency COVID-19 response for the use of self-test kits at home to test for respiratory illness. This may suggest that an outcome of the COVID-19 pandemic is increased health awareness, an acceptability of rapid tests, and mHealth but also a desire to rapidly identify and understand the cause of illness.

6.0 Conclusion

This study confirmed the relevance of established frameworks like the UTAUT for evaluating mHealth tools and self-testing technologies (Napitupulu et al., 2022). These frameworks provide a robust basis for assessing intention to use and overall usability, offering a consistent approach for evaluating health technology. This study provides early insights into the specific needs and preferences of university students and staff, highlighting the unique context of higher education. One key finding was that university staff reported higher satisfaction with the ease of use of the antigen tests compared to the accompanying app, suggesting that while digital platforms are convenient, physical testing methods are perceived as more user-friendly by certain groups. This difference in preference underscores the need for mHealth solutions that prioritise intuitive design and simplicity to accommodate varying levels of digital literacy among users.

This research is not without limitations. The study's sample size was limited and focused exclusively on a university setting, which may impact the generalisability of the findings to other demographics or broader populations and other settings. Further, descriptive statistics provide a snapshot of data but do not give insights into the underlying causes or relationships within the data.

This study presents practical implications for software designers, antigen test developers, policymakers, and healthcare providers, as they highlight areas for improvement in user interface design, the integration of digital and physical testing methods, and strategies for promoting self-testing within different populations. The COVID-19 pandemic highlighted the critical role of technology in managing a global health crisis. If home self-testing and the use of mHealth for symptom disclosure are implemented and used effectively, it can significantly benefit all members of a university community, particularly students. This, in turn, could reduce staff absenteeism, ensuring that classes are held as scheduled without the need for cancellations. With reduced viral transmission, students will be less likely to miss classes due to illness. Therefore, leveraging mHealth for symptom monitoring and disclosure could positively contribute to a healthier campus environment.

As the world anticipates and prepares for future pandemics, there are numerous research opportunities focused on enhancing digital preparedness capability. These opportunities span a range of areas, from improving digital health infrastructure to optimising technology adoption and usage across diverse populations. Future research should focus on assessment of UniHealth in diverse settings, particularly targeting vulnerable populations such a nursing homes and residential care settings. Further, these technologies could be adapted for other contexts and diseases such sexually transmitted diseases or even potentially identifying the need for vaccination boosters through mHealth enabled self-testing for the presence of antibodies against certain circulating viruses.

Declaration of Funding

This research is funded by Science Foundation Ireland - UniHealth 22/NCF/DR/11269.

References

Aburas, A. A., & Ayran, M. (2013). M-Health for Higher Education. *Turkish Online Journal of Distance Education*, *14*(2), 196-207.

Alam, M. Z., Hoque, M. R., Hu, W., & Barua, Z. (2020). Factors influencing the adoption of mHealth services in a developing country: A patient-centric study. International Journal of Information Management, 50, 128-143.

Aicken, C. R., Fuller, S. S., Sutcliffe, L. J., Estcourt, C. S., Gkatzidou, V., Oakeshott, P., ... & Shahmanesh, M. (2016). Young people's perceptions of smartphone-enabled self-testing and online care for sexually transmitted infections: qualitative interview study. *BMC public health*, *16*, 1-11.

Aranha, M., James, K., Deasy, C., & Heavin, C. (2021). Exploring the barriers and facilitators which influence mHealth adoption among older adults: A literature review. Gerontechnology, 20(2).

Asadzadeh, A., & Kalankesh, L. R. (2021). A scope of mobile health solutions in COVID-19 pandemics. Informatics in medicine unlocked, 23, 100558.

Birkmeyer, S., Wirtz, B. W., & Langer, P. F. (2021). Determinants of mHealth success: An empirical investigation of the user perspective. *International Journal of Information Management*, *59*, 102351.

Blebil, A. Q., Dujaili, J. A., Mohammed, A. H., Loh, L. L., Chung, W. X., Selvam, T., & Siow, J. Q. (2023). Exploring the eHealth literacy and mobile health application utilisation amongst Malaysian pharmacy students. *Journal of Telemedicine and Telecare*, *29*(1), 58-71.

Budd, J., Miller, B. S., Weckman, N. E., Cherkaoui, D., Huang, D., Decruz, A. T., ... & McKendry, R. A. (2023). Lateral flow test engineering and lessons learned from COVID-19. Nature Reviews Bioengineering, 1(1), 13-31.

Cervantes-Torres, L., & Romero-Blanco, C. (2023). Longitudinal study of the flash glucose monitoring system in type 1 diabetics: An mHealth ally in times of COVID-19. *Journal of Clinical Nursing*, *32*(13-14), 3840-3851.

Chan F, Thong J, Venkatesh V, Brown SA, Hu PJH, Tam K-Y (2010) Modeling citizen satisfaction with mandatory adoption of an e-government technology. J Assoc Inf Syst 11(10):519–549

Cui, H., Xie, J., Zhu, M., Tian, X., & Wan, C. (2022). Virus transmission risk of college students in railway station during post-COVID-19 era: Combining the social force model and the virus transmission model. Physica A: Statistical Mechanics and its Applications, 608, 128284.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, 319-340.

Fishbein, M., & Ajzen, I. (1975). Belie5 attitude, intention, and behavior: An introduction to theory and research. Reading, MA: Addison-Wesley.

Getachew, E., Adebeta, T., Muzazu, S. G., Charlie, L., Said, B., Tesfahunei, H. A., ... & Manyazewal, T. (2023). Digital health in the era of COVID-19: Reshaping the next generation of healthcare. Frontiers in public health, 11, 942703.

Hall, C. L., Gómez Bergin, A. D., & Rennick-Egglestone, S. (2024). Research Into Digital Health Intervention for Mental Health: 25-Year Retrospective on the Ethical and Legal Challenges. *Journal of Medical Internet Research*, *26*, e58939.

Karahanna, E., & Straub, D. W. (1999). The psychological origins of perceived usefulness and ease-of-use. *Information & management*, *35*(4), 237-250.

Ko, J. S., Stafylis, C., & Klausner, J. D. (2020). Mobile health promotion of human immunodeficiency virus self-testing in the United States. *Mhealth*, *6*.

Maurer, M. M. (2024). *Self-Tests for Health: Extent of Use, User Experience, and Predictors of Use of Various Types of Self-Tests* (Bachelors thesis, University of Twente).

Microsoft (2024). Other performance considerations, Last accessed 01/03/25 https://learn.microsoft.com/en-us/power-apps/maker/canvas-apps/app-performance-considerations

Napitupulu, D., Yacub, R., & Perdana Kusuma Putra, A. H. (2021). Factor Influencing of Telehealth Acceptance During COVID-19 Outbreak: Extending UTAUT Model. *International Journal of Intelligent Engineering & Systems*, *14*(3).

Nasralla, M. M., Khattak, S. B. A., Ur Rehman, I., & Iqbal, M. (2023). Exploring the role of 6G technology in enhancing quality of experience for m-health multimedia applications: a comprehensive survey. *Sensors*, *23*(13), 5882.

Nicolaidou, I., Aristeidis, L., & Lambrinos, L. (2022). A gamified app for supporting undergraduate students' mental health: A feasibility and usability study. *Digital Health*, *8*, 20552076221109059.

Pearson, W. S., Kreisel, K., Peterman, T. A., Zlotorzynska, M., Dittus, P. J., Habel, M. A., & Papp, J. R. (2018). Improving STD service delivery: Would American patients and providers use self-tests for gonorrhea and chlamydia?. Preventive Medicine, 115, 26-30.

Peprah, P., Abalo, E. M., Agyemang-Duah, W., Gyasi, R. M., Reforce, O., Nyonyo, J., ... & Kaaratoore, P. (2019). Knowledge, attitude, and use of mHealth technology among students in Ghana: A university-based survey. *BMC medical informatics and decision making*, *19*, 1-11.

Pittman, T. W., Decsi, D. B., Punyadeera, C., & Henry, C. S. (2023). Saliva-based microfluidic point-of-care diagnostic. Theranostics, 13(3), 1091.

Rouidi, M., Hamdoune, A., Choujtani, K., & Chati, A. (2022). TAM-UTAUT and the acceptance of remote healthcare technologies by healthcare professionals: A systematic review. Informatics in Medicine Unlocked, 32, 101008.

Simmhan, Y., Rambha, T., Khochare, A., Ramesh, S., Baranawal, A., George, J. V., ... & Kiran, R. (2020). GoCoronaGo: privacy respecting contact tracing for COVID-19 management. *Journal of the Indian Institute of Science*, *100*, 623-646.

Smith, P. S., Alaa, A., Sasco, E. R., Bagkeris, E., & El-Osta, A. (2023). How has COVID-19 changed healthcare professionals9 attitudes to self-care? A mixed methods research study. PloS One, 18(7), e0289067. https://doi.org/10.1371/journal.pone.0289067

Undelikwo, V. A., Shilton, S., Folayan, M. O., Alaba, O., Reipold, E. I., & Martínez-Pérez, G. Z. (2023). COVID-19 self-testing in Nigeria: Stakeholders' opinions and perspectives on its value for case detection. Plos one, 18(4), e0282570.

van der Waal, N. E., de Wit, J., Bol, N., Ebbers, W., Hooft, L., Metting, E., & Van der Laan, L. N. (2022). Predictors of contact tracing app adoption: Integrating the UTAUT, HBM and contextual factors. Technology in Society, 71, 102101.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS quarterly, 425-478.

Wu, B., & Chen, X. (2017). Continuance intention to use MOOCs: Integrating the technology acceptance model (TAM) and task technology fit (TTF) model. *Computers in human behavior*, *67*, 221-232.

Ye, J. (2020). The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. JMIR medical informatics, 8(7), e19866.

Zapata, B. C., Fernández-Alemán, J. L., Idri, A., & Toval, A. (2015). Empirical studies on usability of mHealth apps: a systematic literature review. *Journal of medical systems*, *39*, 1-19.

Appendix 1.

The aim of this survey is to understand the usability, accessibility, and the likelihood of adoption of the mHealth application, and the study supplied antigen test.

Demographic Questions:

Please select the Age range of which you belong to:

18- 24, 25-34, 35-44, 45-54, 55-64 and 65 and over

What level of programme are you currently enrolled in:

Undergraduate student, Postgraduate student, Staff Member

	Performance Expectancy
PE1a	Using the UniHealth app enables me to submit my symptoms and self-test results in a meaningful and accessible way.
PE1b	Using the supplied antigen test enables me to self-test in a meaningful and accessible way.
PE1c	Using the UniHealth app would improve my quality of life.
PE2a	Using the UniHealth app would positively impact my health and wellbeing.
PE2b	Using the supplied antigen test would positively impact my health and wellbeing.
PE3a	Using the UniHealth app would give me greater control over my understanding of my symptoms.
PE3b	Using the UniHealth antigen test would give me greater control over my understanding of my symptoms.
	Effort Expectancy
EE1a	I find using the UniHealth app simple.
EE1b	It is easy for me to use the UniHealth app.
EE1c	I rarely have issues using the UniHealth app.
EE2a	I find using the supplied antigen tests simple.
EE2b	It is easy for me to use the supplied antigen tests.
EE2c	I rarely have issues using the supplied antigen tests.
	Social Influence
SI1a	People who influence my behaviour would think that I should use the UniHealth app.
SI1b	People who are important to me would think that I should use the UniHealth app.
SI1c	People who are in my social circle would think that I should use the UniHealth app.
	Facilitating conditions
FC1a	The university has been helpful in encouraging the use of the UniHealth app and supplied antigen test.
FC2a	I have the resources (e.g., access to hardware, internet connection, phone credit) necessary to use the UniHealth app.

FC2b	I have the knowledge necessary to use the UniHealth app.
FC2c	I have the knowledge necessary to use the supplied antigen test.
FC3a	A specific person (or group) is available for assistance with difficulties associated with accessing and using the UniHealth app and supplied antigen test.
FC3b	Instructions about using the UniHealth app and supplied antigen test are available to me.
	Trusting Beliefs
TB1a	The UniHealth app is dependable and reliable.
TB1b	The supplied antigen test is dependable and reliable.
TB2a	I trust the UniHealth app because it serves my best interests.
TB2b	I trust the supplied antigen test because it serves my best interests.
TB3a	Trusting the UniHealth app is easy for me.
TB3b	Trusting the supplied antigen test is easy for me.
	Behavioural Intention
BI1a	I will use the UniHealth app for my health and wellbeing needs.
BI1b	I predict many universities will use the UniHealth app in the future.
BI1c	I will recommend the UniHealth app to others.

On the Social Design Effectiveness of Healthcare Chatbots: A Systematic Literature Review

Research in Progress

Yuanyuan Lai [1] * and Meixi He [2]

[1] *Royal Holloway University of London, School of Business and Management, UK, yuanyuan.lai@rhul.ac.uk*
[2] *Nanjing University of Aeronautics and Astronautics, Nanjing, China, hemeixi@nuaa.edu.cn*

Abstract

Healthcare chatbots are increasingly used to provide immediate, on-demand support for clinical care, mental health, patient engagement, and administrative efficiency. As these digital agents become more integrated into healthcare service delivery, designing chatbots that aim to foster social interactions has attracted interest from both academia and practice to enhance user trust and engagement. Social design elements such as human-like visual appearance, communication style, and personality have been examined by a number of studies in recent years. However, the knowledge related to the effectiveness of these social design cues remains segregated as this topic spans different use scenarios and different research areas including information systems, human-computer interactions, and digital health. To obtain a comprehensive overview of what social designs of healthcare chatbots have been studied so far and how they affect users' interaction outcomes in the context of healthcare services, this study presents a systematic literature review (SLR) to synthesise existing research. Using a comprehensive search across four databases, this review identifies 54 records after the keywords/abstract screening and full-text screening. The next steps are outlined at the end of this work.

Keywords: chatbots, healthcare, social design, anthropomorphism, conversational agents

1.0 Introduction

Conversational artificial intelligence (AI) agents or chatbots have transformational potential to automate basic and repeatable customer service interactions (Schanke et al., 2021). The interest in chatbots from both scientific and industrial sectors has increased substantially in recent years, particularly in areas like retail, healthcare, public management, and education (Feine et al., 2019; Ju et al., 2022). In the healthcare service context, chatbots have emerged as promising tools to complete tasks such as preliminary symptom assessments, automated triage, mental health support, appointment scheduling, and telemedicine support (Lai et al., 2023; W. Liu et al., 2024; Nadarzynski et al., 2019). Driven by technical advancements such as natural language processing, machine learning, and cloud computing (Diederich et al., 2022; Elshan et al., 2022), these digital agents are increasingly capable of simulating interactive, human-like conversations with users (Schuetzler et al., 2018; Seeger et al., 2021), invoking social-emotional responses such as trust (W. Liu et al., 2024) and liking (Lee & Lee, 2023).

It's been acknowledged by researchers that the design and evaluation of conversational agents should include both their technical performance and social interaction aspects (Araujo, 2018; Bickmore & Picard, 2005; Feine et al., 2019). The technical potentials of applying healthcare chatbots have received lots of research efforts (e.g., Ceney et al., 2021; Inkster et al., 2018; Mehta et al., 2021; Nadarzynski et al., 2019), however, the effectiveness of their social design aspects on users' interaction outcomes (e.g., attitudes, perceptions, intentions, and behaviour (Elshan et al., 2022)) remains underexplored. Unlike chatbots in other fields, healthcare chatbots are used in situations that often are emotionally charged and ethically sensitive. Thus, incorporating social cues (e.g., human-like conversational tone, avatars) in chatbots' interface design is therefore considered an important way to maintain and foster more natural interactions to achieve higher user trust and satisfaction (W. Liu et al., 2024; Schillaci et al., 2024; You et al., 2023). For instance, chatbot *Sensely*[1] used by the UK's NHS has added a female doctor's face and name (Olivia) to increase the human touch of the interaction. Nonetheless, as claimed by Schillaci et al. (2024), the impact of such design is contingent on contextual variables like users' characteristics, chatbots' roles, disease types etc.

In addition, previous studies in different research streams and disciplines have given various names to describe the social design aspects of chatbots (e.g., anthropomorphism, anthropomorphic cues, human-like characteristics, humanness, social cues, social presence, empathic), making it difficult to synthesise the effectiveness of social designs for specific application scenarios like healthcare. Furthermore, the heterogeneous functional purposes of healthcare chatbots and many relationships between design variables and use outcomes from different use cases resulted in a fragmented and sparse literature base. Even for the same design elements, different use contexts might generate different or even contradictory research results. Given this, in this study, we use social design to encompass the various social design elements that make the interactions feel genuine, such as chatbots' visual cues, conversational design, personality, and empathy.

[1] https://websdk.sense.ly/latest/?nhsSignupEnabled=true

Several literature reviews on the design of chatbots have emerged in recent years. For example, Elshan et al. (2022) conducted a literature review and classified the identified design elements into verbal elements, auditory elements, invisible elements, visual elements, and interaction design elements with a focus on user acceptance of intelligent agents in general. Diederich et al. (2022) analysed conversational agent research regarding user interaction, context, agent design, as well as user perceptions and outcomes. Although both papers provided insights into the design aspects of conversational agents, neither of them focused on specific application scenarios like healthcare. The user perception and interaction outcomes with healthcare chatbots can be very different from chatbots used in other general domains, and some social design characteristics (e.g., sense of humour) may not match users' needs in this context. Therefore the design configurations need to be analysed separately along with the boundary conditions. To this end, this study aims to address the following two research questions:

RQ1: What social design elements of healthcare chatbots have been examined by previous literature?

RQ2: How does social design affect users' interaction outcomes with healthcare chatbots?

To answer the above two questions, this study will conduct a systematic literature review to identify the influencing mechanisms of various social design elements of healthcare chatbots. This study will provide a comprehensive overview for future healthcare chatbot-related research and offer feasible design suggestions.

2.0 Research Background

2.1 Healthcare Chatbots

To increase the efficiency and cost-effectiveness of service delivery, AI-based solutions are now playing an important role in the healthcare sector (Kumar et al., 2021). Healthcare chatbots can mimic human conversations through voice commands or text messages to offer information and personal assistance (Luo et al., 2019; Schuetzler et al., 2020), allowing for 24/7 continuous support and reducing strain on human professionals. Previous studies suggest that users do obtain meaningful support and experiences through healthcare chatbots. For example, the chatbot *Vik* improved cancer patients' medication adherence through reminders and educational content (Chaix et al., 2019). Miller et al. (2020) found that users reported positive experiences with the symptom assessment chatbot *Ada Health* in a primary care setting in London. Ho et al. (2018) demonstrated that chatbots and humans are both effective in terms of soliciting emotional, relational, and psychological outcomes. Zamora (2017) reports similar findings on using chatbots to fulfil users' emotional needs since the experience with chatbots lacks perceived judgment.

Drawing on the literature together, we categorised healthcare chatbots into four groups (see Figure 1) along the social-orientation and task-orientation dimensions (Pezenka et al., 2024). They are: (1) Symptom assessment chatbots that can pre-diagnose symptoms and provide further guidance on treatment and medication, such as *Ada Health*, *Mediktor* and *Left-handed Doctor* (Lai et al., 2023; You et al., 2023); (2) Mental health support chatbots that use cognitive behavioural therapy to help users identify and cope with mental issues, such as Wysa and Woebot (Inkster et al., 2018); (3) Social companion chatbots like Replika are AI agents that are designed to engage with users in meaningful and empathetic ways to provide social support and companionship (Possati, 2023); (4) Customer service chatbots that are mainly used by health institutions to provide information on hospital billing, service hours, payment, insurance etc. Considering that each type of healthcare chatbot has different usage scenarios, target users, and functions, it naturally requires a different design focus to cater for the corresponding user needs.

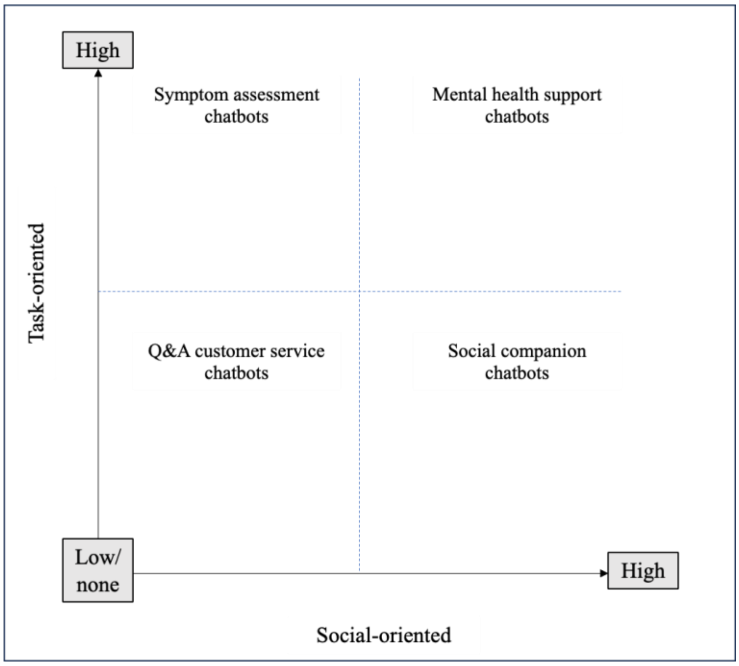

High
Symptom assessment chatbots
Mental health support chatbots
Task-oriented
Q&A customer service chatbots
Social companion chatbots
Low/ none
High
Social-oriented

Figure 1. Types of healthcare chatbots

2.2 Social design of chatbots

Greulich & Schlieter (2023) classified the anthropomorphic designs of chatbots into four categories: identity, verbal, nonverbal and embodiment. Identity refers to providing a human identity for the chatbot such as a human name (Araujo, 2018; W.

Liu et al., 2024), gender and personality (Kang & Kang, 2024). Verbal refers to the verbal anthropomorphic design in the conversation, such as human-like conversation tone (W. Liu et al., 2024), response delays (Diederich et al., 2021) and empathy (B. Liu & Shyam Sundar, 2018). Nonverbal refers to non-verbal cues in the conversation, such as punctuation to highlight emotions (Han et al., 2023) and emoticons to show human-like social responses (Seeger et al., 2021). Embodiment refers to endowing the chatbot with a virtual or physical representation.

In previous studies, the social design of chatbots has been shown to reduce the psychological distance between humans and chatbots (Li & Sung, 2021), enhance social presence (Konya-Baumbach et al., 2023), increase users' trust (Pizzi et al., 2023) and reuse intention (W. Liu et al., 2024). However, according to the Uncanny Valley Theory, overly realistic machines can instead create a feeling of eeriness in users, which can reduce users’ trust, and elicit discomfort or resistance in users (Song & Shin, 2024). An experimental study pointed out that when the perceived control is low, consumers would perceive stronger threats and prefer less anthropomorphic AI service agents (Yang et al., 2022). Not only that, highly anthropomorphic designs raise users' expectations of chatbots, but if chatbots fail to deliver the expected services, users will be more disappointed (Crolic et al., 2022).

3.0 Research Approach

We took a three-step approach to conduct the review following Elshan et al. (2022). In the first stage, we developed the search strings. Given that this research aims to investigate the design aspects of chatbots used in healthcare and medical sectors, the search strings consisted of three parts. The first part relates to the keywords of chatbots including synonyms that indicate text-based conversational agents and used in previous studies (Diederich et al., 2022; Elshan et al., 2022). The second part points to the healthcare and medical service contexts of this research. The last part corresponds to the social and anthropomorphic design focus of this study. This resulted in the following search string:

(("chatbot") OR ("chat bot*") OR ("intelligent agent*") OR ("conversational agent*")) AND (("health*") OR ("medical")) AND (("social design") OR ("anthropomorph*") OR ("human-like*") OR ("humanness"))*

In the second step, we selected relevant publication outlets to capture a representative sample of empirical research on the design and user interaction with healthcare chatbots. To encompass recent developments in this rapidly evolving field, our search included both journals and conference proceedings in the areas of information systems (IS), computer science (CS), human-computer interactions

(HCI), and medicine. This allowed us to incorporate timely studies published through conference proceedings, which often have faster acceptance processes compared to journals, ensuring that recent and significant contributions were not overlooked (Elshan et al., 2022). To collect appropriate studies, we searched through databases including Scopus, Web of Science, PubMed and EBSCO in early March of 2025 and limited the time scope up until the end of February of 2025.

The third step was for paper selections. The query returned 497 records in total. We removed 214 duplicate records which left us 283 articles for the next screening. To eliminate obviously irrelevant studies, in the initial filtering stage, two authors conducted the title-abstract-keywords screening independently for the 214 papers. For this round of screening, we used inclusion criteria: (1) Written in English; (2) Papers that provided empirical insights on user interaction with healthcare chatbots; and (3) Healthcare chatbots are designed for lay people, not for healthcare professionals; (4) Focused on text-based chatbots, not voice-based agents i.e., Apple's Siri, Amazon's Alex. After being assessed against the inclusion criteria and calibration between the two authors, 73 papers were kept for the next round of full-text screening. In this round, we used the exclusion criteria: (1) Studies that don't examine specific social design elements; (2) Studies that have no evidence for the relationships between social design and user outcomes (i.e., studies using thematic analysis to extract major topics from user reviews). 54 papers were kept after this round.

Next Steps: We will add extra sources of articles by performing a backward/forward search and reporting the total number of articles included. Then we will conduct the literature synthesis to summarise the social design elements from the searched literature and their impact on the user interaction outcomes.

References

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183–189. https://doi.org/10.1016/j.chb.2018.03.051

Bickmore, T. W., & Picard, R. W. (2005). Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Transactions on Computer-Human Interaction*, *12*(2), 293–327.

Ceney, A., Tolond, S., Glowinski, A., Marks, B., Swift, S., & Palser, T. (2021).

Accuracy of online symptom checkers and the potential impact on service utilisation. *PLoS ONE*, *16*(7), 1–16. https://doi.org/10.1371/journal.pone.0254088

Chaix, B., Bibault, J. E., Pienkowski, A., Delamon, G., Guillemassé, A., Nectoux, P., & Brouard, B. (2019). When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. *Journal of Medical Internet Research*, *21*(5), 1–7.

Crolic, C., Thomaz, F., Hadi, R., & Stephen, A. T. (2022). Blame the Bot: Anthropomorphism and Anger in Customer-Chatbot Interactions. *Journal of Marketing*, *86*(1), 132–148. https://doi.org/10.1177/00222429211045687

Diederich, S., Benedikt Brendel, A., Morana, S., Kolbe, L., & Benedikt, A. (2022). On the Design of and Interaction with Conversational Agents: An Organizing and Assessing Review of Human-Computer Interaction Research. *Journal of the Association for Information Systems*, *23*(1), 96–138. https://doi.org/10.17705/1jais.00724

Diederich, S., Lembcke, T.-B., Brendel, A. B., & Kolbe, L. (2021). Understanding the Impact that Response Failure has on How Users Perceive Anthropomorphic Conversational Service Agents: Insights from an Online Experiment. *AIS Transactions on Human-Computer Interaction*, *13*(1), 82–103. https://doi.org/10.17705/1thci.00143

Elshan, E., Zierau, N., Engel, C., Janson, A., & Leimeister, J. M. (2022). Understanding the Design Elements Affecting User Acceptance of Intelligent Agents: Past, Present and Future. *Information Systems Frontiers*, *24*, 699–730. https://doi.org/10.1007/s10796-021-10230-9

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human-Computer Studies*, *132*, 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

Greulich, S., & Schlieter, H. (2023). “Look Closer” Anthropomorphic Design and Perception of Anthropomorphism in Conversational Agent Research. *Forty-Fourth International Conference on Information Systems, Hyderabad, India*, 1–17. https://aisel.aisnet.org/icis2023/aiinbus/aiinbus/9

Han, E., Yin, D., & Zhang, H. (2023). Bots with Feelings: Should AI Agents Express Positive Emotion in Customer Service? *Information Systems Research*, *34*(3), 1296–1311. https://doi.org/10.1287/isre.2022.1179

Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, *68*(4), 712–733. https://doi.org/10.1093/joc/jqy026

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *Journal of Medical Internet Research*, *6*(11), 1–14. https://doi.org/10.2196/12106

Ju, J., Meng, Q., Sun, F., Liu, L., & Singh, S. (2022). Citizen preferences and government chatbot social characteristics : Evidence from a discrete choice experiment. *Government Information Quarterly, December*, 101785. https://doi.org/10.1016/j.giq.2022.101785

Kang, E., & Kang, Y. A. (2024). Counseling Chatbot Design : The Effect of Anthropomorphic Chatbot Characteristics on User Self-Disclosure and Companionship. *International Journal of Human–Computer Interaction*, *40*(11), 2781–2795. https://doi.org/10.1080/10447318.2022.2163775

Konya-Baumbach, E., Biller, M., & von Janda, S. (2023). Someone out there? A study on the social presence of anthropomorphized chatbots. *Computers in Human Behavior*, *139*, 107513. https://doi.org/10.1016/j.chb.2022.107513

Kumar, P., Dwivedi, Y. K., & Anand, A. (2021). Responsible artificial intelligence (AI) for value formation and market performance in healthcare: the mediating role of patient’s cognitive engagement. *Information Systems Frontiers*, 1–24.

Lai, Y., Panagiotopoulos, P., & Lioliou, E. (2023). Empowering users with medical artificial intelligence technologies. *Thirty-First European Conference on Information Systems*, 1–17. https://aisel.aisnet.org/ecis2023_rp

Lee, J., & Lee, D. (2023). Telematics and Informatics User perception and self-disclosure towards an AI psychotherapy chatbot according to the anthropomorphism of its profile picture. *Telematics and Informatics*, *85*(August), 102052. https://doi.org/10.1016/j.tele.2023.102052

Li, X., & Sung, Y. (2021). Anthropomorphism brings us closer: The mediating role of psychological distance in User-AI assistant interactions. *Computers in Human Behavior, 118,* 106680. https://doi.org/10.1016/j.chb.2021.106680

Liu, B., & Shyam Sundar, S. (2018). Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. *Cyberpsychology, Behavior, and Social Networking*, *21*(10), 625–636.

https://doi.org/10.1089/cyber.2018.0110

Liu, W., Jiang, M., Li, W., & Mou, J. (2024). How does the anthropomorphism of AI chatbots facilitate users' reuse intention in online health consultation services? The moderating role of disease severity. *Technological Forecasting and Social Change, 203,* 123407. https://doi.org/10.1016/j.techfore.2024.123407

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: machines vs. humans: the impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science, 38*(6), 937–947.

Mehta, A., Andrea, B. ;, Niles, N., Vargas, J. H., Marafon, T., Dotta Couto, D., & Gross, J. J. (2021). Acceptability and Effectiveness of Artificial Intelligence Therapy for Anxiety and Depression (Youper): Longitudinal Observational Study. *Journal of Medical Internet Research, 23*(6), e26771. https://doi.org/10.2196/26771

Miller, S., Gilbert, S., Virani, V., & Wicks, P. (2020). Patients' utilization and perception of an artificial intelligence-based symptom assessment and advice technology in a British primary care waiting room: exploratory pilot study. *Journal of Medical Internet Research, 7*(3), e19713.

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. *Digital Health, 5,* 1–12.

Pezenka, I., Aunimo, L., Janous, G., & Dobrowsky, D. (2024). Emotionality in Task-Oriented Chatbots–The Effect of Emotion Expression on Chatbot Perception. *Communication Studies, 75*(6), 825–843. https://doi.org/10.1080/10510974.2024.2363259

Pizzi, G., Vannucci, V., Mazzoli, V., & Donvito, R. (2023). I, chatbot! the impact of anthropomorphism and gaze direction on willingness to disclose personal information and behavioral intentions. *Psychology & Marketing*. https://doi.org/10.1002/MAR.21813

Possati, L. M. (2023). Psychoanalyzing artificial intelligence: the case of Replika. *AI and Society, 38*(4), 1725–1738. https://doi.org/10.1007/s00146-021-01379-7

Schanke, S., Burtch, G., & Ray, G. (2021). Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research*, *32*(3), 736–751.

Schillaci, C. E., de Cosmo, L. M., Piper, L., Nicotra, M., & Guido, G. (2024). Anthropomorphic chatbots for future healthcare services: Effects of personality, gender, and roles on source credibility, user satisfaction, and intention to use. *Technological Forecasting and Social Change*, *199*, 123025. https://doi.org/10.1016/j.techfore.2023.123025

Schuetzler, R. M., Giboney, J. S., Grimes, G. M., & Nunamaker, J. F. (2018). The influence of conversational agent embodiment and conversational relevance on socially desirable responding. *Decision Support Systems*, *114*, 94–102. https://doi.org/10.1016/j.dss.2018.08.011

Schuetzler, R. M., Grimes, G. M., & Giboney, J. S. (2020). The impact of chatbot conversational skill on engagement and perceived humanness. *Journal of Management Information Systems*, *37*(3), 875–900.

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2021). Texting with Humanlike Conversational Agents: Designing for Anthropomorphism. *Journal of the Association for Information Systems*, *22*(4), 931–967. https://doi.org/10.17705/1jais.00685

Song, S. W., & Shin, M. (2024). Uncanny Valley Effects on Chatbot Trust, Purchase Intention, and Adoption Intention in the Context of E-Commerce: The Moderating Role of Avatar Familiarity. *International Journal of Human–Computer Interaction*, *40*(2), 441–456. https://doi.org/10.1080/10447318.2022.2121038

You, Y., Tsai, C., Li, Y., Ma, F., Heron, C., & Gui, X. (2023). Beyond Self-diagnosis: How a Chatbot-based Symptom Checker Should Respond. *ACM Transactions on Computer-Human Interaction*. https://doi.org/10.1145/3589959

Zamora, J. (2017). I'm Sorry, Dave, I'm afraid I can't do that: Chatbot perception and expectations. *5th International Conference on Human Agent Interaction*, 253–260.

Monetization of Plastic Waste: An Intelligent Sharing Economy Model for Managing Plastic Waste in Ghana

Joseph Kwame Adjei (Computer Science & Information Systems Department, Ashesi University),
Clifford Yeboah (Computer Science & Information Systems Department, Ashesi University) and
Henry Owusu (Computer Science & Information Systems Department, Ashesi University)

Research in Progress

Abstract

Plastic waste is a major environmental and socio-economic problem in Ghana, where urbanization and the increasing use of plastic are exacerbating the inefficiency of waste management systems. This paper proposes a new framework for monetizing the collection and management of plastic waste through the sharing economy. The proposed model uses digital platforms to bring together households, waste collectors, recyclers, and businesses, to monetize plastic waste collection. Drawing on the circular economy, behavioral and institutional issues, the paper outlines ways to monetize plastic waste and to incentivise key players in the plastic waste ecosystem. The framework is designed to increase sustainability, create economic opportunities, and reduce environmental damage. Implementation strategies and policy implications are discussed to highlight the practical relevance of the study.

Keywords: Plastic Waste Management, Sharing Economy, Circular Economy, Digital Platforms, Monetization, Incentivized Recycling

1.0 Introduction

The increasing challenge of plastic waste poses major environmental threat globally. Geyer et al (2017) estimated that less than 10% of the plastic waste produced are recycled (Geyer et al., 2017). In developing countries such as Ghana, the issue has dire consequences, particularly, due to the growing use of plastics and inefficient plastic waste management (PSM). Quartey et al. (2015) noted that various plastics, including low-density polyethylene (LDPE) bags, high-density polyethylene (HDPE), polypropylene, polystyrene, polyvinyl chloride (PVC), and polyethylene terephthalate (PET), are widely used in Ghana for packaging food, water, and groceries (Quartey et al., 2015).

A recent World Economic Forum (WEF) study indicated that, Ghana generates about 840,000 tons of plastic waste annually and only 9.5% is collected and recycled (WEF, 2023). Such improper plastic waste management leads to widespread pollution, causing ecological and health issues, contamination of water bodies, flooding, and biodiversity loss (Alqattaf, 2020). The challenge is exacerbated by structural inefficiencies in waste management systems, due to inadequate resources, lack of policy enforcement, and limited support to the informal sector plastics collectors, who incidentally contribute significantly to effective PSM (Tahiru et al., 2024).

Despite the enormous environmental challenges, plastics such as PET and HDPE are convertible into valuable industrial inputs, construction materials, and renewable energy (Faisal et al., 2023; Smith et al., 2022). Thus, the need to transition from the fragmented approach to PSM to effective systems that recognise and interconnect plastic waste generators, informal collectors and aggregators, recycling plants, policy makers, and the local authorities, in a more streamlined PSM ecosystem.

The sharing economy model presents a promising solution to the (PSM) challenges (Puschmann & Alt, 2016). The sharing economy model has already revolutionized the transport and logistics services (e.g., Uber, Bolt and Glovo), the hospitality services (e.g., Airbnb), in many countries, etc. Although the model was tested in India for waste management (Kabadiwalla Connect), it has not been tested in Africa, particularly Ghana. Moreover, there is limited literature on the application of digital technologies in optimizing waste management logistics (Cohen & Kietzmann, 2014). Thus, the study attempts to address the following research questions: *What key factors contribute to a streamlined plastic waste management ecosystem that are more beneficial to all stakeholders? How can a sharing economy model be effectively designed and implemented to streamline plastic waste management in Ghana?* To address these questions, we explore the fundamental requirements for integrating the sharing economy model in PSM. We then design and implement a sharing economy platform for PSM Ghana.

This study is significant because it responds to the urgent need for systemic change and novel approach to PSM in Ghana, that aligns economic incentives, affordable technological solutions, and address environmental challenges caused by plastic. The relevant literature are analyzed in the next section, followed by a detailed description of our methods and theories examined. We then proceed to craft the conceptual model followed by discussions of the results, our conclusions and recommendations.

2.0 Literature and Theory

This section explores the key concepts in PSM, the sharing economy frameworks, and the theoretical foundations of the proposed model. Each subsection synthesizes key findings from seminal and recent works, establishing the relevance of these concepts to the problem of plastic waste in Ghana.

2.1 The Plastic Waste Challenge

The global plastic waste crisis has reached an alarming level, with the UNDP Environment Program estimating that more than 400 million tons of plastic waste are generated annually (UNDP, 2024). The World Economic Forum (2023) also warned that plastic waste could outnumber fish stocks in the oceans by 2050 if the current rate of plastic waste entering the oceans (WEF, 2023) continues. In Ghana, plastic waste has become a serious environmental and public health problem. Ghana is estimated to have around 90% of the 840,000 tons of plastic waste generated annually not effectively managed, thereby choking drainage systems and eventually reaching water bodies and the ocean (WEF, 2023).

A substantial contributor to plastic waste is the pervasive consumption of sachet water, which reached an estimated 11.3 million litres daily in 2017 (Wardrop et al., 2017). With Ghana’s population growing from 29 million in 2017 to 34 million in 2024 (Data Commons, 2023), the reliance on sachet water as the source of primary drinking has further aggravated plastic pollution (Jambeck et al., 2015). Other sources include the widespread use of plastics for packaging due to its affordability, durability, and convenience (Ugwu & Godwin-Okoubi, n.d.). Abrokwah et al, (2022) also found that rise in income levels, perceptions of hygiene and safety as well as convenience are the main factors driving the patronage of bottled water (Abrokwah et al., 2022).

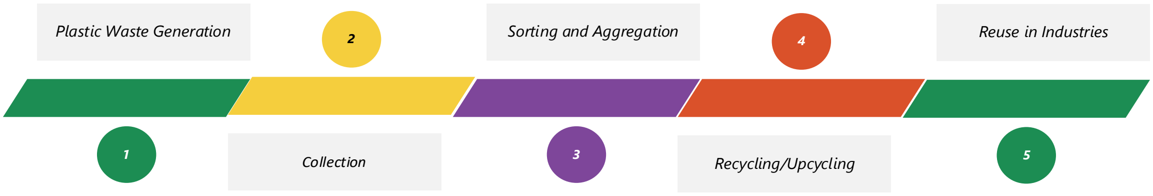


Figure 1: A schematic representation of Plastic Waste Supply Chain

The plastic waste supply chain in Ghana, as illustrated in Figure 2, consists of four key stages: collection, sorting and aggregation, recycling/upcycling, and reuse in industries. For the efficient management of plastic waste, the collection stage which involves gathering plastic waste from households, businesses, and public spaces is often facilitated by informal waste pickers. This is a very critical stage in PSM logistics as it prevents dumping in open spaces and clogging the drainage systems (Wardrop et al., 2017).

In the sorting and aggregation stage, collected plastics are categorized by type and quality, then consolidated for further processing. The recycling/upcycling stage involves transforming sorted plastics into new products, though limited infrastructure in Ghana poses challenges (Abrokwah et al., 2022). Small-scale initiatives and community projects play a vital role in this process. Finally, the reuse stage in industries integrates recycled plastics into manufacturing, reducing the demand for virgin plastics and mitigating environmental impact. Addressing these issues requires improved infrastructure, public awareness, and supportive policies to enhance recycling and reuse efforts. Strengthening the plastic waste supply chain is essential for achieving sustainable waste management in Ghana (Debrah et al., 2021; Okai, 2020).

2.2 Efforts to Address Plastic Waste in Ghana

Several policies and laws have been introduced to govern the regulatory framework of waste and sanitation in Ghana. These include the 1992 Constitution as amended in 1996, the revised National Environmental Sanitation Policy (2010), the Environmental Protection Agency Act 490 (1994), and the Public Health Act (2012) (Kwansa, 2021). The Environmental Sanitation By-Laws (2003) and the National Environmental Policy (2012) also provide additional guidelines for waste management, whereas the Local Governance Act 936 (2016) also mandates Metropolitan, Municipal, and District Assemblies (MMDAs) to implement and enforce sanitation and environmental safety policies in Ghana (Kwansa, 2021).

The Ghana Recycling Initiative by Private Enterprises (GRIPE) is making significant strides in tackling plastic pollution (Kwansa, 2021; Okai, 2020). The GRIPE initiative includes implementing recycling programs in schools, promoting waste separation, and organizing plastic buyback events. The GRIPE program partnered with Environment 360 in 2018 to launch a school recycling program in 19 schools in Tema Newtown with plans to expand to 60 schools by 2020 (Kwansa, 2021). GRIPE collaborated with Premier Waste Services to promote community plastic buyback event, combining plastic collection with public awareness campaigns to educate the community about the importance of proper waste management (“GRIPE Case Study,” 2021).

The National Plastic Action Partnership (NPAP) was also launched in 2019 in collaboration with the World Economic Forum’s Global Plastic Action Partnership (GPAP) focusing on creating a circular economy for plastics in Ghana by mobilizing stakeholders across the plastics value chain (WEF, 2023) (Global Plastic Action Partnership, 2021). Key activities include developing a mobile software package to provide waste pickers with real-time data on plastic prices, improving transparency in the supply chain, and supporting the establishment of recycling infrastructure. NPAP has also facilitated partnerships with international organizations to secure funding for large-scale recycling projects, aiming to reduce plastic leakage into Ghana’s oceans and waterways (Global Plastic Action Partnership, 2021).

Despite these efforts, Ghana’s PSM initiatives remains fragmented. Integrating informal waste pickers, who recover substantial amounts of recyclables into formal systems is a critical opportunity for improving waste recovery rates. Further investments in sorting centers, public education, and stakeholder collaboration are essential to strengthening Ghana’s plastic waste value chain and mitigating environmental damage. Kwansa, (2021) has therefore proposed that to effectively manage plastic waste in Ghana, there should be effective planning, sensitization, and coordination among stakeholders, appropriate technology, enforcements of the waste management regulations, and equitable incentive mechanism for all stakeholders in the ecosystem (Kwansa, 2021).

2.3 Factors Influencing a Streamlined PSM Ecosystem

To address the first research question, *what key factors contribute to a streamlined plastic waste management ecosystem that are more beneficial to all stakeholders?,* we conducted a comprehensive analysis using a Fishbone Diagram (Ishikawa Diagram) (Ishikawa & Loftus, 1990). The Fishbone diagram (aka cause and effect diagram) is a graphical tool for comprehensive analysis of the interplay of major causes and the effect of a phenomena or an event (Coccia, 2020).

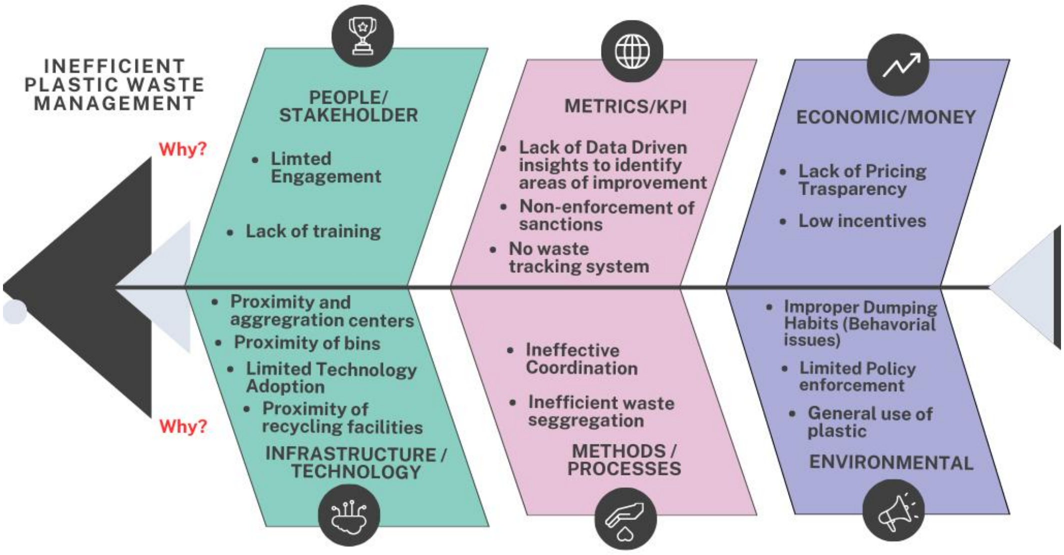

INEFFICIENT PLASTIC WASTE MANAGEMENT
Why?
PEOPLE/ STAKEHOLDER
Limted Engagement
Lack of training
METRICS/KPI
Lack of Data Driven insights to identify areas of improvement
Non-enforcement of sanctions
No waste tracking system
ECONOMIC/MONEY
Lack of Pricing Trasparency
Low incentives
Why?
Proximity and aggregration centers
Proximity of bins
Limited Technology Adoption
Proximity of recycling facilities
INFRASTRUCTURE / TECHNOLOGY
Ineffective Coordination
Inefficient waste seggregation
METHODS / PROCESSES
Improper Dumping Habits (Behavorial issues)
Limited Policy enforcement
General use of plastic
ENVIRONMENTAL

Figure 2: Fishbone diagram showing the primary causes of inefficient PSM (“the why’s”)

The tool was used to identify and categorize the primary factors influencing the PSM ecosystem in Ghana. Each category is further decomposed into sub-factors that collectively shape the efficiency and effectiveness of the PSM system. The diagram highlights main categories: stakeholders, metrics, economic factors, technology and infrastructure, business processes and the environment**.**

2.4 Behavioral and Institutional Issues

Various economic, social, and technical challenges hinder effective PSM in developing countries. De Feo & Ferrara, (2024) have emphasized the need to understand the behavioral patterns to design successful waste management strategies (De Feo & Ferrara, 2024). In Ghana, inadequate waste disposal practices are often linked to a lack of awareness and insufficient institutional support.

Institutional pressures shape organizational and individual behaviors, offering a framework to understand decision-making in waste management. Kitole and Sesabo (2024), studied small-scale recycling firms in Tanzania and noted the impact of coercive, normative, and mimetic pressures on efficient PSM. Coercive pressure arises from formal regulations and policies. Kitole and Sesabo (2024) found it significantly influenced recycling investments in Tanzania. Such coercive pressures could be leveraged to mandate compliance with recycling targets, waste segregation, and proper disposal practices. For instance, policies that penalize improper disposal of plastics or incentivize recycling through monetized proposed sharing economy can drive behavioral change.

Normative pressure stems from social norms and expectations. Kitole and Sesabo (2024) highlighted its role in metal recycling in Tanzania, where social norms and business associations influenced decisions. For the proposed sharing economy model, normative pressure can foster a recycling culture through the implementation of social features, such as community leaderboards, rewards for collective achievements, and educational content on the environmental impact of plastic waste.

Mimetic pressure refers to the tendency of individuals or organizations to imitate the behaviors of successful peers or competitors. Mimetic pressure drives imitation of successful practices. Kitole and Sesabo (2024) showed it positively influenced investments in plastics, papers, and e-wastes. The proposed sharing economy model can emulate successful implementations, such as Kabadiwalla Connect in India.

Integrating coercive, normative, and mimetic pressures creates a comprehensive framework for addressing Ghana’s plastic waste challenges. Coercive pressure ensures regulatory compliance, normative pressure fosters community engagement, and mimetic pressure drives innovation adoption. Together, they enable the scalability and sustainability of the sharing economy model, transforming plastic waste into economic opportunities while mitigating environmental harm.

2.5 Sharing Economy

Speaking of the last decade, massive advancement in the internet and mobile technology has given rise to a concept that has gained massive traction and has come to stay, which is referred to as “sharing economy” (Acquier et al., 2017). This idea encapsulates numerous variations of products and services that are distributed, shared and accessed through collaborative and self-reinforcing practices. It is no surprise the sharing economy model can be found in numerous sectors such as transportation, healthcare, food, fashion, telecommunications, construction, etc., (Munoz & Cohen, 2017). Notable examples range from platforms such as Airbnb, Uber, Bolt, BlaBlaCar and so on and so forth.

Hamari et al (2016), has described sharing or collaborative economy as a peer-to-peer-based sharing or accessing goods and services, coordinated through community-based online services (Hamari et al., 2016). Ritter & Schanz (2019) described value proposition, value creation and delivery, and value capture as the fundamental principles that enable sharing economy concepts (Ritter & Schanz, 2019). Value proposition refers to the reasons a customer will value an organization's offering (Osterwalder & Pigneur, 2010). Value creation and delivery are the actions undertaken to create, produce, sell and deliver products or services to customers, whereas value capture are the processes of securing benefits or profits from value creation and the distribution of those profits among the participating actors, in this case the collectors, aggregators, etc. (Sjödin et al., 2020). These concepts define our proposed sharing economy model for PSM

2.6 Application of the Sharing Economy Models in PSM

Kabadiwalla Connect, based in Chennai, India, makes use of a digital platform to formalize the role of informal waste collectors, known as kabadiwalas, by connecting them with households and businesses for efficient waste collection and recycling. This model demonstrates the potential of technology to bridge the gap between informal and formal waste management sectors, which enhances collection rates and creates economic opportunities for waste pickers. Similarly, Rubicon, a U.S.-based waste management company, employs a cloud-based platform to optimize waste collection logistics which links waste generators with haulers and recyclers while emphasizing sustainability and cost-efficiency (Rubicon, 2025). These cases together highlight the power of digital platforms to streamline stakeholder interactions, a principle adapted in our model for Ghana. However, unlike India's e-waste focus or the U.S.'s advanced infrastructure context, our framework tailors these insights to Ghana's plastic waste challenges, which emphasizes affordability, scalability, and integration of the informal sector, which are key considerations given Ghana's resource constraints and reliance on informal waste pickers.

3.0 Methods and Materials

This study adopts a conceptual approach to develop an innovative sharing economy model for monetizing plastic waste in Ghana. A conceptual methodology is particularly appropriate for addressing systemic challenges like waste management, as it allows for the integration of multiple theories, case studies, and empirical insights to propose a comprehensive and adaptable model (Jabareen, 2009). It is also a useful in creating new knowledge by building on carefully selected sources of information (Jabareen, 2009).

We conducted a thorough review of literature and analysis of best practices in waste management and sharing economy applications. To ground the conceptual model in practical insights, we conducted a comparative analysis of two exemplary sharing cases of the application of sharing economy platforms in waste management. Case studies, such as the Rubicon platform in the United States and Kabadiwalla Connect in India, offered practical insights into the integration of digital platforms into PSM systems in Ghana. The conceptual model is based on the Ishikawa fishbone diagram (Ishikawa & Loftus, 1990) was used to categorise the primary causes of inefficient PSM in Ghana and to define the antecedents of the proposed solution.

3.1 Theoretical Foundations

The circular economy framework emphasizes the elimination of waste and the maximization of material lifespans through closed-loop systems. When applied to PSM, the circular economy framework transforms the perception of plastic waste from a disposable byproduct into a valuable resource. It promotes the development of efficient processes for the collection, sorting, and recycling of plastics, ensuring they are reintegrated into the economy rather than discarded. By viewing plastic waste as a resource, the circular economy provides a robust foundation for designing sustainable, closed-loop systems that maximize material value and minimize environmental impact (Hailemariam & Erdiaw-Kwasie, 2023). This theoretical lens is critical for conceptualizing innovative solutions that address the global plastic waste crisis while fostering economic and environmental sustainability. This shift not only reduces environmental harm but also creates economic opportunities by turning waste into raw materials for new products.

Completing the circular economy concept is the concept of monetization which has various dimensions (Najjar & Kettinger, 2013). Things possessing limited intrinsic value or intrinsically worthless items could be monetized or transformed into a legitimate money making venture, provided the holder is capable of providing the services others want, or can add value to such items. In PSM, financial incentives like earning credits for proper waste segregation can act as nudges, motivating households to engage more actively in recycling efforts.

3.2 A Sharing Economy Based PSM Platform (Bolaman)

The platform connects households, waste collectors, recyclers, and businesses, facilitating real-time interactions and transactions. Studies by (Cohen & Kietzmann, 2014) demonstrate that digital platforms are critical enablers of sharing economy models, fostering transparency and efficiency. Households earn credits for segregating and depositing plastics, aligning with the Thaler & Sunstein (2008) principles of behavioral economics. These incentives encourage participation and behavior change, addressing the issue of low recycling rates. AI-powered analytics optimize waste collection routes and predict material demand, ensuring resource efficiency. This approach aligns with the findings of Jabareen (2009), who emphasize the role of large-scale data in unlocking the potential of circular economy models. The model emphasizes collaboration among formal and informal actors, building on insights from network theory. Mayer and Sparrowe (2013) highlight that strong stakeholder networks are critical for achieving systemic change in resource management.

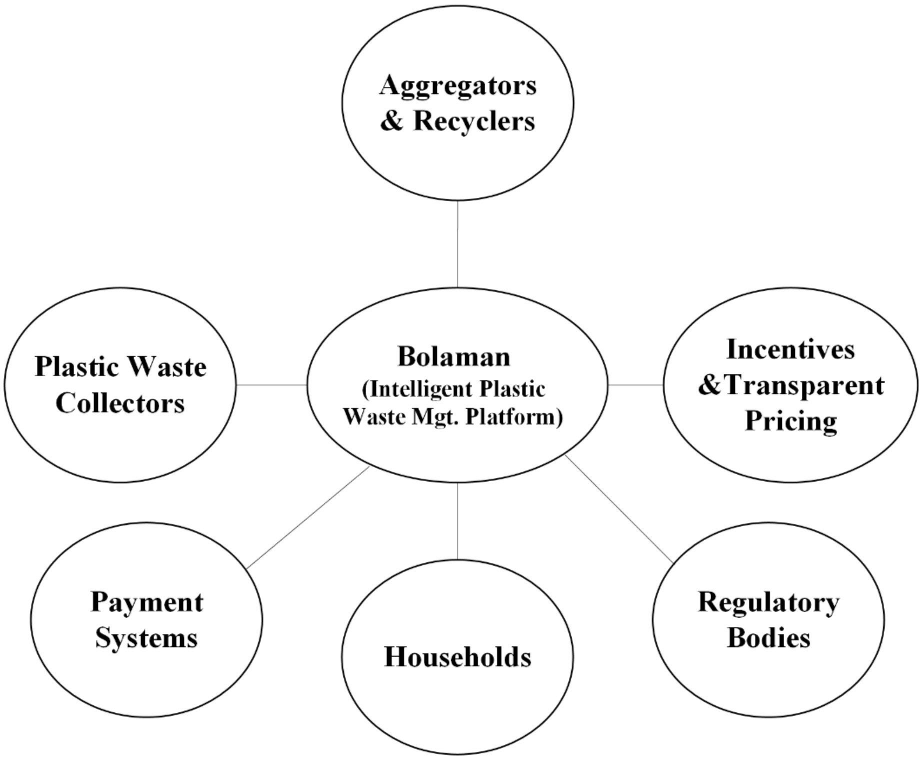


Figure 3: A diagram representing the proposed sharing economy model of Bolaman (Intelligent PSM Platform)

Bolaman is an innovative sharing economy-based platform designed to tackle PSM by fostering collaboration among various stakeholders. The platform connects plastic waste collectors, recyclers, households, and regulatory bodies, creating an efficient ecosystem for managing plastic waste. Waste collectors use intelligent tools to gather and sort plastic materials, which are then processed by recyclers to produce reusable materials. Households are incentivized to participate by segregating and contributing their plastic waste, while regulatory bodies ensure compliance with environmental standards. Bolaman also incorporates a robust payment system to fairly compensate all participants, promoting transparency and encouraging active involvement in the waste management process.

The platform would be built on advanced technologies like IoT and AI to optimize waste collection, sorting, and recycling, enhancing overall efficiency. By promoting community engagement and awareness, the platform would encourage widespread participation in PSM.

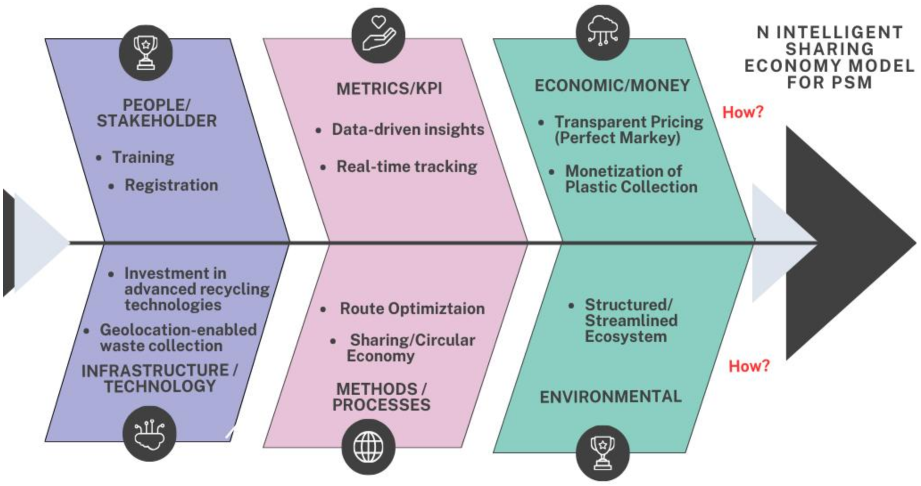

N INTELLIGENT SHARING ECONOMY MODEL FOR PSM
PEOPLE/ STAKEHOLDER
• Training
• Registration
METRICS/KPI
• Data-driven insights
• Real-time tracking
ECONOMIC/MONEY
• Transparent Pricing (Perfect Markey)
• Monetization of Plastic Collection
How?
• Investment in advanced recycling technologies
• Geolocation-enabled waste collection
INFRASTRUCTURE / TECHNOLOGY
• Route Optimiztaion
• Sharing/Circular Economy
METHODS / PROCESSES
• Structured/ Streamlined Ecosystem
ENVIRONMENTAL
How?

Figure 4: Fishbone diagram showing the Intelligent sharing economy platform for PWM

4.0 Discussion

The proposed sharing economy model for monetizing plastic waste in Ghana integrates circular economy, behavioral economics, and network theory to address the country’s plastic waste challenges. The circular economy framework transforms plastic waste into valuable resources, while behavioral economics informs incentive mechanisms like credits and gamification to encourage participation. Network theory ensures efficient stakeholder interactions, connecting households, waste collectors, recyclers, and businesses through digital platforms. AI-powered analytics optimize waste collection routes and predict material demand, enhancing system efficiency. This model not only creates economic opportunities by formalizing the informal sector and reducing production costs, but also mitigates environmental degradation by increasing recycling rates and reducing plastic pollution.

The model also promotes social empowerment by raising awareness, encouraging behavioral change, and integrating informal waste pickers into the formal economy. Collaboration among stakeholders strengthens social cohesion and resilience, while gamification fosters community engagement. Despite challenges such as the need for digital infrastructure and regulatory support, the model's alignment with economic, environmental, and social goals positions it as a transformative solution. By leveraging the sharing economy, Ghana can create a sustainable waste management system that benefits all stakeholders, demonstrating the potential of innovative approaches to address complex environmental challenges.

5.0 Conclusion

This paper has presented an intelligent sharing economy model for the monetization of plastic waste in Ghana, addressing critical gaps in existing waste management practices. By leveraging digital platforms and incentivized participation, the model aligns with sustainable development goals while offering economic and environmental benefits.

This study presents a practical and innovative solution to Ghana's plastic waste crisis. By turning waste into a valuable resource and creating economic opportunities, the model aligns environmental goals with financial incentives. It's a win-win: cleaner communities, healthier ecosystems, and new income streams for informal workers. Future research would focus on designing the system and pilot testing it to assess its feasibility and scalability, providing a foundation for broader implementation. The framework envisions a digital platform that connects households, waste collectors, recyclers, and businesses, incentivizing waste collection and segregation, and facilitating efficient material flows. The study also advances theoretical understanding, offers practical pathways for implementation, and highlights the potential for sharing economy models to transform waste into a valuable resource.

References

Abrokwah, S., Ekumah, B., Adade, R., & Akuoko, I. S. G. (2022). Drivers of single-use plastic waste generation: Lessons from packaged water consumers in Ghana. *GeoJournal*, *87*(4), 2611–2623. https://doi.org/10.1007/s10708-021-10390-w

Acquier, A., Daudigeos, T., & Pinkse, J. (2017). Promises and paradoxes of the sharing economy: An organizing framework. *Technological Forecasting and Social Change*, *125*, 1–10. https://doi.org/10.1016/j.techfore.2017.07.006

Alqattaf, A. (2020). Plastic waste management: Global facts, challenges and solutions. *2020 Second International Sustainability and Resilience Conference: Technology and Innovation in Building Designs (51154)*, 1–7. https://ieeexplore.ieee.org/abstract/document/9319989/

Coccia, M. (2020). Fishbone diagram for technological analysis and foresight. *International Journal of Foresight and Innovation Policy*, *14*(2/3/4), 225. https://doi.org/10.1504/IJFIP.2020.111221

Cohen, B., & Kietzmann, J. (2014). Ride On! Mobility Business Models for the Sharing Economy. *Organization & Environment, 27*(3), 279–296. https://doi.org/10.1177/1086026614546199

Data Commons. (2023). *Ghana—Place Explorer—Data Commons*. https://datacommons.org/place/country/GHA?utm_medium=explore&mprop=count&popt=Person&hl=en

De Feo, G., & Ferrara, C. (2024). Advancing communication in solid waste management: Leveraging life cycle thinking for environmental sustainability. *Environmental Technology Reviews*, *13*(1), 441–460. https://doi.org/10.1080/21622515.2024.2362448

Debrah, J. K., Vidal, D. G., & Dinis, M. A. P. (2021). Innovative use of plastic for a clean and sustainable environmental management: Learning cases from Ghana, Africa. *Urban Science*, *5*(1), 12.

Faisal, F., Rasul, M. G., Jahirul, M. I., & Schaller, D. (2023). Pyrolytic conversion of waste plastics to energy products: A review on yields, properties, and production costs. *Science of The Total Environment, 861*, 160721.

Geyer, R., Jambeck, J. R., & Law, K. L. (2017). Production, use, and fate of all plastics ever made. *Science Advances, 3*(7), e1700782. https://doi.org/10.1126/sciadv.1700782

Ghana Recycling Initiative by Private Enterprises (GRIPE). (2021, January 22). *Circle Economy Foundation*. https://prevent-waste.net/en/ghana-recycling-initiative-by-private-enterprises-gripe/

Global Plastic Action Partnership. (2021). *A Roadmap for Radical Reduction of Plastic Pollution in Ghana* (First Edition). World Economic Forum.

Hailemariam, A., & Erdiaw-Kwasie, M. O. (2023). Towards a circular economy: Implications for emission reduction and environmental sustainability. *Business Strategy and the Environment, 32*(4), 1951–1965. https://doi.org/10.1002/bse.3229

Hamari, J., Sjöklint, M., & Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. *Journal of the Association for Information Science and Technology*, *67*(9), 2047–2059. https://doi.org/10.1002/asi.23552

Ishikawa, K., & Loftus, J. H. (1990). *Introduction to quality control* (Vol. 98). Springer. https://link.springer.com/book/9789401176903

Jabareen, Y. (2009). Building a Conceptual Framework: Philosophy, Definitions, and Procedure. *International Journal of Qualitative Methods, 8*(4), 49–62. https://doi.org/10.1177/160940690900800406

Jambeck, J. R., Geyer, R., Wilcox, C., Siegler, T. R., Perryman, M., Andrady, A., Narayan, R., & Law, K. L. (2015). Plastic waste inputs from land into the ocean. *Science, 347*(6223), 768–771. https://doi.org/10.1126/science.1260352

Kwansa, V. (2021). Ghana's Efforts and Plastic Waste Management Strategies. *Academia Letters*, 2.

Munoz, P., & Cohen, B. (2017). Sustainable Entrepreneurship Research: Taking Stock and Looking Ahead. *Business Strategy and the Environment*, *27*. https://doi.org/10.1002/bse.2000

Najjar, M. S., & Kettinger, W. J. (2013). Data Monetization: Lessons from a Retailer's Journey. *MIS Quarterly Executive, 12*(4).

Okai, D. E. (2020). *Recycling as a strategy for revenue generation and municipal plastic waste management: The case of Accra Metropolitan Area* [PhD Thesis]. https://air.ashesi.edu.gh/items/dfbc8f97-ad10-4402-953c-7a5c1fe46826

Osterwalder, A., & Pigneur, Y. (2010). *Business model generation: A handbook for visionaries, game changers, and challengers*. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=UzuTAwAAQBAJ&oi=fnd&pg=PP1&dq=Osterwalder+and+Pigneur,+2010,+2010,&ots=yZFMFeCc_v&sig=-ZJtAoZWDPiCaQstewIt1LZf1-w

Puschmann, T., & Alt, R. (2016). Sharing economy. *Business & Information Systems Engineering, 58*(1), 93–99.

Quartey, E. T., Tosefa, H., Danquah, K. A. B., & Obrsalova, I. (2015). Theoretical Framework for Plastic Waste Management in Ghana through Extended Producer Responsibility: Case of Sachet Water Waste. *International Journal of Environmental Research and Public Health*, *12*(8), Article 8. https://doi.org/10.3390/ijerph120809907

Ritter, M., & Schanz, H. (2019). The sharing economy: A comprehensive business model framework. *Journal of Cleaner Production*, *213*, 320–331.

Rubicon. (2025). Rubicon | Software for Smart Waste and Recycling Solutions. *Rubicon | Software for Smart Waste and Recycling Solutions*. https://www.rubicon.com/

Sjödin, D., Parida, V., Jovanovic, M., & Visnjic, I. (2020). Value Creation and Value Capture Alignment in Business Model Innovation: A Process View on Outcome-Based Business Models. *Journal of Product Innovation Management*, *37*(2), 158–183. https://doi.org/10.1111/jpim.12516

Smith, R. L., Takkellapati, S., & Riegerix, R. C. (2022). Recycling of Plastics in the United States: Plastic Material Flows and Polyethylene Terephthalate (PET) Recycling Processes. *ACS Sustainable Chemistry & Engineering*, *10*(6), 2084–2096. https://doi.org/10.1021/acssuschemeng.1c06845

Tahiru, A., Cobbina, S. J., & Asare, W. (2024). *Challenges and Opportunities for Waste-to-Energy Integration in Tamale's Waste Management System*. https://www.preprints.org/manuscript/202405.1923

Thaler, R., & Sunstein, C. (2008). Nudge: Improving decisions about health, wealth and happiness. *Amsterdam Law Forum; HeinOnline: Online*, 89. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/amslawf1&section=49

Ugwu, O. S., & Godwin-Okoubi, M. L. O. (n.d.). *Consumer Analysis for the Use of Bottled Packaged Soft Drinks to Plastic Packaged Soft Drinks in Eke Market Afikpo in Ebonyi State, Nigeria*. Retrieved March 10, 2025, from http://arcnjournals.org/images/277514562112016.pdf

UNDP. (2024). *Our planet is choking on plastic*. UNDP Environment Programme. http://unep.org/interactive/beat-plastic-pollution/

Wardrop, N. A., Dzodzomenyo, M., Aryeetey, G., Hill, A. G., Bain, R. E., & Wright, J. (2017). Estimation of packaged water consumption and associated plastic waste production from household budget surveys. *Environmental Research Letters*, *12*(7), 074029.

WEF. (2023, May 31). *Data benefits Ghana's fight against plastic pollution*. World Economic Forum. https://www.weforum.org/impact/data-benefits-ghana-fight-against-plastic-pollution/

Generative AI-Induced Synthetic Data: Explicating the Ethical Implications

Joseph Kwame Adjei (Computer Science & Information Systems Department, Ashesi University), Henry Owusu (Computer Science & Information Systems Department, Ashesi University) and Clifford Yeboah (Computer Science & Information Systems Department, Ashesi University)

Completed Research

Abstract

Generative AI has revolutionized the creation of synthetic data, offering scalable and privacy-preserving solutions for data augmentation, testing, and analytics. However, the growing adoption of generative AI technologies raises critical ethical questions, including biases in generated data, misuse risks, accountability gaps, and potential erosion of trust. This systematic literature review employs Chitu Okoli's method to synthesize the ethical implications of generative AI-induced synthetic data. By analyzing peer-reviewed articles, industry reports, and guidelines, this study categorizes the key ethical concerns, evaluates existing mitigation strategies, and identifies research gaps. The findings contribute to ethical AI discourse by highlighting challenges and proposing avenues for developing responsible generative AI applications. This work provides valuable insights for researchers, practitioners, and policymakers seeking to balance innovation with ethical integrity.

Keywords: Synthetic Data, Artificial Intelligence, Ethics, Generative AI, Bias in data.

1.0 Introduction

Generative Artificial Intelligence (AI) is increasingly transforming the way online content is created and consumed, and the fabrication and use of synthetic data, offering significant data accessibility, and privacy preservation. Synthetic data, which mimics the statistical properties of real-world datasets, is increasingly used in diverse fields, including healthcare, finance, autonomous systems, and marketing. This technology addresses key challenges associated with real-world data, such as data scarcity, privacy regulations, and ethical issues surrounding the use of sensitive information. Generative AI helps businesses develop while protecting privacy and lowering the risks involved in working with actual data by creating realistic looking but artificial datasets.

The appeal of synthetic data lies in its ability to provide robust solutions for various real-world challenges. For example, in healthcare, synthetic data facilitates the development of machine learning models without exposing sensitive patient information, thereby complying with privacy regulations such as HIPAA and GDPR (Beaulieu-Jones et al., 2019). Similarly, in the financial sector, synthetic data is used to simulate trading environments, stress-test algorithms, and enhance fraud detection systems, contributing to improved decision-making and risk management (Patki et al., 2016). These benefits underscore synthetic data’s role as a catalyst for data-driven innovation and efficiency.

Despite its potential, the use of generative AI to produce synthetic data raises significant ethical concerns. Bias in training data can lead to synthetic datasets that perpetuate societal inequities, particularly in high-stakes systems like criminal justice, hiring, or healthcare (Mehrabi et al., 2021). In high-stakes systems like criminal justice, jobs, or healthcare, bias in training data can result in synthetic datasets that reinforce social injustices (Mehrabi et al., 2021). Furthermore, there are risks to public confidence and social stability from the possible malevolent use of synthetic data for things like creating deepfakes or disseminating false information (Chesney & Citron, 2019a). These dangers show that in order to guarantee the appropriate application of generative AI technology, strict ethical review and strong governance procedures are required.

Accountability in generative AI systems further complicates ethical oversight. Since many generative artificial intelligence models function as “black boxes,” it can be difficult to track down decision-making procedures or attribute responsibility for ethical harms (Floridi & Cowls, 2022). This opacity erodes stakeholder trust and makes it more difficult for regulators to create precise rules for the moral application of synthetic data.

Although the ethical implications of synthetic data have gained attention in recent years, the existing literature is often fragmented, addressing isolated concerns without providing a comprehensive analysis of the broader ethical landscape. This study seeks to address this gap by synthesizing the main issues surrounding generative AI-generated synthetic data, categorizing them into major themes such as bias, misuse risks, accountability, and trust. Furthermore, the study evaluates current strategies for mitigating these ethical concerns and identifies critical areas for future research.

This work contributes to the conversation around ethical AI by providing researchers, practitioners, and policymakers with useful insights. It emphasizes the need for interdisciplinary approaches to addressing ethical challenges and proposes a framework for aligning technological innovation with societal values. By situating the discussion within a broader socio-technical context, this study lays the groundwork for the ethical creation and application of generative AI technology.

2.0 Literature Review

The literature review synthesizes existing knowledge on generative AI-induced synthetic data and its ethical implications. It explores the key themes of bias, misuse risks, accountability, and trust, providing a comprehensive analysis of current research. This section also categorizes reviewed papers to highlight contributions and research gaps.

2.1 Method

This systematic literature review (SLR) was designed to rigorously synthesize existing knowledge on the ethical effects of synthetic data produced by generative AI. The methodology followed a structured process encompassing planning, literature selection, and data synthesis to ensure replicability and reliability. The primary objective was to identify key ethical themes, evaluate mitigation strategies, and uncover research gaps, guided by the research question: *What are the ethical implications of generative AI-induced synthetic data, and how can they be addressed effectively?*

The literature search involved comprehensive database queries across Scopus, IEEE Xplore, SpringerLink, and Web of Science. Keywords such as “generative AI,” “synthetic data,” “ethical implications,” “bias in AI,” “accountability in AI,” and “trust in AI” were used to retrieve relevant studies. Articles were included if they were peer-reviewed or high-quality industry reports published between 2017 and 2024. As depicted in **Figure 1**, from an initial pool of 658 articles, 112 were retained after title and abstract screening. A full-text review further refined the selection to 62 articles that directly addressed the ethical dimensions of generative AI-induced synthetic data.

Data analysis involved extracting and coding information to identify recurring themes and categories. An inductive approach grouped findings into four key themes: bias, misuse risks, accountability, and trust. Mitigation strategies and research gaps were also systematically documented. This structured approach ensured a comprehensive synthesis of the ethical landscape surrounding generative AI-induced synthetic data, laying a robust foundation for discussion and future research.

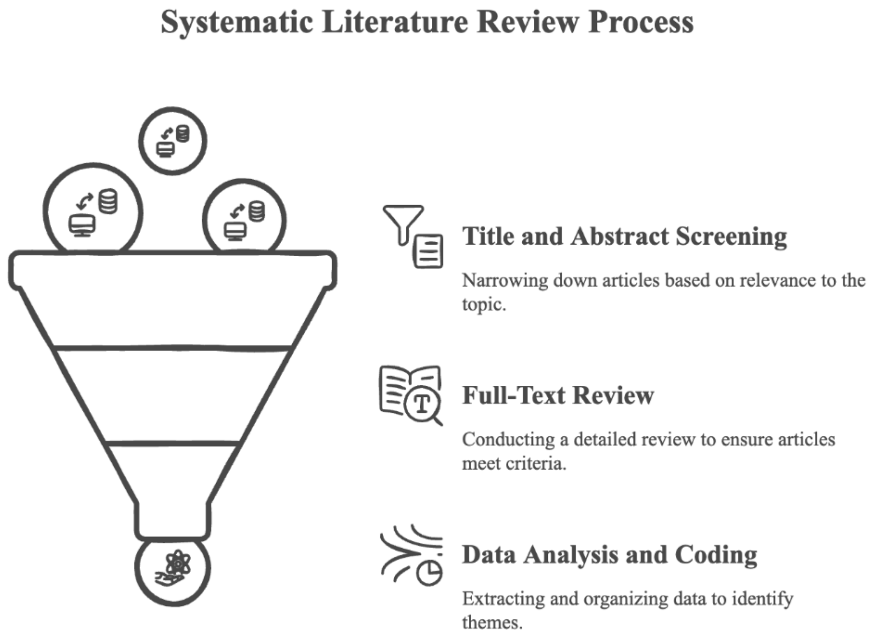

Systematic Literature Review Process
Title and Abstract Screening
Narrowing down articles based on relevance to the topic.
Full-Text Review
Conducting a detailed review to ensure articles meet criteria.
Data Analysis and Coding
Extracting and organizing data to identify themes.

Figure 1. Systematic Literature Review Process

2.2 Overview of Generative AI and Synthetic Data

To create realistic synthetic data, generative AI uses sophisticated models like variational autoencoders (VAEs), diffusion models, and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). Synthetic data replicates the statistical properties of real datasets without exposing sensitive information, making it a valuable tool for privacy-preserving applications (Beaulieu-Jones et al., 2019). Key applications include healthcare. According to Chen et al., (2021), synthetic data facilitates healthcare research without violating privacy laws, enabling machine learning models to generalize well in clinical settings (Chen et al., 2021). It has also been applied to simulate financial scenarios, improve fraud detection systems, and enhance algorithmic trading strategies (Patki et al., 2016). Similarly, synthetic data also support the training of autonomous vehicles by simulating diverse driving scenarios (Dosovitskiy et al., 2017).

2.3 Ethical Implications of Generative AI-Induced Synthetic Data

Bias and Fairness

Generative AI systems derive their behavior from the data they are trained on, which often contains inherent biases. These biases can inadvertently transfer into the synthetic data they generate, perpetuating or even amplifying systemic inequities. For example, Mehrabi et al., (2021) illustrate how inequalities in artificial data could deepen disparities in critical domains such as hiring, lending, and criminal justice, disproportionately affecting underrepresented or marginalized groups. Efforts to mitigate such biases are ongoing, with researchers developing fairness-aware algorithms designed to balance equity and utility. However, challenges remain in striking a balance between reducing bias and preserving the utility or fidelity of the generated data. According to Dwork et al., (2012), while methods like differential privacy can address privacy concerns, their interplay with fairness objectives often creates complex trade-offs that require careful management.

Misuse Risks

The misuse of synthetic data for malicious purposes is a growing concern, with applications such as deepfakes, fabricated media, and misinformation campaigns highlighting the potential societal risks. (Chesney & Citron, 2019b) emphasize the importance of addressing these challenges through the implementation of verification mechanisms to detect and prevent harm caused by synthetic data misuse. For instance, malicious actors have exploited synthetic data to fabricate identities or spread disinformation, eroding trust in digital ecosystems. Recent studies, such as Pan et al., (2023) suggest that incorporating provenance tracking – a system for documenting the origin and lineage of data – could significantly enhance the authenticity and traceability of synthetic datasets, thereby deterring misuse.

Accountability and Governance

The complexity and opacity of generative AI systems pose significant challenges to accountability. Determining responsibility for ethical harms caused by synthetic data is complicated by the lack of transparency in how these systems function. In order to help users and regulators better understand how decisions are produced and potential hazards, Floridi and Cowls, (2022) support incorporating explainability and transparency into the design of generative AI systems. However, the regulatory landscape for synthetic data remains underdeveloped. Without robust governance frameworks, gaps persist in areas such as ethical oversight, data ownership, and responsibility allocation. Policymakers are urged to prioritize the creation of comprehensive regulations to address these deficiencies and align synthetic data use with societal values.

Trust and Public Perception

Building public trust in synthetic data hinges on the responsible communication of its benefits and risks, as well as the implementation of robust mechanisms to ensure ethical practices. Binns, (2018) stresses the importance of clear, transparent communication to demystify synthetic data and address public scepticism. Many individuals are unaware of synthetic data's capabilities, leading to fears of manipulation or loss of control. Public education campaigns, alongside ethical governance initiatives, could help reduce resistance and foster greater acceptance. Furthermore, trust can be bolstered by providing assurances about the ethical stewardship of synthetic data, such as through the establishment of independent oversight bodies or certifications. These steps are essential to ensuring that synthetic data fulfils its potential as a transformative tool while minimizing ethical and societal risks.

2.4 Mitigation Strategies

Fairness-aware algorithms play a crucial role in reducing biases within synthetic data while maintaining its statistical utility. These methods are essential to ensure that synthetic data does not exacerbate existing inequities, particularly in sensitive domains such as healthcare and finance. Researchers like (Dwork et al., 2012) and Mehrabi et al. (2021) have explored techniques that embed fairness into data generation processes, ensuring equitable outcomes without compromising the usability of the data for analytical purposes.

Transparency tools, such as explainable AI (XAI), enhance the interpretability of synthetic data generation processes. By making these processes more understandable, stakeholders can better assess the reliability and ethical implications of the data. Doshi-Velez & Kim (2017), emphasize the importance of XAI in fostering trust and accountability in AI systems, which is particularly relevant in the context of synthetic data (Doshi-Velez & Kim, 2017).

Provenance tracking provides a mechanism for verifying the authenticity of synthetic data by tracing its origins and transformations throughout the generation pipeline. This ensures that the data is credible and has not been tampered with or misused. Pan et al. (2023), highlight the importance of provenance tracking as a safeguard against ethical and security vulnerabilities in synthetic data applications (Pan et al., 2023).

Finally, regulatory initiatives are critical for establishing ethical guidelines and policies for the generation and use of synthetic data. Such frameworks, as proposed by

Floridi and Cowls (2022), are necessary to govern the responsible application of synthetic data technologies. These initiatives ensure that synthetic data aligns with societal values and ethical principles, fostering greater trust and accountability in its usage.

2.5 Categorization of Reviewed Papers

The table below categorizes the reviewed papers based on their focus areas, methodologies, and key contributions.

Category	**Paper**	**Focus Area**	**Key Contributions**
Generative AI Models	(Goodfellow et al., 2014)	GANs and synthetic data generation	Introduced GANs, foundational for realistic synthetic data generation.
Healthcare Applications	Beaulieu-Jones et al. (2019)	Privacy-preserving healthcare data	Demonstrated synthetic data use in clinical research without compromising privacy.
Bias and Fairness	Mehrabi et al. (2021)	Bias in generative AI models	Explored how biases in training data propagate to synthetic datasets and mitigation strategies.
Misuse Risks	Chesney & Citron (2019)	Deepfakes and misinformation	Highlighted risks of misuse, including deepfakes, and proposed verification mechanisms.
Accountability	Floridi et al. (2018)	Governance and explainability	Proposed principles for embedding transparency and accountability in AI systems.
Transparency	Doshi-Velez & Kim (2017)	Explainable AI (XAI)	Advocated for explainability as a core principle to enhance trust in AI systems.
Provenance Tracking	Pan et al., (2023)	Data authenticity and provenance	Proposed mechanisms for ensuring synthetic data authenticity and traceability.
Fairness in AI	Dwork et al. (2012)	Fairness-aware algorithms	Introduced fairness metrics and algorithms for bias

			mitigation in machine learning systems.

Table 1 - Categorisation of Reviewed Papers

2.6 Research Gaps

As depicted in **Figure 2**, the long-term societal impacts of synthetic data adoption remain largely unexplored. While synthetic data offers immediate benefits, such as enhanced privacy and accessibility, its broader implications on social structures, labor markets, and decision-making processes are unclear. There is a lack of longitudinal studies examining how reliance on synthetic datasets might influence systemic inequities or societal trust in digital ecosystems. Addressing this gap requires interdisciplinary research to understand and mitigate potential negative consequences.

Global regulatory frameworks for synthetic data generation and use are underdeveloped and fragmented. This lack of standardization creates inconsistencies in ethical guidelines and accountability mechanisms, particularly in high-stakes domains like healthcare and finance. Disparities in legal oversight also hinder international collaboration and leave room for regulatory loopholes, such as misuse of synthetic data for malicious purposes. Establishing coordinated global policies is essential for fostering equitable and ethical synthetic data practices.

Public perception of synthetic data is another critical yet underexplored area. Trust in synthetic data systems is influenced by factors such as transparency, explainability, and ethical assurances, but empirical studies investigating these dynamics are scarce. There is also limited understanding of public awareness regarding synthetic data's capabilities and risks. Research in this area is vital to align synthetic data technologies with societal expectations and build confidence among users.

Finally, the ethical concerns surrounding emerging applications of synthetic data, such as synthetic biology, education, and creative industries, remain inadequately studied. These fields introduce unique challenges, including issues of biosecurity, fairness, intellectual property, and authenticity. Tailored ethical frameworks are needed to address the distinct risks and opportunities in these domains, ensuring responsible and innovative use of synthetic data across sectors.

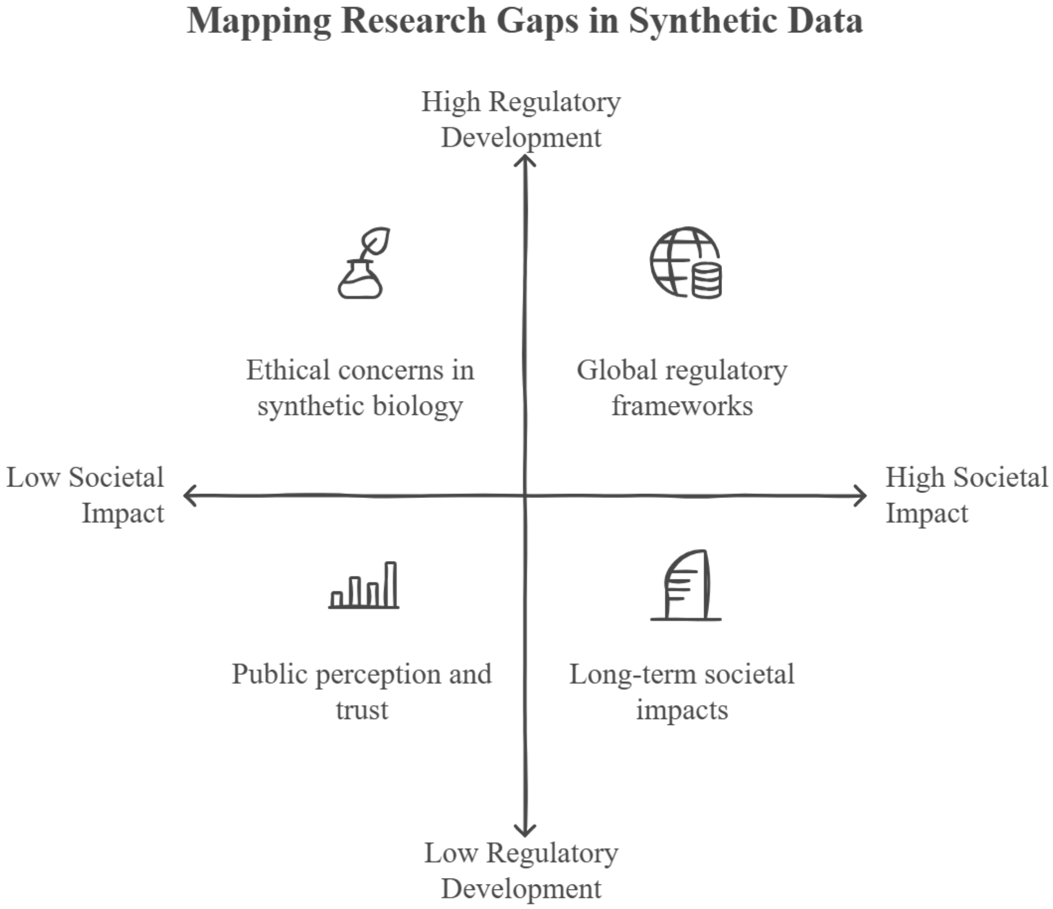

Mapping Research Gaps in Synthetic Data
High Regulatory Development
Ethical concerns in synthetic biology
Global regulatory frameworks
Low Societal Impact
High Societal Impact
Public perception and trust
Long-term societal impacts
Low Regulatory Development

Figure 2. **Mapping Research Gaps in Synthetic Data**

3.0 Discussion

3.1 Bias and Fairness

Bias in synthetic data continues to be a significant concern, especially when such data is deployed in high-stakes domains like healthcare, criminal justice, and hiring. The role of fairness-aware algorithms is pivotal in addressing this issue. Current methods, while promising, often lack generalizability, requiring domain-specific adaptations. For example, in healthcare, (Mehrabi et al., 2021) argue that biases in synthetic datasets could worsen health disparities if left unchecked, such as by underrepresenting minority populations in diagnostic training data. Addressing these challenges necessitates interdisciplinary collaboration among AI researchers, ethicists, domain experts, and policymakers. Such collaboration ensures that fairness interventions are not only technically sound but also aligned with societal and ethical standards. Moreover, adopting frameworks like algorithmic impact assessments could provide systematic evaluations of fairness across diverse applications.

3.2 Misuse Risks

The misuse of synthetic data, particularly for creating deepfakes and facilitating misinformation campaigns, poses serious societal risks. Deepfakes, for instance, have been exploited to manipulate public opinion and fabricate evidence, highlighting the urgent need for robust detection and verification systems. Chesney & Citron, (2019a) recommend a multi-pronged approach to mitigate these risks, including the development of AI-powered verification tools, public awareness campaigns to improve digital literacy, and regulatory measures to set boundaries on the use of synthetic data technologies. However, finding the right balance between fostering innovation and imposing restrictions remains a challenge. Over-regulation could stifle innovation, while under-regulation may allow harmful misuse to proliferate. Policymakers must, therefore, consider nuanced and adaptive governance models that evolve alongside advancements in generative AI.

3.3 Accountability and Governance

The opaque nature of many generative AI algorithms makes accountability a pressing concern. Determining who is responsible for ethical harms caused by synthetic data is often challenging, particularly when decisions are based on highly complex or black-box models. Floridi & Cowls (2022), advocate embedding explainability and traceability into generative AI processes as part of a broader push for ethical compliance. This aligns with the concept of "ethics by design," which emphasizes the proactive integration of ethical considerations during AI development rather than treating them as an afterthought. Traceability, for instance, could involve creating data provenance systems that log the origins and transformations of synthetic data. Additionally, governance frameworks should include independent oversight and mechanisms for redress to address ethical lapses effectively.

3.4 Trust and Public Perception

Public trust is essential for the broad acceptance and ethical use of synthetic data. However, scepticism and fear often arise from a lack of understanding of its potential benefits and risks. Binns, (2018) emphasizes the need for transparency and education to bridge this trust gap. Educational initiatives could include public workshops, interactive tools, and media campaigns to demystify synthetic data. Moreover, the development of user-friendly explainability tools could empower users to understand how synthetic data is generated and used. Ethical certifications or standards for synthetic data could also play a role in enhancing trust by signaling adherence to best practices. Such measures not only build confidence among end-users but also encourage responsible behavior among developers and organizations (Olateju et al., 2024).

3.5 Toward a Responsible Framework

Synthesizing the findings across the dimensions of bias, misuse risks, accountability, and trust underscores the urgent need for a comprehensive ethical framework to guide the responsible use of generative AI-induced synthetic data (Eacersall et al., 2024). Such a framework must address the unique ethical complexities associated with synthetic data while balancing innovation with accountability (Hao et al., 2024). As shown in **Figure 3**, to be effective, the framework should integrate technical, regulatory, and educational strategies that work cohesively to tackle challenges at multiple levels.

Technical Strategies

Technical solutions form the backbone of any ethical framework. They ensure that synthetic data systems operate with integrity, transparency, and accountability. The growing use of synthetic data requires robust mechanisms to address ethical concerns such as bias, explainability, and provenance. Fairness-aware algorithms detect and mitigate biases in synthetic data generation (Lepri et al., 2018, 2021). These algorithms help maintain equitable representation in training datasets, especially in domains like healthcare and finance where biases have serious consequences. Fairness-aware methods support ethical decision-making and reduce discriminatory outcomes. Explainability tools allow stakeholders to understand the process behind synthetic data generation. User-friendly features like visual tools show how data is generated, increasing trust and accountability. Provenance tracking systems record the origins, transformations, and applications of synthetic data. They maintain a transparent chain of custody that upholds data authenticity and prevents misuse. Blockchain technology can create immutable records of synthetic data provenance, ensuring traceability and reliability. Integrating fairness-aware algorithms, explainability tools, and provenance systems upholds ethical standards while enabling innovation. Technical solutions like these are essential for responsible AI development and long-term trust in data-driven systems.

Regulatory Strategies

Robust regulatory measures are crucial for setting clear boundaries and standards for the ethical use of synthetic data. Adaptive policies require regulatory frameworks that evolve alongside advancements in generative AI technology, addressing both current and emerging ethical challenges with laws mandating disclosure in sensitive areas like journalism or healthcare. Independent oversight bodies composed of interdisciplinary experts from AI, law, ethics, and sociology should monitor and enforce compliance with ethical standards and include grievance redressal systems for addressing lapses. International collaboration is essential as synthetic data systems operate across borders, with harmonized global standards preventing regulatory arbitrage and initiatives like global data ethics councils playing a pivotal role.

Educational Strategies

Education and awareness are indispensable for fostering public trust and ensuring the responsible use of synthetic data. Public awareness campaigns should demystify synthetic data and generative AI for the general public by explaining their benefits and risks using accessible language and examples to reduce skepticism and promote informed discourse. Digital literacy programs that enhance understanding among stakeholders – including end-users, professionals, and policymakers – can empower them to engage with synthetic data technologies responsibly, such as training journalists to identify and report deepfakes or educating developers on fairness-aware algorithms. Transparency initiatives like open-access repositories of synthetic data and detailed documentation of AI models help build confidence by providing clear guidelines on how synthetic data is generated, validated, and applied.

An Integrated Framework

A truly responsible framework for generative AI-induced synthetic data must seamlessly integrate technical interventions to ensure that the underlying systems are robust, reliable, and fair, regulatory measures to provide the necessary guardrails to prevent misuse and hold stakeholders accountable and educational initiatives to promote understanding and trust, empowering society to engage with synthetic data in meaningful ways.

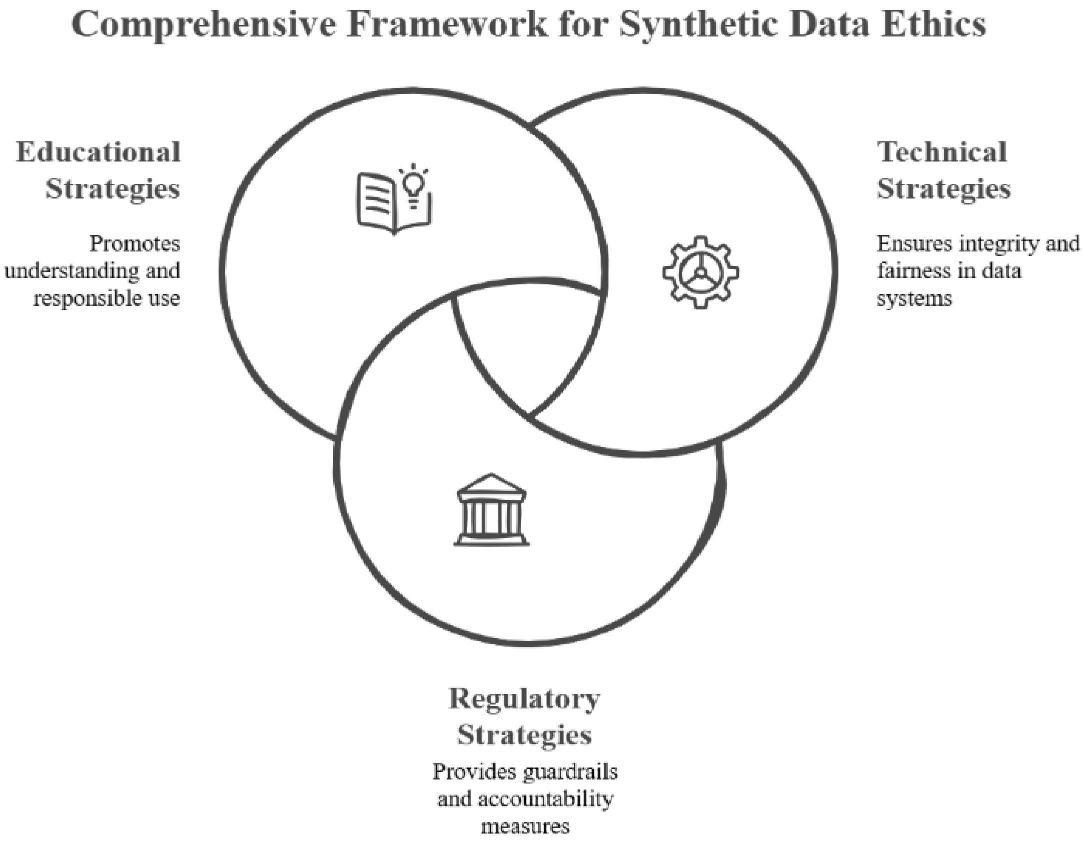

Comprehensive Framework for Synthetic Data Ethics
Educational Strategies
Promotes understanding and responsible use
Technical Strategies
Ensures integrity and fairness in data systems
Regulatory Strategies
Provides guardrails and accountability measures

Figure 3. Comprehensive Framework for Synthetic Data Ethics

3.6 Ethical Framework for Generative AI

Figure 4 presents a comprehensive ethical framework for generative AI-induced synthetic data, designed to ensure that the deployment and application of AI systems align with the values and principles that govern responsible innovation. This framework breaks down the ethical considerations into five key components that work synergistically to address the multifaceted challenges inherent in AI technologies. These components are interconnected and together provide a holistic approach to ensuring the ethical use of AI-driven synthetic data:

The ethical tenets that underpin the framework direct all facets of the creation and application of generative AI. These fundamental ideas, fairness, accountability, transparency, and privacy ensure that AI systems are developed and used in a way that respects social norms and human dignity. Fairness aims to prevent bias and ensure equal treatment across different demographic groups, while accountability establishes responsibility for the outcomes generated by AI systems. Transparency fosters openness about how AI models function and make decisions, allowing for greater trust and understanding. Privacy safeguards sensitive data and ensures that AI systems respect individuals' rights to confidentiality and personal security. The ethical principles provide a moral compass, ensuring that all AI systems, particularly those generating synthetic data, align with broader social values, such as justice, respect for human rights, and equality.

Risk assessment focuses on identifying potential risks connected to the creation, use, and consequences of generative AI technologies. These risks include biases embedded in AI models, the potential for misuse, and vulnerabilities that could lead to data breaches or harm. It highlights the necessity of continuously assessing AI systems in order to identify hazards early in their lifecycle. Through rigorous risk assessment, developers and organizations can anticipate and mitigate harmful effects, such as discrimination or the generation of misleading data, which could undermine public trust and perpetuate inequalities. This component encourages a proactive approach to risk management, promoting continuous improvement of AI models and strategies to prevent undesirable outcomes, such as reinforcing existing social biases or enabling malicious activities.

Governance mechanisms provide the framework within which AI systems must operate to ensure adherence to ethical principles. These mechanisms encompass policies, laws, regulations, and standards that ensure AI systems are held accountable for their actions and outputs. They include regulatory compliance requirements that align with national and international legal standards, as well as industry best practices that guide the development and use of generative AI. Governance also involves organizational policies designed to ensure internal accountability, such as oversight committees, audits, and ethical review boards. These structures help ensure that AI systems are developed and deployed in ways that align with the agreed-upon ethical standards, minimizing the risks of unethical behaviors or consequences.

The ethical framework recognizes the involvement of diverse stakeholders who collectively shape the development and oversight of AI systems. These stakeholders include AI developers, researchers, corporations, regulators, policymakers, advocacy groups, and the general public. Each group has a distinct role in the AI ecosystem and a responsibility to ensure the ethical use of synthetic data. Developers and researchers create the AI models and are responsible for integrating ethical principles into their designs. Policymakers and regulators set legal boundaries and frameworks that ensure AI systems operate within safe and ethical parameters. Advocacy groups help raise awareness of the social implications of AI, ensuring that marginalized or vulnerable populations are considered in the development of AI technologies. Lastly, the public serves as both consumers and critics, holding organizations accountable for the real-world impacts of AI. By promoting collaboration among all stakeholders, the framework ensures that generative AI systems are continuously scrutinized, improved, and adjusted based on diverse perspectives and expert opinions.

Effective implementation strategies are critical to operationalizing the ethical framework and ensuring its long-term success. These strategies involve coordinated efforts to integrate ethical principles into every phase of AI development and deployment. They include fostering collaboration among stakeholders, ensuring that diverse viewpoints are considered, and building systems for monitoring and evaluating AI models in real-world settings. Additionally, implementation strategies emphasize education and training on ethical practices for those involved in the development and use of AI. This includes not only technical training on how to design ethical AI systems but also educating the broader public about the implications of generative AI and how to engage with these technologies responsibly. Transparent reporting and open communication about AI processes and decisions are essential to building trust with stakeholders and ensuring that AI systems are continuously refined in line with ethical standards.

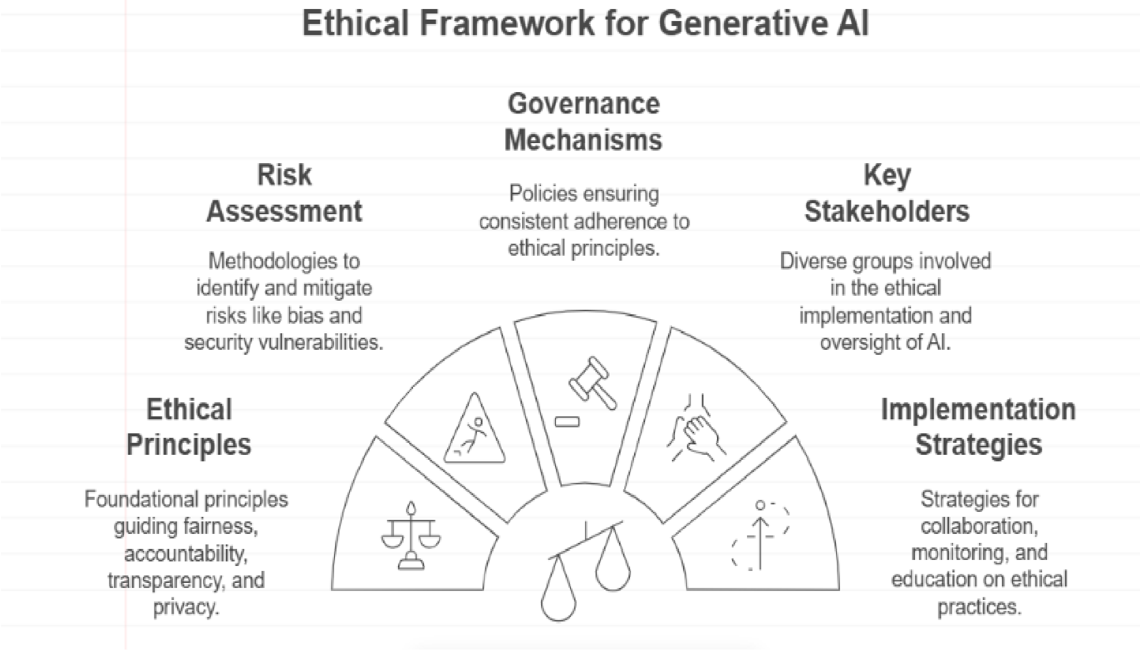

Ethical Framework for Generative AI
Governance Mechanisms
Policies ensuring consistent adherence to ethical principles.
Risk Assessment
Methodologies to identify and mitigate risks like bias and security vulnerabilities.
Key Stakeholders
Diverse groups involved in the ethical implementation and oversight of AI.
Ethical Principles
Foundational principles guiding fairness, accountability, transparency, and privacy.
Implementation Strategies
Strategies for collaboration, monitoring, and education on ethical practices.

Figure 4. Ethical Framework for Generative AI

3.7 Summary of Contributions

This study makes several significant contributions to the discourse on generative AI and synthetic data, addressing the ethical complexities that arise from its development and application. One of the key contributions of this study is its comprehensive ethical categorization of the challenges associated with generative AI-induced synthetic data. By synthesizing and organizing the ethical implications into four critical areas – bias and fairness, misuse risks, accountability and governance, and trust and public perception – the study offers a structured framework for understanding the ethical landscape. This categorization not only facilitates a deeper comprehension of these issues but also serves as a foundation for future research and policymaking, providing a roadmap for addressing these concerns systematically.

Another notable contribution is the study's critical evaluation of existing mitigation strategies aimed at addressing the challenges posed by synthetic data.

Through an analysis of current ethical frameworks and technical interventions, this research highlights both their effectiveness and limitations. For example, fairness-aware algorithms, while effective in some domains, require domain-specific adaptations to ensure efficacy. Similarly, provenance tracking and detection tools show promise in mitigating misuse risks but face challenges in scalability and implementation. This evaluation offers a roadmap for enhancing ethical practices, emphasizing areas where existing strategies need improvement and where new solutions must be developed.

The study also identifies critical gaps in the existing body of knowledge, particularly in areas that have been underexplored or overlooked. For instance, while much research focuses on immediate technical challenges, there is a lack of comprehensive studies on the long-term societal impacts of synthetic data adoption. Similarly, the fragmented state of regulatory oversight across different jurisdictions poses a significant challenge, as there is no unified global framework to address the ethical and legal complexities of generative AI systems. By bringing attention to these gaps, the study underscores the need for interdisciplinary collaboration and the development of robust frameworks that can address emerging ethical concerns effectively.

Lastly, this research provides actionable insights for a wide range of stakeholders, including researchers, practitioners, and policymakers. For researchers, it highlights the importance of developing fairness-aware algorithms and scalable verification tools. For practitioners, the study emphasizes the need for ethical design practices and transparency in deploying generative AI systems. Policymakers, on the other hand, are encouraged to prioritize the creation of adaptive and inclusive regulatory frameworks that balance innovation with ethical accountability. By bridging theoretical understanding with practical recommendations, this study aims to guide the ethical and responsible integration of synthetic data into diverse real-world applications, fostering sustainable and inclusive technological innovation.

4.0 Conclusion

Generative AI-induced synthetic data holds immense potential to transform industries, from healthcare and finance to entertainment and education. However, its transformative capabilities are accompanied by major ethical issues, including issues of inequality, misuse, responsibility, and public trust. This study offers a systematic exploration of these challenges through a literature-based analysis, categorizing them into four primary areas and evaluating existing mitigation strategies.

By synthesizing insights from recent and reliable literature, the research provides actionable recommendations for stakeholders aiming to develop and deploy responsible generative AI systems. For instance, embedding explainability and fairness-aware algorithms into generative AI processes can enhance accountability and reduce bias. Similarly, implementing provenance tracking and verification tools can mitigate misuse risks, while public education campaigns can foster trust and improve societal understanding of synthetic data.

In addition to addressing current challenges, the study identifies critical research gaps, such as the lack of empirical studies on public perception, the absence of comprehensive global regulatory frameworks, and the need to explore the long-term societal impacts of synthetic data adoption. Addressing these gaps will require a collaborative, interdisciplinary approach, drawing on expertise from fields such as AI, ethics, law, and sociology.

As generative AI advances, maintaining ethical integrity in its applications will be crucial to maximize its benefits and lowering any risks. By prioritizing the development of ethical frameworks, inclusive policies, and educational initiatives, stakeholders can guide the responsible adoption of synthetic data, paving the way for sustainable technological innovation that aligns with societal values and priorities. Future research must focus on creating a global dialogue around these issues, fostering shared accountability, and ensuring that generative AI systems serve as tools for positive, equitable transformation.

References

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., & Greene, C. S. (2019). Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. *Circulation: Cardiovascular Quality and Outcomes*, *12*(7), e005122. https://doi.org/10.1161/CIRCOUTCOMES.118.005122

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 149–159. https://proceedings.mlr.press/v81/binns18a.html

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), 493–497. https://doi.org/10.1038/s41551-021-00751-8

Chesney, R., & Citron, D. (2019a). Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics. *Foreign Affairs*, *98*, 147.

Chesney, R., & Citron, D. (2019b). Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics. *Foreign Affairs*, *98*, 147.

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning* (No. arXiv:1702.08608). arXiv. https://doi.org/10.48550/arXiv.1702.08608

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An Open Urban Driving Simulator. *Proceedings of the 1st Annual Conference on Robot Learning*, 1–16. https://proceedings.mlr.press/v78/dosovitskiy17a.html

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

Eacersall, D., Pretorius, L., Smirnov, I., Spray, E., Illingworth, S., Chugh, R., Strydom, S., Stratton-Maher, D., Simmons, J., Jennings, I., Roux, R., Kamrowski, R., Downie, A., Thong, C. L., & Howell, K. A. (2024). *Navigating Ethical Challenges in Generative AI-Enhanced Research: The ETHICAL Framework for Responsible Generative AI Use* (No. arXiv:2501.09021). arXiv. https://doi.org/10.48550/arXiv.2501.09021

Floridi, L., & Cowls, J. (2022). A Unified Framework of Five Principles for AI in Society. In *Machine Learning and the City* (pp. 535–545). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119815075.ch45

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, *27*. https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., Zhou, Z., & Tang, H. (2024). *Synthetic Data in AI: Challenges, Applications, and Ethical Implications* (No. arXiv:2401.01629). arXiv. https://doi.org/10.48550/arXiv.2401.01629

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy & Technology*, *31*(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

Lepri, B., Oliver, N., & Pentland, A. (2021). Ethical machines: The human-centric use of artificial intelligence. *iScience*, *24*(3), 102249. https://doi.org/10.1016/j.isci.2021.102249

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, *54*(6), 115:1-115:35. https://doi.org/10.1145/3457607

Olateju, O. O., Okon, S. U., Olaniyi, O. O., Samuel-Okon, A. D., & Asonze, C. U. (2024). Exploring the Concept of Explainable AI and Developing Information Governance Standards for Enhancing Trust and Transparency in Handling Customer Data. *Journal of Engineering Research and Reports*, *26*(7), 244–268. https://doi.org/10.9734/jerr/2024/v26i71206

Pan, B., Stakhanova, N., & Ray, S. (2023). Data Provenance in Security and Privacy. *ACM Computing Surveys*, *55*(14s), 1–35. https://doi.org/10.1145/3593294

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410. https://doi.org/10.1109/DSAA.2016.49

Data-Driven Supply Chain Orientation: The Role of Big Data Analytics, Challenges and Opportunities

Hakeem Sunmola (University of Birmingham), Ali Esfahbodi (University of Birmingham) and Yufeng Zhang (University of Birmingham)

Completed Research

Abstract

This paper conducts a systematic literature review of the transformative impact of big data analytics on supply chain orientation. Following a structured methodology based on the PRISMA framework, 60 articles were analysed to identify themes and synthesis results. The review shows that data analytics enables organisations to achieve more data-driven and more strategically aligned supply chain orientation through improved decision-making, supply chain orientation visibility, risk management and efficiency. However, challenges such as data heterogeneity, technological infrastructure, governance and human resources need to be addressed to fully exploit big data in supply chain orientation. This analysis reveals how practitioners can work with big data analytics to enhance their supply chain activities and gain a competitive edge in a dynamic environment today.

Keywords: Big Data analytics, Supply Chain Orientation, Data-driven Decision Making, Agility, Collaboration, Integration, Security.

1. Introduction

Data generation in the digital revolution era has created a domain for organisations to utilise to enhance their business operations. This phenomenon, which is commonly termed Big Data (BD), has driven organisations to analyse their supply chains and represents a tremendous opportunity for organisations to gain further insights, optimisation of operations and competitive advantage (Kamble and Gunasekaran, 2020). One such area where BD may transform practices is supply chain management (SCM), leveraging big data anlayrics (BDA) to enable businesses to transform their supply chain orientation (SCO) to be more agile, efficient and resilient in today's dynamic global market. SCO is the alignment of an organisation's supply chain activities with its overall business objectives and refers to a holistic

management of the flow of goods and information from raw material sourcing to final product delivery (Esper et al., 2010). An effective Data-Driven SCO (DDSCO) builds collaboration

and trust among supply chain partners, optimises processes and increases responsiveness to customer demands (Liu et al., 2023). In today's competitive environment a strong DDSCO is a

prerequisite for operational excellence, cost reduction and customer satisfaction, as the convergence of BD to help organisations strengthen their DDSCO offers a unique opportunity to unlock new levels of efficiency and responsiveness (Mentzer et al., 2001; Liu et al., 2023). BDA, which can analyse enormous amounts of structured and unstructured data, provides the tools to understand the supply chain (Khan et al., 2024). Insights gained from this data can be used to optimise key processes like demand planning, inventory management and logistics (Karaboga et al., 2023; Younis and Wuni, 2023) Moreover, BDA enables proactive risk management, supply chain orientation visibility (SCOV), and increased collaboration between stakeholders (Ghaleb et al., 2021; Bechtsis et al., 2022). While the potential of BD in SCM is well recognised, its specific roles, opportunities and challenges in SCO need to be further investigated. Much of the existing research addresses BD applications in isolated supply chain functions or broader business contexts (Addo-Tenkorang and Helo, 2016; Nguyen et al., 2018). This paper addresses this gap by focusing on how BDA transforms SCO and what this transformation means for organisations looking to improve their data-driven decision-making (DDDM) capabilities. Thus, the underlying research goal is to understand how BDA transforms SCO, and what are the key challenges and opportunities for businesses looking to improve their DDDM capabilities? The research questions (RQ) to explore this research objective are threefold:

> RQ1. How does BDA enable more proactive and predictive decision-making in SCM compared to traditional approaches?
>
> RQ2. How can BDA optimise key processes within a SCO, such as demand planning, inventory management, and logistics?
>
> RQ3. What are the unique data integration and interoperability challenges faced by organisations seeking to leverage BDA across diverse systems and partners within a SCO?

While previous review studies BDA in SCM have provided valuable insights into its applications (e.g. Hazen et al., 2014; Schoenherr and Speier-Pero, 2015; Zhong et al., 2016; Nguyen et al., 2018), however, to the authors knowledge, a review considering the linkage to BDA with DDSCO has yet to be explored. Additionally, this study aims to fill this gap by systematically analysing how BDA supports SCOV and facilitates real-time decision-making, demand forecasting, risk management, and sustainability. Given the rapid acceleration of digital transformation post-COVID-19, this research also explores emerging challenges and strategic opportunities that organizations must consider to optimize supply chain performance. By categorizing recent literature into distinct themes, this study provides a structured, up-to-date analysis that extends the theoretical and practical understanding of BDA's role in modern supply chains, highlighting the growing body of knowledge by examining the role and opportunities of BDA in transforming DDSCO, while also presenting the challenges associated with its adoption (Akter et al., 2016; Aydiner et al., 2019). The reviewed literature clearly demonstrates BDA's transformative potential to revolutionise SCM through the optimisation of DDSCO, enhancing efficiency, improving sustainability, and gain a competitive advantage.

This paper is structured to provide a comprehensive analysis of the topic. Following this introduction, a literature review will examine existing research on BDA and DDSCO. The subsequent sections will delve into the methodological approach of this paper, findings and discussions regarding the specific roles, opportunities, and challenges of BDA in the context of DDSCO. The paper will conclude with the summary of findings, implications for practice, limitations, and future research directions. By exploring these key aspects, this research aims to contribute to a deeper understanding of how BDA can be leveraged to achieve a more DDSCO.

2. Literature Review

2.1. Big Data

The importance of BD in the past decade can not be underestimated, its introduction has revolutionised the way businesses leverage data. While the terminology vary the core idea is using advanced analytics to extract valuable insights from large and complex datasets (De Mauro et al., 2015). De Mauro et al. (2015, p.9) defines BD as "information assets characterised by high volume, velocity, and variety, requiring specific technology and analytical methods for value extraction." BDA has emerged from advancements in business analytics and business intelligence, offering a more sophisticated approach to analysing and managing complex datasets (Zameer et al., 2020b). This evolution has led to diverse analytical methodologies and tools designed for the challenges posed by BD.

These advancements have paved the way for DDDM in various sectors, including supply chain management (SCM). Wang et al. (2022) reviewed the widespread use of BDA in SCM and logistics, highlighting its potential for competitive advantage. They highlight that by leveraging BD, logistics providers can improve their capabilities, responding to demand and capacity variations in real-time with dynamic pricing and product offerings. This is supported in the literature by Le (2023) who argues that BDA is fundamental to future sustainable SCM, while (Tripathi et al., 2024) emphasise its potential for improving accuracy in supply chain operations. Research recognises the significance of studying the integration of BDA in the context of SCM, risk management, and agility (Papadopoulos et al., 2017) and human resource management for organisational sustainability (El-Kassar and Singh, 2019), and poignantly the role it plays in an organisations low-carbon sustainability efforts (Singh et al., 2018; Liu, 2019).

2.2 Data-Driven Supply Chain Orientation

The concept of SCO has evolved alongside the rise of BD, leading to the emergence of DDSCO (Liu et al., 2023). Early research on SCO focused on the role of information technology in enhancing economic and environmental performance (Jadhav et al., 2019). Digital supply chain systems can help to facilitate information flows that enable firms to match supply and demand, ultimately enhancing performance (Wu et al., 2006). In a similar vein, literature highlighted the contribution of information technology-enabled supply chain platforms to reducing production costs and improving economic performance (Prajogo and Olhager, 2012; Qrunfleh and Tarafdar, 2014). Whilst, the emphasis of the role of information sharing within the supply chain for achieving sustainability has also been highlighted (Lai et al., 2015; Jadhav et al., 2019). This has therefore led to the concept of DDSCO to emerge, recognising the transformative potential of BDA in SCM.

DDSCO, refers to the presence of an infrastructure to process and analyse BD and the strategic intent to leverage the resulting insights to improve supply chain performance (Chavez et al., 2017; Yu et al., 2018). This involves not only the technological capabilities to handle BD, but also the organisational commitment to utilise data-driven insights for strategic decision-making and performance improvement (Liu et al., 2023). The importance of BD in SCM is underscored by its potential to enable new ways of organising and analysing supply chain processes to facilitate improved managerial decision-making (Han et al., 2024), innovation (Qrunfleh and Tarafdar, 2014) and market sensing (Wamba et al., 2020) leading to improved performance (Kamboj and Rana, 2023). However, despite its potential, BDA is still evolving (Kristoffersen et al., 2020).

3. Research Methodology

3.1 Data extraction, screening and evaluation

This study employs a Systematic Literature Review (SLR) approach and tests the research questions posed, acknowledging the need for Systematic and thorough retrieval of relevant studies Durach et al. (2017). Following the 'Preferred Reporting Items for Systematic Reviews & Meta-Analyses' (PRISMA) framework this research follows a structured approach to ensure clarity, and replicability (Moher et al., 2009). The study selection follows the four phases of PRISMA: Identification, Screening, Inclusion and Eligibility.

Furthermore, predefined inclusion and exclusion criteria (Table 1 below) were applied to ensure relevance and quality of studies selected. For instance, the period 2020-2024 was selected to capture BDA developments in SCM as the technological change in this area is rapid especially AI, IoT and real time analytics are emerging while considering the recent literature to reflect the current trends, opportunities and challenges. Moreover, the COVID-19 pandemic that started in 2020 further digitalized supply chains, making this period relevant for understanding BDA in modern SCM practices.

Table 1. Inclusion / Exclusion Criteria for SLR

Criterion	Inclusion	Exclusion
Topic	• Explicitly addresses data-driven approaches in SCO / SCM. • Focuses on the application of BD in SCM.	• Discusses traditional SCM or data analysis without explicitly linking them to the specified digital technologies. • Focuses on general digital transformation or data analytics without specific applications in SCM.
Types of studies	• Empirical studies (quantitative or qualitative) • Case studies • Review articles/meta-analyses	• Opinion pieces, editorials, blog posts • Purely theoretical papers without strong empirical evidence.

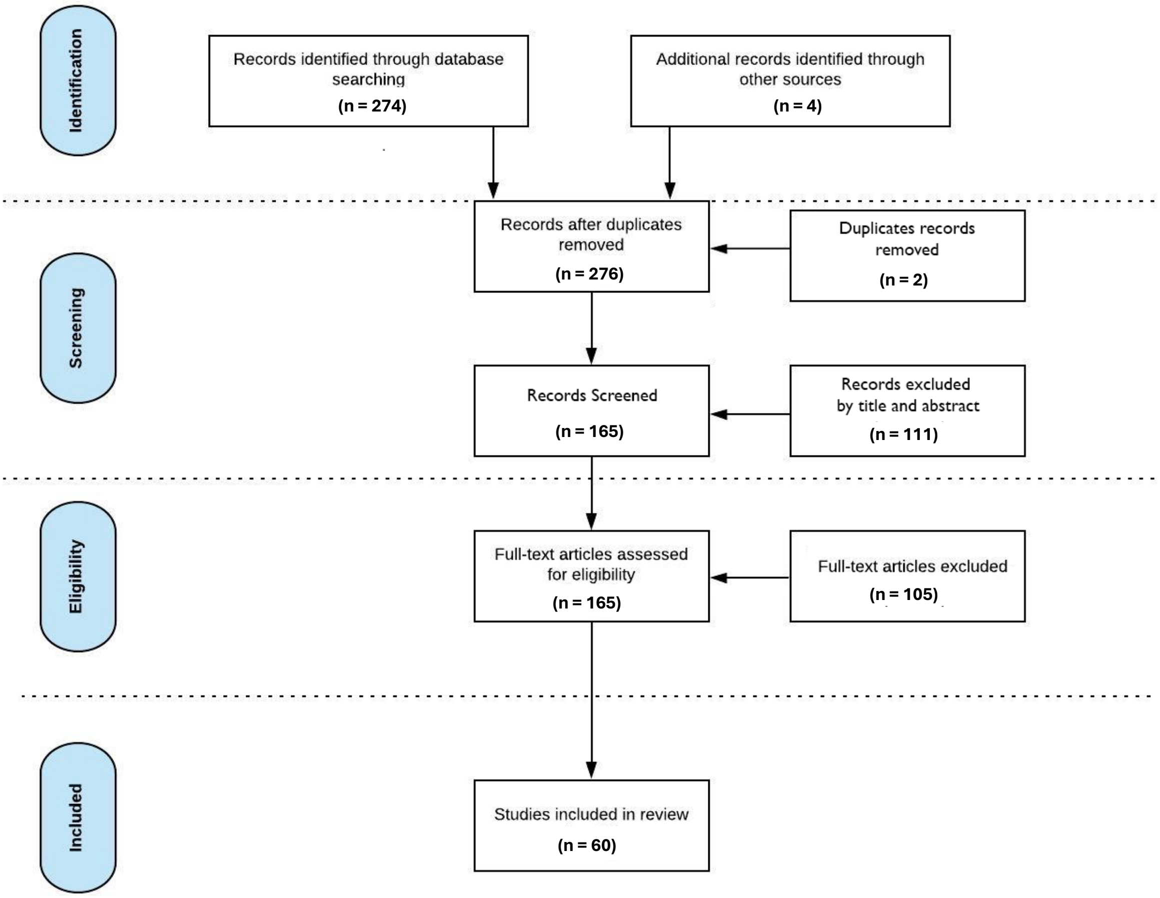
Publication characteristics	• Peer-reviewed journal articles or conference proceedings • Published in English • Published between 2020-2024	• Non-peer-reviewed publications (white papers, reports) • Articles in predatory or low-quality journals • Duplicate publications

The identification phase commenced with a comprehensive search across six prominent scientific databases: Wiley, Scopus, Taylor & Francis, Emerald, ScienceDirect, and Web of Science. These databases were strategically selected due to their extensive indexing capabilities across various fields pertinent to this review, including operations and supply chain management. To maximise the retrieval of relevant articles, a targeted search string was formulated, incorporating the primary key terms, “Supply Chain Management”, "Supply Chain Orientation", and “Big Data” along with related words and phrases for "Big Data." The following search strings were adopted: (a) ("Supply Chain Management" OR “Supply Chain Orientation”) AND ("Big Data" OR "Digital Technology" or “Digital Technologies”). In databases that allowed for the wildcard functionality, such as ScienceDirect, “Digital Technolog*” was utilised, as shown in Table 2 below. This comprehensive search strategy, limited to English language publications, yielded a total of 274 records. Hand searching and citation chaining were employed as a means to supplement the database search (Craane et al., 2012), resulting in the identification of 4 additional records.

Table 2. Search String used for SLR

Search String Component	Terms Included
Supply Chain Focus	"Supply Chain Management" OR "Supply Chain Orientation"
Big Data Focus	"Big Data" OR "Digital Technology" OR "Digital Technologies" OR "Digital Technolog*"

Screening phase included identifying duplicate records and removing them from the initial pool of 278 identified records. 2 duplicate articles were excluded for 276 unique records. Then, the Eligibility phase performed a title and abstract analysis of the remaining 276 studies to determine relevance to the research questions. That excluded another 111 records deemed out of scope of review and left 165 articles to assess. The final phase, Inclusion, required a full-text examination of these 165 articles to confirm contextual and empirical relevance to the research questions. Based on the inclusion and exclusion criteria in Table 2 this in-depth assessment excluded 105 articles. In the end, 60 articles were sufficient to provide a balanced account of the role of BD in DDSCO (See Figure 1 below), as per the PRISMA flow diagram.



Identification
Records identified through database searching
(n = 274)
Additional records identified through other sources
(n = 4)
Screening
Records after duplicates removed
(n = 276)
Duplicates records removed
(n = 2)
Records Screened
(n = 165)
Records excluded by title and abstract
(n = 111)
Eligibility
Full-text articles assessed for eligibility
(n = 165)
Full-text articles excluded
(n = 105)
Included
Studies included in review
(n = 60)

Figure 1. Systematic Literature Review PRISMA (adapted from Moher et al. (2009)

3.2 Thematic Analysis

After systematic retrieval, selection and evaluaiton of relevant literature, a thematic analysis approach was applied to organise and synthesise results. Thematic analysis is a qualitative research method for identifying, analysing and reporting patterns (themes) within data (Naeem et al., 2023). It provides a flexible yet rigorous method of analysing textual data that identifies key concepts, patterns, and relationships (Naeem et al., 2023). The organisation, coding and analysis of the selected articles were carried out using NVivo software version 12. NVivo facilitates a systematic and rigorous process for identifying key concepts, patterns, and relationships within the literature. Following the approach of Sabharwal and Miah (2021), the analysis started with familiarisation with the data - which was achieved by repeatedly reading the selected articles to get a sense of the content and initial ideas and patterns. Then key phrases, sentences or paragraphs related to the research questions were coded to identify common themes and ideas. These codes were then grouped and classified into possible themes based on common patterns and meanings. The identified themes were reviewed and refined to ensure they represented the data and were relevant to the research questions. Following this, themes were defined and given brief informative names that best reflected their essence. Finally, the final report was structured around the identified themes and provided a coherent and insightful analysis of literature. This is discussed in the following sections.

3.2.1 Thematic Analysis Procedure

This study follows Braun and Clarke's (2006) thematic analysis framework, adapted for a systematic literature review to ensure rigor and replicability. The six-stage process, data familiarization, coding, theme identification, review, definition, and synthesis, was structured using NVivo to enhance transparency and pattern recognition across the literature. A hybrid deductive-inductive approach guided the coding, balancing predefined themes from supply chain management and digital transformation with emergent insights from the data. This integration enabled both structured categorization and flexibility in capturing novel trends in BDA and digital technologies within SCO.

3.2.2 Coding Process

The coding process was systematically conducted using NVivo to ensure transparency and consistency. Full-text articles were imported, and relevant sections on BDA, digital technologies, and SCO were segmented for structured analysis. Initial codes were assigned to key concepts and organized under parent nodes representing overarching themes, with child nodes capturing specific dimensions. Themes were refined iteratively, consolidating overlapping categories and identifying patterns across studies. To enhance reliability, a subset of articles was double-coded, with discrepancies resolved through discussion. NVivo’s structured framework enabled systematic traceability, linking themes directly to their references for methodological rigor.

3.2.3 Themes and Sub-Themes

NVivo facilitated a systematic and transparent thematic analysis by structuring themes and sub-themes with corresponding references from the reviewed literature. This ensured consistency, traceability, and a clear audit trail throughout the coding process. The hierarchical organization of themes captured key patterns and relationships, allowing for a structured synthesis of findings. Table 3 provides an example of this coding framework. This method enhanced analytical rigor, ensuring insights were systematically categorized and reliably interpreted.

Table 3. Extract of NVivo systematic structure

Author	Theme (Parent Node)	Sub-Theme (Child Node)	Key Insights
Yu et al. (2020)	Opportunities	Collaboration and Integration	Hospitals can leverage BDA to enhance supplier coordination and improve healthcare service efficiency.
Maheshwari et al. (2020)	Challenges	Data Governance and Security	BDA adoption is hindered by data privacy concerns and regulatory compliance issues.
Margaritis et al. (2022)	Opportunities	Supply Chain Visibility & Transparency	BDA enhances visibility by integrating data from different FSC stages, leading to better coordination.

4. Findings

4.1 Descriptive Statistics

4.1.1 Year of Publication

The distribution of publications over time provides valuable insights into the evolution and maturity of research on BDA in DDSCO. Figure 2 presents the number of publications included in this review by their year of publication.

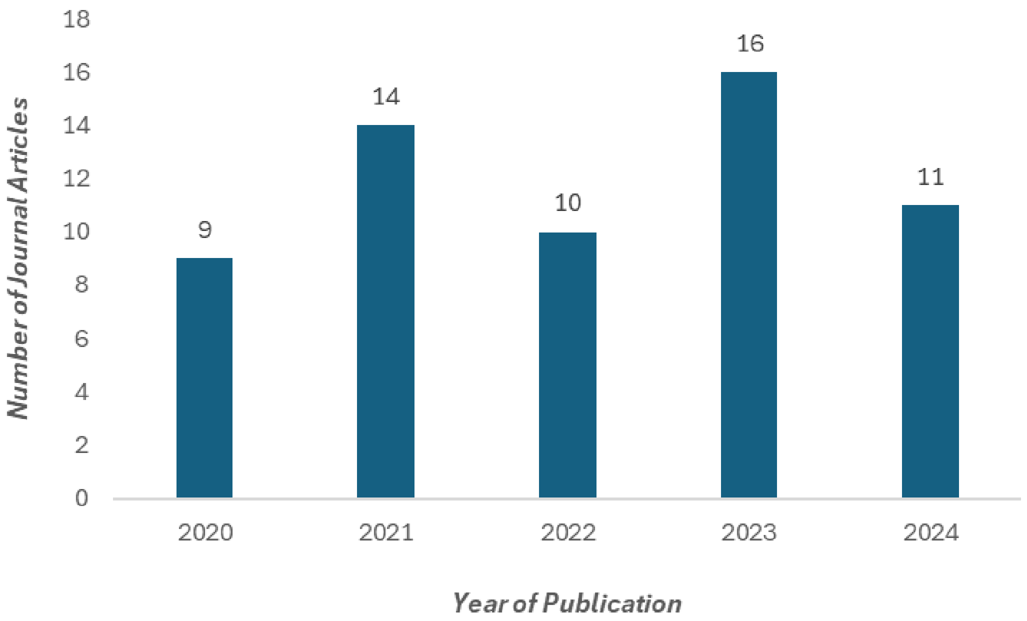


Figure 2. Year of Publication of Articles from SLR

The number of publications increased gradually from 2020 to 2023 (with a small decline in 2024). This trend indicates a recent interest in the application of BD in the context of SCO. The peak in 2023 reflects increased research activity, perhaps reflecting an awareness of BD's potential to transform SCM practices. The slight dip in 2024 could be the publication lag for recent studies since some relevant articles may still be under review. This trend indicates that this field remains in its early growth phase (Singh et al., 2022), and there is still much scope for further development of knowledge and application of BD in SCO for better supply chain operations and strategic goals.

4.1.2 Publication Distribution

Analyzing the publications indicated that the 60 articles included in this review were published in 45 journals. These (39) publications typically contained one article relevant to this review and thus represented a broad spectrum of research in this area. These publications include journals as Omega, Journal of Cleaner Production, Total Quality Management Journal, Supply Chain Management: An International Journal and Benchmarking: An International Journal. This diversity demonstrates the interdisciplinary nature of research in BD and SCO originating from operations management, sustainability, information technology and engineering (Tseng et al., 2021, 2021; Yu et al., 2021). Nevertheless, a fairly small number of publications (six) comprised a bigger part of the literature and each contained 2 or more articles. These dominant publication outlets are shown in Table 4 below.

Table 4. Publications with 2 or more Journal Articles

Publication	Frequency
International journal of production research	6
Sustainability	4
Technological forecasting & social change	3
The international journal of logistics management	2
Annals of operations research	2
Computers & industrial engineering	2

4.1.3 Analysis of Keywords

Keyword analysis revealed the main themes and concepts from the reviewed literature. Keywords provided by authors of the 60 articles were analysed to determine frequency and visualise relationships. Figure 3 displays a word cloud based on these keywords illustrating key themes and concepts.

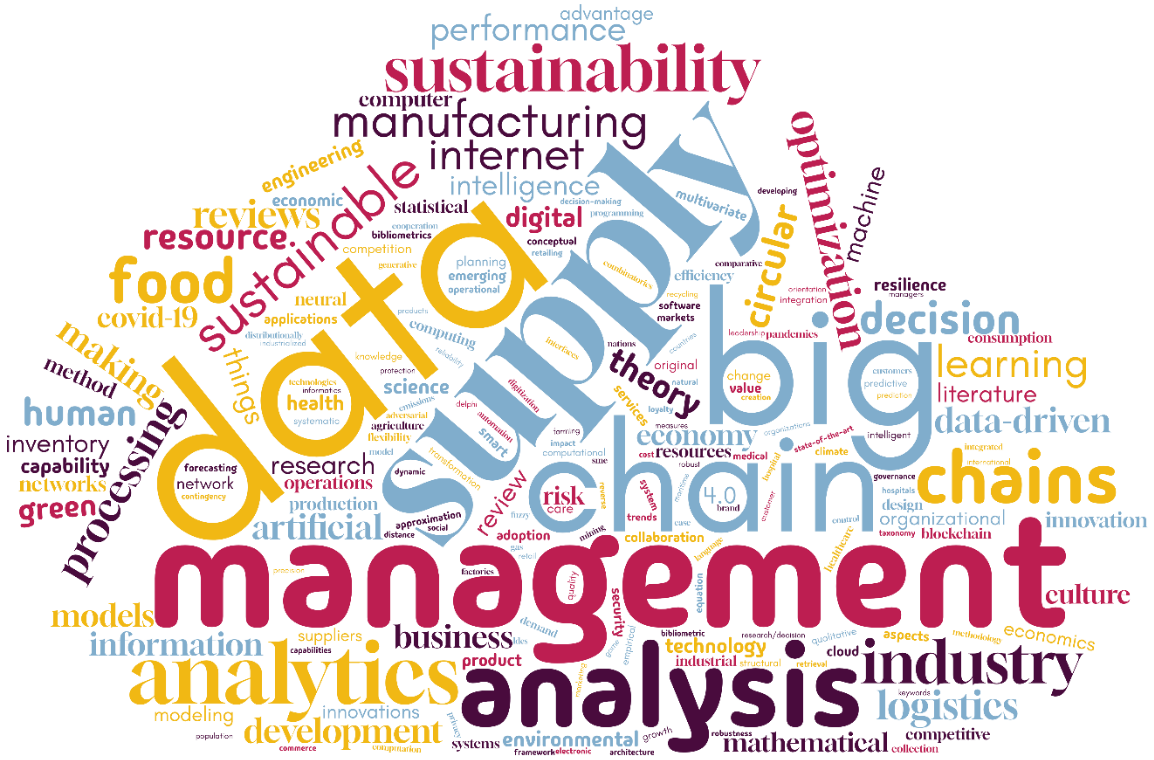


Figure 3. Word Cloud – Keywords from the SLR articles

As depicted in the word cloud, the most frequent keywords include "data", "supply chain", "management". "big", and "analysis". These keywords highlight the core focus of the reviewed literature on the application of BDA in SCM. Additional keywords: "sustainability”, "performance”, "industry”, "logistics”, "manufacturing”, "artificial”, "data-driven”, "decision”, "development”, "learning”, "business”, "information”, "intelligence”, "optimization”, "processing”, "circular”, "economy”, "environmental”, "human”, “competitive” and "advantage”, further emphasize the complex nature of this particular research discipline. The word cloud displays the main themes and concepts from literature and illustrates how BDA relates to various aspects of SCM and to DDSCO.

4.1.4 Themes from the SLR

The three main themes identified in the SLR are: 1) the role of BDA in Supply Chain SCO, 2) the opportunities presented by BDA in SCO, and 3) the challenges associated with implementing BDA in SCO. These themes, with their respective sub-themes, are visually represented in Figure 4.

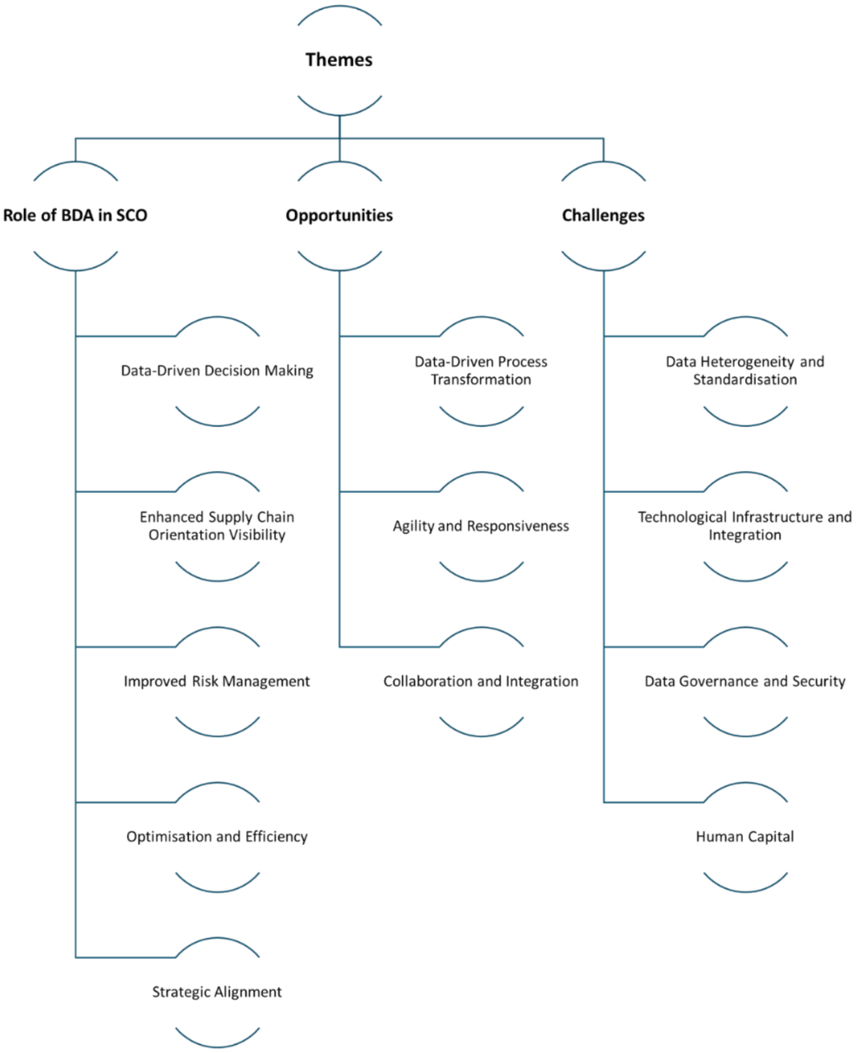

Themes
Role of BDA in SCO
Opportunities
Challenges
Data-Driven Decision Making
Enhanced Supply Chain Orientation Visibility
Improved Risk Management
Optimisation and Efficiency
Strategic Alignment
Data-Driven Process Transformation
Agility and Responsiveness
Collaboration and Integration
Data Heterogeneity and Standardisation
Technological Infrastructure and Integration
Data Governance and Security
Human Capital

Figure 4. Themes of the SLR

5. Discussion

5.1 The Role of BDA in SCO

5.1.2 Data-Driven Decision Making

BDA is fundamentally transforming the way organisations approach SCO, enabling a move from reactive to proactive and predictive decision-making. This transformation is driven by BD informing real-time about dynamic conditions, demand fluctuations and possible disruptions. As Maheshwari et al. (2021) highlight, the analysis of real time data streams from sensors, social media, and technological systems such as point-of-sale terminals enabling fast responses to changing market conditions and customer demands. This real-time capability enables organisations to change production schedules, rerouting shipments or pricing strategies in response to real-time market feedback (Kokkinou et al., 2023). Moreover, utilising real-time data analysis is crucial for understanding and reducing the "ripple effect" of disruptions in the supply chain so that organisations can adapt to unplanned events (Dolgui and Ivanov, 2021). A case study conducted by Sundarakani et al. (2021) signifies this whereby retailer can sense a sudden spike in demand for a specific product because of trending social media posts and use real-time data analysis to adjust stock levels and avoid stockouts. The ability for supply chains to become much agile as a results, improves customer satisfaction and also enhances financial performance outcomes (Kokkinou et al., 2023).

The application of predictive analytics further enhances the impact of BD in SCO. Analysing historical data and current information enables organisations to forecast demand, predict risks and manage inventory levels (Ma et al., 2020). This proactive is essential for planning and reducing uncertainties (Younis and Wuni, 2023). This allows organisations such as those in the manufacturing industry to forecast demand for a new product launch using predictive analytics based on historical sales data and market trends (Wang et al., 2022). This allows optimal production planning, avoiding overstocking or understocking and a successful product launch. In their study on the application of Industry 4.0 enablers in SCM, Younis and Wuni (2023) highlight the increasing significance of predictive analytics in supply chain operations. Emergence of new performance measures based on predictive analytics in BD-driven supply chains (Kamble and Gunasekaran, 2020) further illustrates this trend towards proactive performance management and better decision making. BDA also supports the monitoring and measurement of key performance indicators (KPIs) for SCM. This provides feedback on the performance of strategies and areas for improvement. Kamble and Gunasekaran (2020) describe a framework for a BD-driven supply chain performance measurement system and highlight the importance of data-driven performance evaluation for understanding and optimising supply chain effectiveness. This continuous monitoring enables businesses to track important metrics as order fulfilment rates, delivery times, stock turnover and customer satisfaction in real time. Identifying trends and anomalies in these KPIs can help businesses to address potential issues, optimise processes and improve overall supply chain performance (Alsadi et al., 2021). Furthermore, Chenger and Pettigrew (2023) state that for SCM to be effective, an organisational culture of continuous improvement must be adopted to leverage BD for DDSCO to optimise resilience in the supply chain.

5.1.3 Enhanced Supply Chain Orientation Visibility

In addition to operational decision-making, BDA also improves SCOV. This is distinct from general supply chain visibility in that it focuses on how data-driven insights can be used to achieve the SCOs strategic goals and objectives. By providing a unified, data-driven view of operations, BD promotes transparency and facilitates better coordination, better risk management, and better decision making across the supply chain while adhering to the overall SCO strategy (Chatterjee et al., 2022). Khoei et al. (2023) highlight the need for BDA to be embedded in supply chain processes to improve efficiency, quality and flexibility - core elements of SCO - where multiple stakeholders and geographically dispersed operations make it difficult to keep an overview. BDA aggregates data from multiple sources to give a single view of the supply chain and allows cross - partner collaboration and coordination within the context of the SCO strategic direction. Margaritis et al. (2022) point out in their study on BD applications in food SCM, the importance of data-driven approaches to improve transparency, traceability and visibility in the food supply chain for a more informed and strategically aligned SCO. In addition, a data-sharing model put-forward by Bechtsis et al. (2022) highlights the facilitation of decision-making, collaboration and SCOV in supply chains, where data-driven approaches foster more connected and resilient supply chain operations that result in a stronger and more strategically aligned SCO.

Real-time data and technologies like RFID and IoT sensors enable SCOV to gain a holistic picture of the supply chain ecosystem. This increased visibility enables organisations to follow the activities, performance and likely risks of their supply chain partners and drive more strategic, aligned decision making. For instance, a manufacturer can see in real time the production capacity and inventory levels of the suppliers and adjust his production schedule to avoid production disruptions. This increases efficiency and builds collaboration and trust between partners. In the exact same manner, retailers can share real time data on consumer need and preferences with their suppliers and adjust production and distribution plans as necessary. This collaborative approach creates a more responsive and agile supply chain where all partners contribute to meeting customer demands effectively. This aspect of BD in SCM is very important, especially to improve SCOV for DDSCO for crisis response. With real-time visibility across the supply chain ecosystem, BD enables organisations to make informed decisions in line with their overall SCO strategy - even during disruptions. For example, during the COVID-19 pandemic, organisations with high SCOV identified vulnerabilities quickly, assessed the impact across supply chain partners and proactively adjusted their strategies to maintain business continuity and customer satisfaction. This proactive approach, supported by real-time monitoring and prediction of disruptions, can improve resilience and service levels as demonstrated by (Behera and Ramanathan, 2022). Through BD, Real-time data and technologies for SCOV can assist DDSCO regarding supply chain sustainability. Organisations can have visibility into environmental practices and performance of suppliers, assisting partner selection and collaboration that aligns with their own sustainability goals (Zekhnini et al., 2023). This data driven approach optimises resource allocation and waste minimisation, promoting a more sustainable and responsible supply chain ecosystem (Tseng et al., 2021).

5.1.4 Improved Risk Management

BDA aims at enhancing risk management capabilities which are a key element of a robust SCO. Using data-driven insights, organisations can go from reactive to proactive risk mitigation (Kokkinou et al., 2023). This is done through predictive risk assessment, where BDA discovers potential vulnerabilities based on historical data, market trends and external factors; such as weather or geopolitical events. For instance, the oil and gas industry might do predictive risk analysis to determine disruptions in their supply chain caused by natural disasters, political instability or regulatory changes in various areas (Chenger and Pettigrew, 2023). This allows them to develop contingency plans, diversify their supplier base or create buffer stocks to absorb these risks and maintain business continuity (Ma et al., 2020). Bechtsis et al. (2022) further highlight the need for data-driven approaches to enhance risk management capabilities in a strategically focused SCO by proposing a data-sharing and monetisation framework to address supply chain resilience, security and sustainability. Moreover, BDA conducts scenario planning to assess how disruptions might affect an organisation's ability to achieve its SCO goals. This allows development of contingency plans and business continuity thereby making the supply chain more resilient and flexible (Kokkinou et al., 2023). This proactive approach builds supply chain resilience and operational efficiency in line with the aim of SCO to maintain a robust, responsive and adaptable SCM (Tseng et al., 2021). Similarly, Zhao and You (2020) highlight BDs application to handling fluctuating customer demand. They propose the use of artificial intelligence (AI) to complement BD in making informed supply chain decisions despite uncertainty, to achieve strategic supply chain goals.

5.1.5 Optimisation and Efficiency

BDA plays an important role to drive optimisation and efficiency in key processes in a well aligned SCO, this leads to better resource utilisation and cost reduction as well as to improved supply chain performance. Enabling a data-driven approach to SCO, BDA enables organisations to align their operational decisions with their strategic goals resulting in a more integrated supply chain (Thekkoote, 2022). For instance, BDA can produce more accurate demand forecasts enabling better production planning, inventory management and ultimately a more flexible and efficient supply chain that can respond to changing market conditions and customer requirements (Luo et al., 2023).

In addition BDA can also optimise logistical processes, warehouse operations, and delivery schedules to achieve higher efficiency and lower costs (Verma et al., 2023). This optimisation of supply chain processes is brought about through DDSCO to ensure alignment across the supply chain to create a responsive and agile supply chain, ultimately contributing to overall business objectives such as increased profitability, enhanced customer satisfaction, and improved competitive advantage. For example, a organisation who focuses on same-day delivery could use BDA to plan delivery routes based on real-time traffic patterns, delivery time windows, and vehicle capacity, reducing transportation cost, improving delivery efficiency and customer satisfaction (Babu et al., 2024). However, to fully and effectively utilise BD for optimisation & efficiency, Ji et al., (2022) and highlight the need to ensure there is the right partner alignment of SCO across the supply chain, by ensuring organisations match BDA capabilities in the supply chain to realise maximum optimisation benefits, efficiency and performance supporting the strategic objectives of the supply chain. This is supported by Gligor et al. (2022) who highlights the importance of SCO-Fit.

5.1.6 Strategic Alignment

BDA promotes a culture of informed decision-making based on evidence and insights, rather than relying on managerial intuition (Joemsittiprasert et al., 2019). By integrating data analysis into the fabric of the organisation, businesses can embed a culture where data is used to inform strategic choices, operational decisions, and performance evaluations. This data-driven culture is fundamental for leveraging BDA in the SCO and gaining a competitive advantage. Karaboga et al. (2023) highlight the mediating effect of data-driven culture on BDA management capability and firm performance, emphasizing the need for organisational culture to unleash the full potential of BD for a strategically aligned and effective SCO.

5.2 Opportunities

5.2.1 Data-Driven Process Transformation

One of the key opportunities offered by BDA is to optimise processes and achieve the strategic goals of SCO. By leveraging BDA, organisations can align their operational decisions with their SCO objectives, leading to a more cohesive and effective supply chain. Maheshwari et al. (2021) express that BDA can help to optimise business operations, reduce costs and facilitate decision-making in SCM. This optimisation can be achieved by process improvements, automation, waste reduction and cost reduction initiatives supported by DDSCO insights. For example, a manufacturer could use BDA to analyse production data and identify areas where production processes can be streamlined to reduce lead times and improve output, as Younis and Wuni (2023) highlighted in the potential of Industry 4.0 enablers to improve decision-making capabilities and process optimisation. Similarly, a retailer can set up an automated inventory replenishment system based on BDA to predict demand and execute orders when inventory levels fall below a certain threshold, as (Gawankar et al., 2020) put-forward on BDA’s potential for proactive transformation. Also, BDA can identify waste streams in the supply chain so organisations can reduce waste, reduce their environmental footprint and improve sustainability (Dwivedi et al., 2022). By optimising processes, automating tasks and reducing waste BDA can lead to significant cost reductions throughout the supply chain, improving profitability and competitive position (Ma et al., 2020).

5.2.2 Agility and Responsiveness

BDA enables organisations to develop agile, responsive supply chains that respond to dynamic market conditions, customer demands and disruptions. Such agility is fundamental for the SCO's objectives of flexibility, adaptability and responsiveness to dynamic market forces. Thekkoote (2022) stresses the agility and flexibility to respond to market changes and customer demands and identifies the role of BDA in achieving these capabilities. By analyzing real time data, businesses can react quickly to shifts in demand, supply and other unexpected events. This agility is crucial to maintain customer satisfaction and reduce disruption impact that could hamper SCO goals. For example, BD coupled with digital twin data-driven models improve resilience and service levels during pandemics (Behera and Ramanathan, 2022).

A successful digital twin SCM system requires internal and external data integration to maintain synchronisation across diverse SC elements, enhancing decision-making capabilities (Zaidi et al., 2024). Kokkinou et al. (2023) explored the relationship between BDA and digital decision culture (DDC), and the development of supply chain robustness and resilience. They addressed how proactive DDDM can support organisations to survive disruptions and gain competitive advantage. They highlight that DAC in combination with DDC improves supply chain robustness through the promotion of proactive measures such as safety stock and redundancies that are required to respond to supply chain disruptions. BDA can also be applied to identify and assess potential supply chain risks to enable organisations to take appropriate measures against them and improve the supply chains resilience (Huang et al., 2020).

5.2.3 Collaboration and Integration

BDA fosters greater collaboration and integration among stakeholders in the supply chain, which is essential for achieving a truly integrated and responsive SCO. This collaborative ecosystem, facilitated by BDA, enables organisations to break down data silos, enhance knowledge sharing, and improve coordination among supply chain members (Delgosha et al., 2020; Pan et al., 2022; Younis and Wuni, 2023).

Information sharing is a key enabler of collaboration, and BDA empowers organisations to share real-time data and insights with their partners. This leads to better coordination, reduced lead times, and improved overall supply chain performance, supporting the SCO's aim for a responsive and efficient network (Liu et al., 2022). Furthermore, BDA facilitates the interconnection of different systems and processes within the supply chain, creating a more connected and cohesive ecosystem that supports the objective of achieving SCOV and control (Alsadi et al., 2021). Yu et al.'s (2021) study evidences this in the healthcare context, demonstrating how BDA significantly impacts hospital supply chain integration and operational flexibility by fostering collaboration both within and between organisations. Similarly Tripathi et al. (2024) highlight the importance of collaboration in achieving an integrated responsive supply chain. Gligor et al. (2022) express the importance of SCO-Fit, which emphasises the importance of customer and supplier fit for achieving agile supply chains. The implementation of BDA supports this enabling the analysis of data on supplier performance, enabling organisations to select the right partners and adjust collaboration strategies (Ji et al., 2022).

5.3 Challenges

5.3.1 Data Heterogeneity and Standardisation

Heterogeneity of data across the supply chain is a major challenge for BDA exploitation for SCO. Diverse data sources from supply chain stakeholders (suppliers, manufacturers, distributors, customers) use disparate formats, structures and standards making data integration and analysis complex (Delgosha et al., 2020). This lack of standardisation results in interoperability problems and inconsistencies and errors in analysis. Moreover, data may remain trapped in isolated systems or "data silos" within departments or partner organisations, preventing a holistic view of the supply chain and limiting BDA leverage (Del Giudice et al., 2021; Margaritis et al., 2022). Data standardisation may hinder effective analysis and decision making (Wang et al., 2022; Luo et al., 2023)This need for standardisation is essential to enable information flow and collaboration between supply chain stakeholders to achieve SCO. Hajek and Abedin (2020) demonstrate this with their study of inventory backorder prediction, where they confront the issue of imbalanced datasets. Whilst Rui and Li (2024) note the difficulties in utilising BDA to unstructured and noisy data, particularly in the context of demand forecasting.

5.3.2 Technological Infrastructure and Integration

Another major challenge is the technological infrastructure required to support BDA in SCO. Diverse technologies (IoT devices, ERP systems, cloud platforms) must be integrated within the supply chain, demanding considerable effort and systems reconfiguration. Younis and Wuni (2023) point out the technological complexity of integrating Industry 4.0 technologies, including BDA, into existing SCM processes, highlighting that this can be particularly challenging for SMEs or organisations with less digital maturity. The issue of disconnected system is further compounded by the need to integrate other emerging technologies such as blockchain and digital twins to support BDA (Sundarakani et al., 2021; Zaidi et al., 2024). Ironically, BD requires big storage facilities, as data volumes increase, the infrastructure must scale to meet the growing requirements, which can be a major investment and requires careful planning to ensure the technological infrastructure can cope with organisational growth, which could be a significant barrier for SMEs to implement BDA (Behera and Ramanathan, 2022). Coupled with the resistance to change and investment requirements, organisations need to adopt a technology embracing culture to successfully leverage BDA for SCO. Several implementation barriers related to the adoption of BD in supply chains, including technological, organisational, and financial hurdles, as well as the complexity of integrating these technologies into existing supply chain structures (Verma et al., 2023; Pratap et al., 2024).

5.3.3 Data Governance and Security

Information sharing is a key concept for SCO, but sharing sensitive data across the supply chain raises data breach, unauthorised access and misuse concerns. Data security and privacy among systems and partners is an important concern for data integrity and compliance with regulations. Security and privacy has been highlighted as as key organisational challenge associated with BDA adoption in SCM, emphasising the need for robust security measures to protect sensitive information and maintain data integrity (Maheshwari et al. 2021). This is particularly important in SCO where data is shared across multiple partners and tiers to support strategic decisions that impact real or indirect performance of the organisation. Wang et al. (2022) review of BDA for intelligent manufacturing systems data security and governance challenges and introduce the possibility of blockchain to address these issues. This privacy and security concern is particularly relevant in the healthcare sector as BD risks such as data access problem, and effective utilisation (Kokshagina et al., 2024) and smart technologies (Chang et al., 2023) bring to light the need for secure and transparent protection of sensitive health information. Hajiheydari et al. (2021) also point out the need for effective security and clear communication to curb scepticism and resistance among practitioners when using technologies in the healthcare sector.

Therefore, clear data ownership and access control policies are required to prevent misuse and enable responsible data handling in a SCO context where data is shared by several partners. Additionally, as SCO places a premium on alignment across the supply chain, many organisations operate internationally and are subject to various data privacy regulations, such as GDPR (Thekkoote, 2022). However, other digital technologies such as blockchain can help in the successful implementation of BDA for greater security and transparency (Sundarakani et al., 2021).

5.3.4 Human Capital

Effectively leveraging BDA for SCO requires skilled workforce in data science, data engineering and analytics but organisations often face a talent shortage and lack of specialist talent (Delgosha et al., 2020). This skills gap can prevent strategic alignment of supply chain with overall business objectives as envisaged in SCO. Maheshwari et al. (2021), identify skills shortage as a major organisational challenge associated with BDA adoption and call on organisations to provide training and development to plug this gap and ensure a capable workforce to support a DDSCO strategy. Thekkoote (2022) echoes this concern, noting the lack of qualified personnel for data analytics as a major challenge in leveraging BD for supply chain management.

Likewise, data literacy development in the supply chain is required for data interpretation and decision making by SCO. Chenger and Pettigrew (2023) also highlight the skills gap and data-driven culture that organisations need to develop to exploit BD for supply chain optimisation and resilience - two key components of a successful SCO strategy. Dwivedi et al. (2022) describe knowledge & skill gaps in sustainable production within the Industry 4.0 framework and propose a set of strategies to train and support industries in applying circular economy principles with BD technologies. The challenges associated with the human capital gap in BD adoption, specifically highlighting the need for employee training and addressing the limitations of technological infrastructure in developing healthcare sectors (Ghaleb et al., 2021). The challenge in implementing BDA in SCO often necessitates changes in processes and organisational culture internally and across supply chain members. Managing this change and ensuring support from top management and other stakeholders is essential for successful adoption and implementation. The potential for cultural resistance to data-driven culture and BDA adoption is signified by the perception of the benefits of digital technologies , particularly in traditional SMEs, underscoring the need for effective change management strategies to overcome these barriers and foster a data-driven culture that supports the strategic alignment goals of SCO (Awan et al., 2023).

6 Conclusion

6.1 Summary of Findings

This SLR shows that BDA is fundamentally transforming SCO in terms of DDDM, SCO visibility, risk management, efficiency and strategic alignment. Powered by BD, real-time insights, predictive analytics and performance measurement enable organisations to make informed decisions, predict trends in the future and manage their supply chains proactively. Improved SCO visibility - through end-to-end transparency, real-time tracking and demand sensing - offers a holistic view of supply chain operations facilitating better coordination, risk management and decision making. BDA also provides risk management capabilities such as predictive risk assessment, scenario planning and early warning systems to organisations to prevent disruptions and improve supply chain resilience. Moreover, BD drives optimisation and efficiency for key processes within the SCO which improves resource utilisation, reduces costs and improves performance. Fostering a data-driven culture and providing insights for strategic planning, BDA enables organisations to align their SCO with long-term goals and market dynamics to create a competitive advantage.

6.2 Implications for Practice

The findings of this research provide key implications for organisations implementing or optimising SCO through BDA three-phased approach is recommended to ensure a structured progression from foundational investments to advanced analytics and full-scale BDA integration, enabling firms to maximise BDA’s potential while addressing implementation challenges.

The first phase focuses on building foundational capabilities, including technological infrastructure, data literacy programs, and a data-driven culture to ensure effective BDA integration. Organisations must align BDA investments with key SCO dimensions such as collaboration, trust, and shared values (Chenger and Pettigrew, 2023), Prioritisation should start with core infrastructure to enable more sophisticated analytics in later stages. The second phase leverages BDA for process optimisation, enhancing supply chain agility, demand forecasting, and risk management. Implementing real-time DDSCO and process automation can improve efficiency and responsiveness. However, data integration and standardisation remain critical to overcoming fragmentation. A robust digital backbone is necessary to facilitate seamless BDA adoption (Sundarakani et al., 2021). The final phase involves scaling BDA adoption and addressing key challenges. Effective data governance and security are crucial to protecting information integrity, while investment in skilled talent ensures proper utilisation of BDA capabilities. Firms must also manage cultural resistance to data-driven decision-making, fostering organisational buy-in through change management strategies (Awan et al., 2023). By following this three-phased approach, organisations can systematically align BDA investments with SCO objectives, ensuring a resilient, agile, and data-driven supply chain while overcoming adoption barriers.

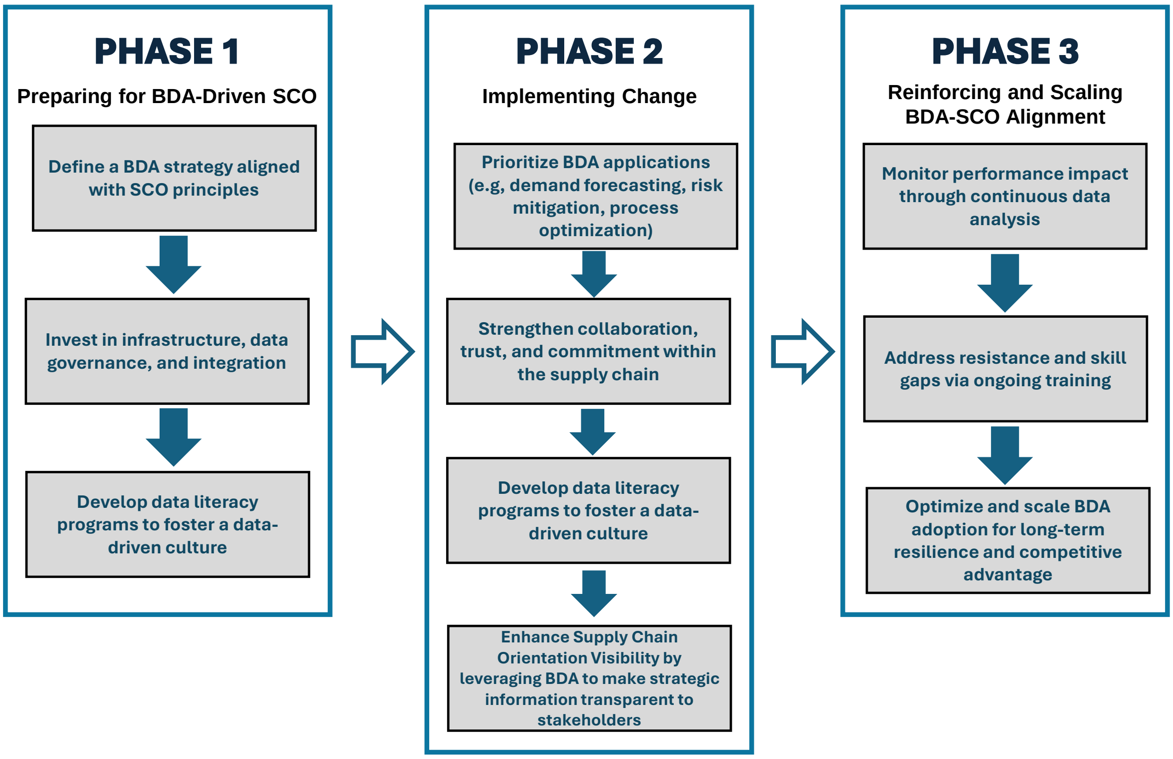

PHASE 1
Preparing for BDA-Driven SCO
Define a BDA strategy aligned with SCO principles
Invest in infrastructure, data governance, and integration
Develop data literacy programs to foster a data-driven culture
PHASE 2
Implementing Change
Prioritize BDA applications (e.g, demand forecasting, risk mitigation, process optimization)
Strengthen collaboration, trust, and commitment within the supply chain
Develop data literacy programs to foster a data-driven culture
Enhance Supply Chain Orientation Visibility by leveraging BDA to make strategic information transparent to stakeholders
PHASE 3
Reinforcing and Scaling BDA-SCO Alignment
Monitor performance impact through continuous data analysis
Address resistance and skill gaps via ongoing training
Optimize and scale BDA adoption for long-term resilience and competitive advantage

Figure 5. Three-Phase Approach to Data-Driven Supply Chain Orientation (DDSCO)

6.3 Limitations and Future Research Directions

This research has limitations. For one, the literature review was limited to selected databases and publication outlets and may not have adequately represented the range of research on BDA and SCO. Furthermore, the analysis was primarily academic and potentially missed out some important insights from industry reports and practitioner publications. The study was limited to articles published in 2020 - 2024 to first understand the most recent developments on BD especially post COVID-19, but this may have excluded some seminal papers published earlier. Highly cited older papers identified by citation chaining were included to alleviate this limitation. Specific industry applications of BDA in SCO were also not explored specifically or extensively. Therefore, several directions for future research arise from this study. Firstly, further research on specific applications of BDA in specific industry sectors could provide more tailored guidance to practitioners. Second, ethical implications of BD use in SCO such as data privacy and algorithmic bias need to be explored for responsible and sustainable practices. Finally, the role of emerging technologies such as Artificial Intelligence and Machine Learning together with BDA could be explored to further optimise SCO.

6.4 Concluding Remarks

This research demonstrated that BDA has the potential to shape the future of SCO. A data-driven approach can provide considerable opportunities to enhance supply chain operations, efficiency, collaboration and agility to obtain a competitive edge in a dynamic world environment. While challenges remain for leveraging BD effectively, proactive measures such as data management, technological infrastructure, governance, and human resources could enable successful implementation and exploitation of this transformative technology.

References

Addo-Tenkorang, R. and Helo, P.T. (2016), “Big data applications in operations/supply-chain management: A literature review.” *Computers & Industrial Engineering*, 101: 528–543. doi:10.1016/j.cie.2016.09.023.

Akter, S., Wamba, S.F., Gunasekaran, A., Dubey, R. and Childe, S.J. (2016), “How to improve firm performance using big data analytics capability and business strategy alignment?” *International Journal of Production Economics*, 182: 113–131. doi:10.1016/j.ijpe.2016.08.018.

Alsadi, A.K., Alaskar, T.H. and Mezghani, K. (2021), “Adoption of Big Data Analytics in Supply Chain Management: Combining Organizational Factors With Supply Chain Connectivity.” *International journal of information systems and supply chain management*, 14 (2): 88–107. doi:10.4018/IJISSCM.2021040105.

Awan, U., Braathen, P. and Hannola, L. (2023), “When and how the implementation of green human resource management and data-driven culture to improve the firm sustainable environmental development?” *Sustainable development (Bradford, West Yorkshire, England)*, 31 (4): 2726–2740. doi:10.1002/sd.2543.

Aydiner, A.S., Tatoglu, E., Bayraktar, E., Zaim, S. and Delen, D. (2019), “Business analytics and firm performance: The mediating role of business process performance.” *Journal of Business Research*, 96 (C): 228–237.

Babu, M.M., Rahman, M., Alam, A. and Dey, B.L. (2024), “Exploring big data-driven innovation in the manufacturing sector: evidence from UK firms.” *Annals of operations research*, 333 (2–3): 689–716. doi:10.1007/s10479-021-04077-1.

Bechtsis, D., Tsolakis, N., Iakovou, E. and Vlachos, D. (2022), “Data-driven secure, resilient and sustainable supply chains: gaps, opportunities, and a new generalised data sharing and data monetisation framework.” *International journal of production research*, 60 (14): 4397–4417. doi:10.1080/00207543.2021.1957506.

Behera, A.K. and Ramanathan, K. (2022), “An effective disaster recovery model in supply chain management at times of pandemic.” *Cardiometry*, (25): 502–510. doi:10.18137/cardiometry.2022.25.502510.

Braun, V. and Clarke, V. (2006), “Using thematic analysis in psychology.” *Qualitative Research in Psychology*, 3 (2): 77–101. doi:10.1191/1478088706qp063oa.

Chang, V., Doan, L.M.T., Ariel Xu, Q., Hall, K., Anna Wang, Y. and Mustafa Kamal, M. (2023), “Digitalization in omnichannel healthcare supply chain businesses: The role of smart wearable devices.” *Journal of business research*, 156: 113369-. doi:10.1016/j.jbusres.2022.113369.

Chatterjee, S., Chaudhuri, R., Shah, M. and Maheshwari, P. (2022), “Big data driven innovation for sustaining SME supply chain operation in post COVID-19 scenario: Moderating role of SME technology leadership.” *Computers & industrial engineering*, 168: 108058–108058. doi:10.1016/j.cie.2022.108058.

Chavez, R., Yu, W., Jacobs, M.A. and Feng, M. (2017), "Data-driven supply chains, manufacturing capability and customer satisfaction." *Production Planning & Control*, 28 (11–12): 906–918. doi:10.1080/09537287.2017.1336788.

Chenger, D. and Pettigrew, R.N. (2023), "Leveraging data-driven decisions: a framework for building intracompany capability for supply chain optimization and resilience." *Supply chain management*, 28 (6): 1026–1039. doi:10.1108/SCM-12-2022-0464.

Craane, B., Dijkstra, P.U., Stappaerts, K. and De Laat, A. (2012), "Methodological quality of a systematic review on physical therapy for temporomandibular disorders: influence of hand search and quality scales." *Clinical Oral Investigations*, 16 (1): 295–303. doi:10.1007/s00784-010-0490-y.

De Mauro, A., Greco, M. and Grimaldi, M. (2015), "What is big data? A consensual definition and a review of key research topics." *AIP Conference Proceedings*, 1644 (1): 97–104. doi:10.1063/1.4907823.

Del Giudice, M., Chierici, R., Mazzucchelli, A. and Fiano, F. (2021), "Supply chain management in the era of circular economy: the moderating effect of big data." *The international journal of logistics management*, 32 (2): 337–356. doi:10.1108/IJLM-03-2020-0119.

Delgosha, M.S., Hajiheydari, N. and Fahimi, S.M. (2020), "Elucidation of big data analytics in banking: a four-stage Delphi study." *Journal of Enterprise Information Management*, 34 (6): 1577–1596. doi:10.1108/JEIM-03-2019-0097.

Dolgui, A. and Ivanov, D. (2021), "Ripple effect and supply chain disruption management: new trends and research directions." *International journal of production research*, 59 (1): 102–109. doi:10.1080/00207543.2021.1840148.

Durach, C.F., Kembro, J. and Wieland, A. (2017), "A New Paradigm for Systematic Literature Reviews in Supply Chain Management." *Journal of Supply Chain Management*, 53 (4): 67–85. doi:10.1111/jscm.12145.

Dwivedi, A., Moktadir, M.A., Chiappetta Jabbour, C.J. and de Carvalho, D.E. (2022), "Integrating the circular economy and industry 4.0 for sustainable development: Implications for responsible footwear production in a big data-driven world." *Technological forecasting & social change*, 175: 121335-. doi:10.1016/j.techfore.2021.121335.

El-Kassar, A.-N. and Singh, S.K. (2019), "Green innovation and organizational performance: The influence of big data and the moderating role of management commitment and HR practices." *Technological Forecasting and Social Change*, 144: 483–498. doi:10.1016/j.techfore.2017.12.016.

Esper, T.L., Defee, C.C. and Mentzer, J.T. (2010), "A framework of supply chain orientation." *The international journal of logistics management*.

Gawankar, S.A., Gunasekaran, A. and Kamble, S. (2020), "A study on investments in the big data-driven supply chain, performance measures and organisational performance in Indian retail 4.0 context.” *International journal of production research*, 58 (5): 1574–1593. doi:10.1080/00207543.2019.1668070.

Ghaleb, E.A., Dominic, P.D., Fati, S.M., Muneer, A. and Ali, R.F. (2021), “The assessment of big data adoption readiness with a technology–organization–environment framework: A perspective towards healthcare employees.” *Sustainability*, 13 (15): 8379-. doi:10.3390/su13158379.

Gligor, D., Feizabadi, J., Pohlen, T., Maloni, M. and Ogden, J.A. (2022), “The impact of the supply chain orientation fit between supply chain members: A triadic perspective.” *Journal of Business Logistics*, n/a (n/a). doi:10.1111/jbl.12304.

Hajek, P. and Abedin, M.Z. (2020), “A Profit Function-Maximizing Inventory Backorder Prediction System Using Big Data Analytics.” *IEEE access*, 8: 58982–58994. doi:10.1109/ACCESS.2020.2983118.

Hajiheydari, N., Delgosha, M.S. and Olya, H. (2021), “Scepticism and resistance to IoMT in healthcare: Application of behavioural reasoning theory with configurational perspective.” *Technological Forecasting and Social Change*, 169: 120807. doi:10.1016/j.techfore.2021.120807.

Han, G., Pan, X. and Zhang, X. (2024), “Big data-driven risk decision-making and safety management in agricultural supply chains.” *Quality assurance and safety of crops & food*, 16 (1): 121–138. doi:10.15586/qas.v16i1.1445.

Hazen, B.T., Boone, C.A., Ezell, J.D. and Jones-Farmer, L.A. (2014), “Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications.” *International Journal of Production Economics*, 154: 72–80. doi:10.1016/j.ijpe.2014.04.018.

Huang, R., Qu, S., Gong, Z., Goh, M. and Ji, Y. (2020), “Data-driven two-stage distributionally robust optimization with risk aversion.” *Applied soft computing*, 87: 105978-. doi:10.1016/j.asoc.2019.105978.

Jadhav, A., Orr, S. and Malik, M. (2019), “The role of supply chain orientation in achieving supply chain sustainability.” *International Journal of Production Economics*, 217: 112–125.

Ji, G., Yu, M., Tan, K.H., Kumar, A. and Gupta, S. (2022), “Decision optimization in cooperation innovation: the impact of big data analytics capability and cooperative modes.” *Annals of operations research*, 333 (2–3): 871–894. doi:10.1007/s10479-022-04867-1.

Joemsittiprasert, W., Sommanawat, K. and Vipaporn, T. (2019), “*Can big data benefits bridge between data driven supply chain orientation and financial performance? Evidence from manufacturing sector of Thailand*.”, 8: 597–609.

Kamble, S.S. and Gunasekaran, A. (2020), “Big data-driven supply chain performance measurement system: a review and framework for implementation.” *International journal of production research*, 58 (1): 65–86. doi:10.1080/00207543.2019.1630770.

Kamboj, S. and Rana, S. (2023), “Big data-driven supply chain and performance: a resource-based view.” *TQM journal*, 35 (1): 5–23. doi:10.1108/TQM-02-2021-0036.

Karaboga, T., Zehir, C., Tatoglu, E., Karaboga, H.A. and Bouguerra, A. (2023), “Big data analytics management capability and firm performance: the mediating role of data-driven culture.” *Review of managerial science*, 17 (8): 2655–2684. doi:10.1007/s11846-022-00596-8.

Khan, W., Nisar, Q.A., Roomi, M.A., Nasir, S., Awan, U. and Rafiq, M. (2024), “Green human resources management, green innovation and circular economy performance: the role of big data analytics and data-driven culture.” *Journal of environmental planning and management*, 67 (10): 2356–2381. doi:10.1080/09640568.2023.2189544.

Khoei, M.A., Aria, S.S., Gholizadeh, H., Goh, M. and Cheikhrouhou, N. (2023), “Big data-driven optimization for sustainable reverse logistics network design.” *Journal of ambient intelligence and humanized computing*, 14 (8): 10867–10882. doi:10.1007/s12652-022-04357-z.

Kokkinou, A., Mandemakers, A. and Mitas, O. (2023), “Developing resilient and robust supply chains through data analytic capability.” *Continuity & resilience review (Online)*, 5 (3): 320–342. doi:10.1108/CRR-07-2023-0013.

Kokshagina, O., Le Masson, P. and Luo, J. (2024), “Beyond the data fads: Impact of big data on contemporary innovation and technology management.” *Technovation*, 134: 103026-. doi:10.1016/j.technovation.2024.103026.

Kristoffersen, E., Blomsma, F., Mikalef, P. and Li, J. (2020), “The smart circular economy: A digital-enabled circular strategies framework for manufacturing companies.” *Journal of Business Research*, 120: 241–261. doi:10.1016/j.jbusres.2020.07.044.

Lai, K., Wong, C.W.Y. and Lam, J.S.L. (2015), “Sharing environmental management information with supply chain partners and the performance contingencies on environmental munificence.” *International Journal of Production Economics*, 164: 445–453. doi:10.1016/j.ijpe.2014.12.009.

Le, T.T. (2023), “Linking big data, sustainable supply chain management and corporate performance: the moderating role of circular economy thinking.” *The international journal of logistics management*, 34 (3): 744–771. doi:10.1108/IJLM-01-2022-0011.

Liu, P. (2019), “Pricing policies and coordination of low-carbon supply chain considering targeted advertisement and carbon emission reduction costs in the big data environment.” *Journal of Cleaner Production*, 210: 343–357. doi:10.1016/j.jclepro.2018.10.328.

Liu, X., Li, S., Wang, X. and Zhang, C. (2023), “Data-driven supply chain orientation and innovation: the role of capabilities and information complexity.” *European journal of innovation management*. doi:10.1108/EJIM-01-2023-0045.

Liu, Y., Fang, W., Feng, T. and Gao, N. (2022), “Bolstering green supply chain integration via big data analytics capability: the moderating role of data-driven decision culture.” *Industrial management + data systems*, 122 (11): 2558–2582. doi:10.1108/IMDS-11-2021-0696.

Luo, D., Thevenin, S. and Dolgui, A. (2023), “A state-of-the-art on production planning in Industry 4.0.” *International journal of production research*, 61 (19): 6602–6632. doi:10.1080/00207543.2022.2122622.

Ma, S., Zhang, Y., Liu, Y., Yang, H., Lv, J. and Ren, S. (2020), “Data-driven sustainable intelligent manufacturing based on demand response for energy-intensive industries.” *Journal of cleaner production*, 274: 123155-. doi:10.1016/j.jclepro.2020.123155.

Maheshwari, S., Gautam, P. and Jaggi, C.K. (2021), “Role of Big Data Analytics in supply chain management: current trends and future perspectives.” *International journal of production research*, 59 (6): 1875–1900. doi:10.1080/00207543.2020.1793011.

Margaritis, I., Madas, M. and Vlachopoulou, M. (2022), “Big Data Applications in Food Supply Chain Management: A Conceptual Framework.” *Sustainability*, 14 (7): 4035-. doi:10.3390/su14074035.

Mentzer, J.T., DeWitt, W., Keebler, J.S., Min, S., Nix, N.W., Smith, C.D. and Zacharia, Z.G. (2001), “Defining supply chain management.” *Journal of Business Logistics*, 22 (2): 1–25. doi:10.1002/j.2158-1592.2001.tb00001.x.

Moher, D., Liberati, A., Tetzlaff, J. and Altman, D.G. (2009), “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.” *BMJ*, 339: b2535. doi:10.1136/bmj.b2535.

Naeem, M., Ozuem, W., Howell, K. and Ranfagni, S. (2023), “A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research.” *International Journal of Qualitative Methods*, 22: 16094069231205789. doi:10.1177/16094069231205789.

Nguyen, T., Zhou, L., Spiegler, V., Ieromonachou, P. and Lin, Y. (2018), “Big data analytics in supply chain management: A state-of-the-art literature review.” *Computers & Operations Research*, 98: 254–264. doi:10.1016/j.cor.2017.07.004.

Pan, Q., Luo, W. and Fu, Y. (2022), “A csQCA study of value creation in logistics collaboration by big data: A perspective from companies in China.” *Technology in society*, 71: 102114-. doi:10.1016/j.techsoc.2022.102114.

Papadopoulos, T., Gunasekaran, A., Dubey, R. and Fosso Wamba, S. (2017), “Big data and analytics in operations and supply chain management: managerial aspects and practical challenges.” *Production Planning & Control*, 28 (11–12): 873–876. doi:10.1080/09537287.2017.1336795.

Prajogo, D. and Olhager, J. (2012), “Supply chain integration and performance: The effects of long-term relationships, information technology and sharing, and logistics integration.” *International Journal of Production Economics*, 135 (1): 514–522. doi:10.1016/j.ijpe.2011.09.001.

Pratap, S., Jauhar, S.K., Gunasekaran, A. and Kamble, S.S. (2024), “Optimizing the IoT and big data embedded smart supply chains for sustainable performance.” *Computers & industrial engineering*, 187: 109828-. doi:10.1016/j.cie.2023.109828.

Qrunfleh, S. and Tarafdar, M. (2014), “Supply chain information systems strategy: Impacts on supply chain performance and firm performance.” *International Journal of Production Economics*, 147: 340–350. doi:10.1016/j.ijpe.2012.09.018.

Rui, G. and Li, M. (2024), “Utilizing Internet Big Data and Machine Learning for Product Demand Forecasting and Analysis of Its Economic Benefits.” *Tehnički vjesnik*, 31 (4): 1385–1394. doi:10.17559/TV-20240318001408.

Sabharwal, R. and Miah, S.J. (2021), “A new theoretical understanding of big data analytics capabilities in organizations: a thematic analysis.” *Journal of Big Data*, 8 (1): 159. doi:10.1186/s40537-021-00543-6.

Schoenherr, T. and Speier-Pero, C. (2015), “Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential.” *Journal of Business Logistics*, 36 (1): 120–132. doi:10.1111/jbl.12082.

Singh, A., Kumari, S., Malekpoor, H. and Mishra, N. (2018), “Big data cloud computing framework for low carbon supplier selection in the beef supply chain.” *Journal of Cleaner Production*, 202: 139–149. doi:10.1016/j.jclepro.2018.07.236.

Singh, C.K., Barme, E., Ward, R., Tupikina, L. and Santolini, M. (2022), “Quantifying the rise and fall of scientific fields.” *PLOS ONE*, 17 (6): e0270131. doi:10.1371/journal.pone.0270131.

Sundarakani, B., Ajaykumar, A. and Gunasekaran, A. (2021), “Big data driven supply chain design and applications for blockchain: An action research using case study approach.” *Omega (Oxford)*, 102: 102452-. doi:10.1016/j.omega.2021.102452.

Thekkoote, R. (2022), “Understanding big data-driven supply chain and performance measures for customer satisfaction.” *Benchmarking : an international journal*, 29 (8): 2359–2377. doi:10.1108/BIJ-01-2021-0034.

Tripathi, S., Bachmann, N., Brunner, M., Rizk, Z. and Jodlbauer, H. (2024), “Assessing the current landscape of AI and sustainability literature: identifying key trends, addressing gaps and challenges.” *Journal of big data*, 11 (1): 65–68. doi:10.1186/s40537-024-00912-x.

Tseng, M.-L., Tran, T.P.T., Ha, H.M., Bui, T.-D. and Lim, M.K. (2021), “Sustainable industrial and operation engineering trends and challenges Toward Industry 4.0: a data driven analysis.” *Journal of industrial and production engineering*, 38 (8): 581–598. doi:10.1080/21681015.2021.1950227.

Verma, A., Kuo, Y.-H., Kumar, M.M., Pratap, S. and Chen, V. (2023), “A data analytic-based logistics modelling framework for E-commerce enterprise.” *Enterprise information systems*, 17 (6). doi:10.1080/17517575.2022.2028195.

Wamba, S.F., Dubey, R., Gunasekaran, A. and Akter, S. (2020), “The performance effects of big data analytics and supply chain ambidexterity: The moderating effect of environmental dynamism.” *International Journal of Production Economics*, 222: 107498. doi:10.1016/j.ijpe.2019.09.019.

Wang, J., Xu, C., Zhang, J. and Zhong, R. (2022), “Big data analytics for intelligent manufacturing systems: A review.” *Journal of manufacturing systems*, 62: 738–752. doi:10.1016/j.jmsy.2021.03.005.

Wu, F., Yeniyurt, S., Kim, D. and Cavusgil, S.T. (2006), “The impact of information technology on supply chain capabilities and firm performance: A resource-based view.” *Industrial Marketing Management*, 35 (4): 493–504. doi:10.1016/j.indmarman.2005.05.003.

Younis, H. and Wuni, I.Y. (2023), “Application of industry 4.0 enablers in supply chain management: Scientometric analysis and critical review.” *Heliyon*, 9 (11): e21292–e21292. doi:10.1016/j.heliyon.2023.e21292.

Yu, W., Chavez, R., Jacobs, M.A. and Feng, M. (2018), “Data-driven supply chain capabilities and performance: A resource-based view.” *Transportation Research Part E: Logistics and Transportation Review*, 114: 371–385. doi:10.1016/j.tre.2017.04.002.

Yu, W., Zhao, G., Liu, Q. and Song, Y. (2021), “Role of big data analytics capability in developing integrated hospital supply chains and operational flexibility: An organizational information processing theory perspective.” *Technological forecasting & social change*, 163: 120417-. doi:10.1016/j.techfore.2020.120417.

Zaidi, S.A.H., Khan, S.A. and Chaabane, A. (2024), “Unlocking the potential of digital twins in supply chains: A systematic review.” *Supply Chain Analytics*, 7: 100075-. doi:10.1016/j.sca.2024.100075.

Zekhnini, K., Chaouni Benabdellah, A. and Cherrafi, A. (2023), “A multi-agent based big data analytics system for viable supplier selection.” *Journal of intelligent manufacturing*. doi:10.1007/s10845-023-02253-7.

Zhao, S. and You, F. (2020), “Distributionally robust chance constrained programming with generative adversarial networks (GANs).” *AIChE journal*, 66 (6): n/a. doi:10.1002/aic.16963.

Zhong, R.Y., Newman, S.T., Huang, G.Q. and Lan, S. (2016), “Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives.” *Computers & Industrial Engineering*, 101: 572–591. doi:10.1016/j.cie.2016.07.013.

Precise Inventory Forecasting based on Sales History and Reviews: Intellectual Inventory Planning by Machine Learning and OpenAI Integration

Honglei Li
Department of Computer & Information Sciences
University of Northumbria at Newcastle Upon Tyne
Honglei.Li@northumbria.ac.uk

Sowjanuya Jamadar
Accenture UK
sowjanya.94@gmail.com

Completed Research

Abstract

Effective stock management is expected to solver the issue of overstocking and understock that may cause revenue losses. This research aims to present an integrated strong inventory management system using cutting-edge technology implementation applied on the publicly available market data. Raw sales data is pre-processed and fine-tuned, time series analysis and a range of machine learning strategies were applied on tuned dataset to produce accurate demand forecasts. Furthermore, we integrated Open AI for evaluating the product's generalized review to help with decision-making in the simplest and most efficient way. It is demonstrated that the simplest strategy to control stock levels to maximize profitability could be achieved by integrating the of projected sales and customer reviews.

Keywords: Demand forecasting, OpenAI, Sentiment analysis, Machine Learning techniques, and Inventory management.

1.0 Introduction

The use of generative artificial intelligence (AI) and big data analysis has significantly transformed industries across the board. Machine learning (ML) is now central to the development of forecasting techniques that analyse all aspects of management, providing valuable insights for businesses to understand and respond to fluctuations in their operations. Inventory planning is a critical function for any business, as it helps determine when to reorder products and in what quantities, while considering various control mechanisms (Tsou 2008). Many approaches view this challenge as a multi-objective optimization, where the goal is to minimize ordering and logistics costs while maximizing service levels. Sales forecasting plays a key role in helping businesses plan their marketing, sales management, manufacturing, purchasing, and storage activities, ultimately leading to increased revenue and reduced production losses (Mentzer and Bienstock 1998).

Existing research highlights two major factors that influence customer purchasing decisions, customer recommendations through word-of-mouth and the effect of commercial advertisements and mass media. Researchers have combined fundamental models, including machine learning and sentiment analysis of customer reviews, along with sentiment index methods, to improve the accuracy of sales forecasting. OpenAI's ChatGPT offers a variety of features in a user-friendly format, with an API available for easy integration. This has inspired the creation of a high-performance inventory management system that merges two key concepts: sales forecasting and sentiment analysis, while incorporating the latest deep learning models and OpenAI technologies.

Traditional inventory management systems typically address challenges through demand forecasting or sentiment analysis of ratings and reviews. However, these methods often lead to problems such as overstocking, stockouts, and inaccurate forecasting, highlighting the need for a shift in approach. This research seeks to tackle these issues by integrating machine learning (ML) techniques with OpenAI's natural language processing capabilities. By combining sentiment analysis with sales forecasting, this approach offers a fresh perspective on inventory planning, aiming to improve accuracy and efficiency.

2.0 Literature Review

2.1 Overview of Inventory management

Inventory management, as described by Muller (2019), is a broad concept encompassing all procedures and best practices for monitoring inventory throughout a product's entire life cycle within a business. Effective inventory management helps maintain accurate supply records, determine appropriate pricing, and address fluctuations in demand without compromising product quality or customer satisfaction. In industries such as retail, food services, hospitality, and manufacturing, inventory shortages can have severe consequences. Koumanakos (2008) found that inventory management, whether efficient or inefficient, is a key factor influencing a firm's performance, although macroeconomic conditions and business-specific factors also play a role. The primary goal of inventory supervisors is to minimize costs or maximize profits while ensuring customer needs are met. Muller (2019) emphasized that balancing the risks of overstocking and shortages is particularly challenging for businesses with complex logistics systems. Inventory is generally considered a current asset with a tangible value, but experts debate whether it should also be viewed as a liability. To qualify as a current asset, stock must be regularly counted and recorded.

The inventory management systems adopted by businesses vary based on industry size, company type, and the quantity of goods involved. Inventory tracking is typically managed using one of two main approaches: periodic or perpetual updates. Common inventory management methodologies include ABC (Always Better Control) Analysis, the Economic Order Quantity (EOQ) Model, the FSN (Fast, Slow, and Non-moving) Method, Just-in-Time (JIT) inventory control, and the Material Requirement Planning (MRP) Method.

To address challenges such as fluctuating market prices, complex storage processes, and resource limitations, logistics companies rely on efficient inventory management systems. Automated inventory tracking and organization enhance operational efficiency, benefiting businesses by streamlining production and reducing manual labor. According to Ugbebor et al. (2024), cloud-based inventory management solutions are increasingly preferred over on-premises alternatives due to cost savings.

2.2 Strategic Insights of sales forecasting for Inventory Control

Forecasting future sales graphs requires a combination of expert business knowledge, present-day sales, past sales information, and predictive analytics. Sales forecasting is critical to the long-term profitability and survival of any business. Keeping track of sales is primarily used for assessing whether the business will meet its sales targets and whether the stock levels are adequate. Sales forecasting is critical because it allows businesses to adjust as necessary to boost revenues from sales. McCarthy et al. (2006) presented the findings of a survey intended to investigate how forecasting of sales practices in management evolved over the last 20 years. A web-based poll of forecasting managers was used to investigate patterns regarding forecasting management, regularity, fulfilment, utilization, and correctness among companies from various industries. The results showed that people were less familiar with methods for forecasting and had a lower percentage of forecasted accuracy. Decker and Gnibba-Yukawa (2010) stated that truthful forecasts for sales are critical and looked at real-world research on consumer behaviour in high-technology markets. The findings demonstrated exactly how to get around the drawbacks of traditional diffusion models by developing a utility-based model that specifically incorporates industry-specific purchasing behaviour and generates precise projections of sales for technological goods in the growth stage. The functionality of the new model has been demonstrated using real-world sales data from the CD, DVD player and digital camera markets. Kourentzes et al. (2020) proposed a method for combining competing inventory objectives, such as meeting customer demand while eradicating surplus stock, and then using the resulting expense function to determine inventory the best settings for forecasting algorithms. The proposed optimization renders the use of specific criteria unreliable, forcing one to revert to more broadly applicable strategies such as cross-validated procedures. Finally, examined the inventory decision-making when establishing an adequate simulation-based costing function. Voulgaridou et al. (2009) put forward new product sales forecasting using the standard MCDA (Multicriteria Decision Making Analysis method) in yet another scientific domain, allowing stakeholders to think about not only past statistics but also current circumstances, patterns, as well as an expert's implicit insights. Consequently, the overall forecasting and the final choice have improved.

2.3 Overview of Sentiment Analysis and its importance in inventory management

Sentiment analysis is a simple yet powerful method for assessing the quality of products by analyzing consumer opinions, especially people's attitudes and inclinations through data mining techniques. Wood et al. (2013) explained how sentiment analysis is utilized in operations and sales planning, particularly for demand forecasting. Their study was the first to highlight the role of sentiment assessment in enhancing Sales and Operations Planning (S&OP) by identifying market fluctuations that impact supply chains. However, despite generating resources to support this methodology, further empirical research is needed to validate these claims.

Sentiment analysis helps gauge public sentiment and reactions to specific products, individuals, or ideas. It delves into customer emotions, offering businesses deeper insights into how consumers perceive their brand. According to IDC, approximately 80% of historical data will be unstructured by 2025, posing a challenge for companies looking to extract meaningful insights. Sentiment analysis plays a crucial role in addressing this issue by transforming raw data into actionable information. Before leveraging sentiment analysis effectively, businesses must first understand its purpose. It has become a vital tool for identifying and resolving product-related issues. With the advancement of AI-based tools, sentiment analysis is no longer a complex task, with companies like Clootrack, Gavagai, and The Lionbridge Company, Inc. offering such services (Ahmed et al. 2022).

Word-of-mouth (WOM) is a significant factor influencing consumer purchasing decisions. With the rise of internet technologies, online word-of-mouth has gained popularity through blogs and customer reviews. To encourage consumers to share feedback, many e-commerce platforms, retailers, and shipping companies have established online review systems. As a result, customer behavior and purchasing patterns have evolved. Online reviews now play a critical role in consumer decision-making, influencing choices on product purchases, movie selections, stock investments, and more (Ryu and Han 2010). Potential buyers rely on public reviews to guide their purchasing decisions, as these reviews often convey personal emotions such as satisfaction, frustration, disappointment, or praise. Over the past decade, sentiment analysis techniques have been widely used to assess emotions expressed in online reviews (Prabowo and Thelwall 2009). Multiple studies suggest that online word-of-mouth significantly impacts consumer behavior and product sales (Liu 2006), demonstrating that factors like review volume, feedback quality, and expressed opinions influence purchasing trends.

2.4 Machine learning for sales forecasting

Traditional forecasting methods such as auto-regressive models, ARIMA, and seasonal-ARIMA (SARIMA) often fail to capture all the hidden patterns within time series data, particularly in quantitative predictions for the automotive industry. Shetty and Buktar (2022) analysed ARIMA and SARIMA models using a dataset from an Indian automobile company and compared their performance with the deep learning model Long Short-Term Memory (LSTM). Their findings indicate that LSTM outperforms ARIMA by 92% and SARIMA by 42.5% in terms of accuracy. Traditional forecasting techniques face difficulties in handling large datasets for sales predictions, whereas newer data mining methods offer more effective solutions.

To determine the most suitable predictive machine learning model for different scenarios, Cheriyan et al. (2018) evaluated three approaches: Generalized Linear Modeling (GLM), Decision Tree (DT), and Gradient Boost Tree (GBT). Among these, the Gradient Boost Algorithm demonstrated the highest accuracy in estimating and predicting future sales, making it the most effective model. Artificial Neural Networks (ANNs) have emerged as a powerful technology for modelling and identifying complex data patterns over the past decade. Their ability to simulate intricate, nonlinear relationships without requiring assumptions about the underlying data-generating process gives them a significant advantage over traditional economic modeling techniques. Hornik et al. (1989) compared ANNs with exponential smoothing, ARIMA models, and multivariate regression, concluding that, on average, ANNs outperform conventional statistical methods.

In modern business environments, managers are responsible for both decision-making and prediction-making—two essential administrative functions. According to Alon et al. (2001), integrating prediction and decision-making methods enables researchers to leverage prior knowledge more effectively. The Analytical Hierarchy Process (AHP) is a comprehensive system designed to evaluate multiple factors, making it an essential tool in decision-making. In deep learning applications, AHP is used to establish, assess, and categorize measurements. To assess the effectiveness of this approach, an automobile price prediction model was implemented, with results compared against neural networks to evaluate performance.

2.5 Investigating OpenAI and its importance in Inventory Management

Artificial intelligence (AI) technology has brought about an immense change in businesses. OpenAI, a preeminent AI research institution that has been crucial in fostering inventiveness and altering the supply chain scene, is one such powerful in effect. According to Hendriksen (2023), Intelligent warehouse management, improved demand forecasting and planning, optimized the distribution and logistics, collaborative decision-making, risk management, and resilience can all be invented with OpenAI. By means of the Azure OpenAI Service, OpenAI's partnership with Microsoft Azure will revolutionize the supply chain sector by giving companies the opportunity to utilize strong and easily deployable artificial intelligence (AI) capabilities. The future of supply chain administration could be drastically altered by the integration of OpenAI's models with the Microsoft Azure OpenAI Service, which offers previously unheard-of potential for increasing customer satisfaction, cutting costs, and improving operational efficiency. According to Borodavko et al. (2021), inventory optimization can be created by utilizing AI insights to eliminate traditional manual procedures.AI can optimize inventory management in three different ways. Organizing the restocking of inventory, Arrival time estimate of a product and safety stock control. Customers’ preferences for how and when to receive their products should also be considered in this customer-behaviour-centric model. Inventory management systems can increase store levels of inventory by using AI to analyse customer satisfaction choices and purchasing patterns.

AI can enable deep learning designs to outperform machine learning in the solution of complex mathematical issues. Desai and Oza (2021) clearly explains, the deep learning models have been combined with automatic optimization of the method for extracting features as opposed to machine learning models. Producing Using deep learning algorithms, the autoregressive language model Pre-trained Transformer 3 (GPT-3) generates text that resembles that of a human. The autoregressive model uses random interpreting for predicting the result. Working with an extensive collection of NLP data and fine-tuning it for a specific purpose to yield results is something that is difficult in today's world. With 175 billion parameters and over 10,000 non-sparse linguistic models, GPT-3 is a model built using deep learning that performs evaluations swiftly and efficiently. Many of GPT-3's applications demonstrate how impressive it is; the details are shown in the following examples. Deliver regular expressions using a variety of use cases expressed in simple English sentences. GPT-3 can automatically create charts and plots from simple English. It can be applied to the developing of interactive quizzes and tailored e-learning programs. By outlining the data source and the desired model output, this GPT-3 can write a model using machine learning (ML). It functions as a more accurate, easier and sophisticated endorser, an Interactive Voice Response (IVR) operator, and a self-learning resume developer despite of guidance and huge data training .

The opinions and sentiments of individuals who initially used ChatGPT, an artificial intelligence-driven advanced language model created by OpenAI, present insightful information about the advantages and disadvantages of this advancement in technology. A research project on the sentiment analysis of tweets on ChatGPT was carried out by Korkmaz et al. (2023), in order to accurately evaluate users' opinions during the very first two months after ChatGPT's launch. Pre-processed and sentiment-analysed, about 78,8000 English tweets were utilised with word-based sentiment dictionaries (AFINN, Bing, and NRC). According to the research, the majority of ChatGPT's original users thought the service was effective and they were pleased. However, some users also displayed negative emotions like worry and fear. Reinforcement learning (RL) fine-tuning of conceptual framework for language is developed for natural language processing (NLP) tasks such as substantial sentiment analysis as part of the OpenAI application. Many fine-tune modelling applications, such as face recognition, summarization of texts, replication activity tasks, and sentence assessment, are motivated by open AI systems (Desai and Oza 2021).

3. Research Methods

Artificial intelligence (AI) has significantly transformed businesses across various industries. One of the key players in this technological shift is OpenAI, a leading AI research institution that has played a crucial role in driving innovation and reshaping supply chain operations. According to Hendriksen (2023), OpenAI's AI-driven solutions enhance intelligent warehouse management, demand forecasting, distribution and logistics optimization, collaborative decision-making, risk management, and supply chain resilience. Through its collaboration with Microsoft Azure via the Azure OpenAI Service, OpenAI is poised to revolutionize the supply chain sector by offering businesses powerful and easily deployable AI capabilities. The integration of OpenAI’s models with Microsoft Azure is expected to enhance customer satisfaction, reduce costs, and improve overall operational efficiency. Borodavko et al. (2021) highlights that AI-driven insights can replace traditional manual processes, leading to more efficient inventory optimization. AI can improve inventory management in three keyways: streamlining the restocking process, estimating product arrival times, and controlling safety stock levels. Additionally, AI-powered inventory management systems analyze customer preferences and purchasing patterns to ensure inventory availability aligns with consumer demand.

Deep learning models have proven to be more effective than traditional machine learning techniques in solving complex mathematical problems. Desai and Oza (2021) explain that deep learning models incorporate automatic feature extraction, enhancing their performance over conventional machine learning models. One notable example is GPT-3, an autoregressive language model that generates human-like text using deep learning algorithms. Built with 175 billion parameters and over 10,000 non-sparse linguistic models, GPT-3 processes extensive natural language data and fine-tunes it for specific tasks. GPT-3’s capabilities include generating regular expressions from simple English sentences, creating charts and plots, developing interactive quizzes, and designing personalized e-learning programs. Additionally, GPT-3 can be used to build machine learning models based on predefined data sources and desired outputs. It also serves as an intelligent recommender, an Interactive Voice Response (IVR) operator, and a self-learning resume generator, operating with minimal guidance and extensive data training (Brown et al. 2020).

The release of ChatGPT, an advanced AI-driven language model developed by OpenAI, has generated valuable insights into its advantages and limitations. Korkmaz et al. (2023) conducted a sentiment analysis of tweets during the first two months following ChatGPT's launch. Using word-based sentiment dictionaries (AFINN, Bing, and NRC), they analyzed approximately 788,000 English-language tweets. Their findings revealed that most early users had a positive experience and were satisfied with ChatGPT's performance. However, some users expressed negative emotions, such as concern and apprehension. OpenAI's applications also extend to reinforcement learning (RL), where fine-tuned language models are developed for natural language processing (NLP) tasks, including large-scale sentiment analysis. Additionally, OpenAI's fine-tuned models are widely used in applications such as facial recognition, text summarization, task automation, and sentence evaluation (Desai and Oza 2021).

3.1 Data sources

Since a combined dataset of sales and reviews was not publicly available, this research has made use of two separate datasets. The following is a list of the datasets.

1. Business Data for the Automobile Sector:

This data set is taking from the publicly available source, Kaggle, having various datasets for every technology, where community of people learn the different projects and competitions performed ("Automotive sector comprehensive business data-2023", Kaggle.com)

This dataset contains information on the automotive industry, including multiple kinds of cars as well as comprehensive sales data from 2016 to 2019 in several CSV files with various details, as shown below:

Customer: Provides details of customers who buy the products as customer_id, employee_id, first name, last name, dob, phone, email, address.

Employee: This file provides employee_id, store_id, first name, last name, dob, phone, email, status, salary, street, city and country, the details of employees working in the store for selling car products.

Store: Store file gives store_id, employee_id, store name, phone, street, city, country, email, postcode details representing store address, employees associated with every store

Category: The category files describe the car category and selling details as category_id, category_name, rating, quantity_sold , being_manufactured, total_sold_value.

Product: The product is associated car models for every category of car type. This file details product_id, provider_id, category_id, product_name, model, year, color, km, price, stock.

Here the stock parameter is manually added to the product file for stock level classification and prediction purpose.

Provider: This file gives the details of providers of the car to the store.

Order: This acts as main data where it provides the sales data like order date, shipped date, status of sale and associated customer id who purchased, employee id who sold the car, store id in which the sale has happened. Here the review parameter is added manually from the second data set.

Five data files are used in this analysis: order, product, category, store, customer. Details of the dataset are provided below in Table 1.

S. No	Dataset	Used/ Not Used	Variables (Columns)	Count of data (Rows)
1	Order	Used	10	1700
2	Product	Used	10	16723
3	Category	Used	7	2672
4	Store	Used	9	1317
5	Customer	Used	9	5727
6	Employee	Not Used	12	4250
7	Provider	Not Used	8	1426

Table 1. Details of the dataset 1

2. Car Reviews Database

Since reviews are essential for sentiment analysis, and the previously mentioned dataset lacks them, data from the "Sentiment Analysis of Car Reviews" dataset on Kaggle—which includes details on car models, manufacturing years, and reviews—has been incorporated. This dataset is used to construct an optimal dataset for implementing the inventory system.

3.2 Big Data Analytics

Rather than being a single tool, this inventory planning system integrates with multiple frameworks and databases, such as NoSQL, as well as visualization software like Tableau and cluster management resources like YARN. For handling large-scale data operations, frameworks such as Hadoop, MapReduce, and Spark efficiently manage both structured and unstructured data through batch and stream processing. Due to its ability to analyze large datasets rapidly, this technology helps organizations reduce costs, enhance product development, and gain valuable market insights (Tableau).

The Apache Spark data processing framework is utilized in this big data analytics project due to its ease of use, flexibility, seamless integration, and fast adaptability to SQL and ANSI SQL. It is also capable of handling both structured and unstructured data (Salloum et al. 2016). Apache Spark is emerging as a leading big data analytics framework, offering a strong foundation and libraries for broadcasting, graph visualization, and predictive modelling.

3.4 Data Analysis Methods

3.4.1 Heat Map and its Correlation

Firstly, scatter matrix is plotted to identify pair-wise relationships for the variables no. of models in a category, rating of category, quantity sold, and total sold value. Pearson and Spearman (Cortez et al. 2018). The correlation coefficients used are Pearson coefficient, and spearman's rank coefficient.

3.4.2 Time Series Analysis

Before conducting time series analysis, car sales trends are examined month by month for each year, with the data points plotted as shown in Figure 1. In this dataset, car sales represent a recurring sequence of sales data over multiple years, recorded at equal intervals. By applying time series analysis techniques, underlying trends in the data have been identified. A time series consists of recurring data, and understanding its characteristics over time is essential for accurate forecasting. The two most widely used forecasting techniques are exponential smoothing and moving averages. For this dataset, the moving averages approach is the most suitable. The additive time series decomposition method is employed to identify patterns and trends within the data.

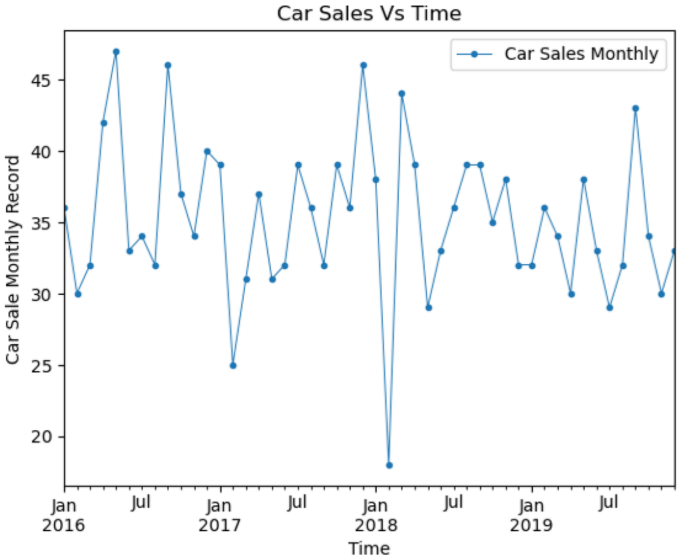
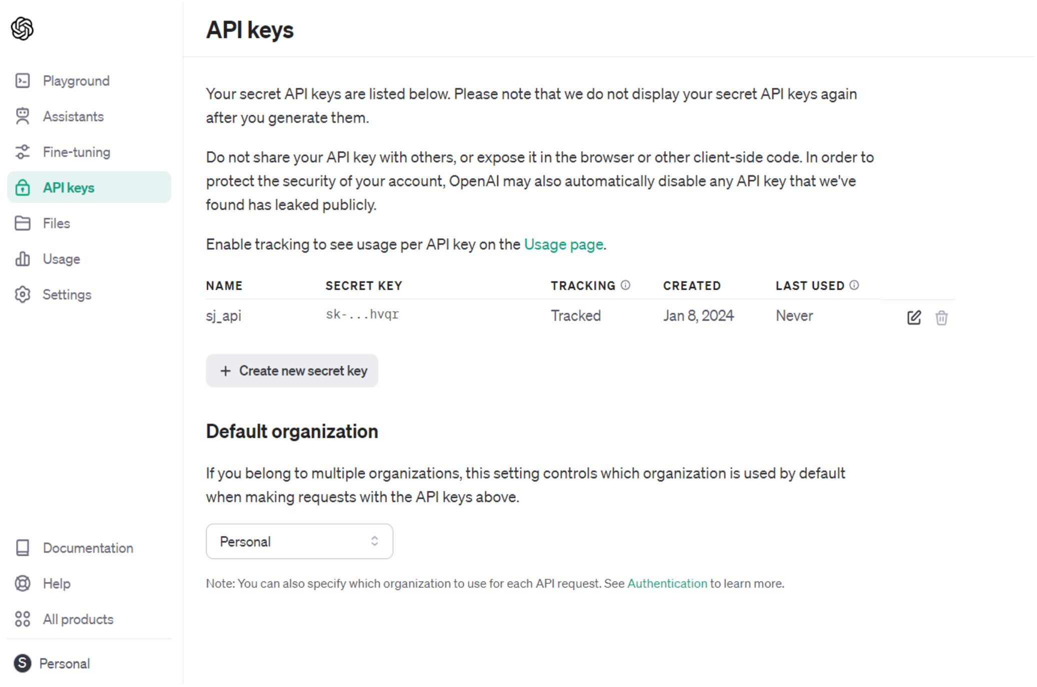


Figure 1: Car sales over the time

The prediction models used this this study include ARIMA model (Ariyo et al. 2014). The machine learning models used in this study are random forest and gradient boosted decision tree model.

3.7 Process of Integrating OpenAI

ChatGPT by OpenAI offers a broad range of versatile capabilities that are domain-neutral and can assist with a variety of everyday tasks and challenges. For this reason, OpenAI provides a user-friendly interface that allows anyone to log in and easily ask questions or interact with ChatGPT. Additionally, OpenAI has developed an API that simplifies the integration of GPT models into various applications, enabling the use of the model within any desired software. For personal use, the API key provides limited trial access for a certain period. For official or commercial use, subscription plans are available for purchase on the OpenAI website. In this research, only the most basic trial version of the API key has been created for integration. The page displaying the generated API key along with the personal account details is shown in Figure 2.



API keys

Your secret API keys are listed below. Please note that we do not display your secret API keys again after you generate them.

Do not share your API key with others, or expose it in the browser or other client-side code. In order to protect the security of your account, OpenAI may also automatically disable any API key that we've found has leaked publicly.

Enable tracking to see usage per API key on the Usage page.

NAME	SECRET KEY	TRACKING ⓘ	CREATED	LAST USED ⓘ
sj_api	sk-...hvqr	Tracked	Jan 8, 2024	Never

Default organization

If you belong to multiple organizations, this setting controls which organization is used by default when making requests with the API keys above.

Personal

Note: You can also specify which organization to use for each API request. See Authentication to learn more.

Figure 2. OpenAI API key creation.

4 Results and Analysis

4.1 Big data analysis models evaluation

4.1.1Heat Map with Correlations results

The relationship between the various variables of car categories is displayed in the scattered matrix plot for the variables "number of models," "rating," "quantity sold," and "total value sold." The scatter matrix shown in Figure 3, which has the values of each variable plotted on the x and y axes, indicates that there is no correlation between the variables and that all the data clusters provide non-linear values.

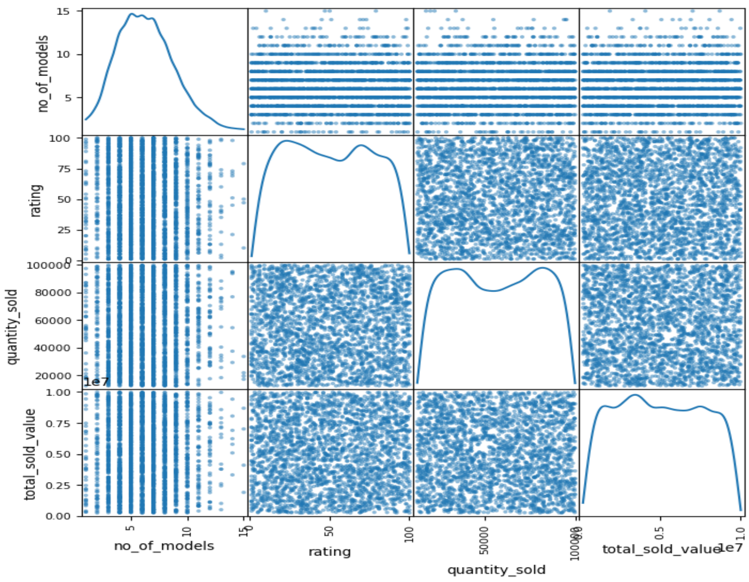


Figure 3. Scatter Matrix for variables

Two correlation models are used to assess a stronger and more accurate correlation between the variables. The correlation heatmaps exhibit a value bar ranging from 1 to 0, signifying the degree of correlation. A dark blue hue denotes a high correlation, while a light-yellow colour indicates no correlation. The Pearson method for vectorized data resulted in matrix values for all the variables shown in Figure 4. The values -0.02 to 0.04 indicate negative correlation and no correlation among each of the parameters.

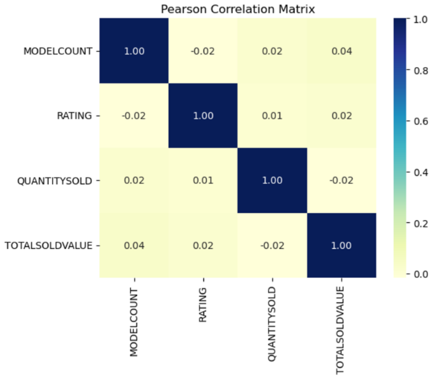


Figure 4. Pearson Correlation Matrix Heat Map

The Spearman correlation matrix, as seen in Figure 4.3, displays the same outcome as the Pearson coefficient for vectorized data of the variables, showing no correlation and a negative correlation between the r values of -0.02 to 0.03.

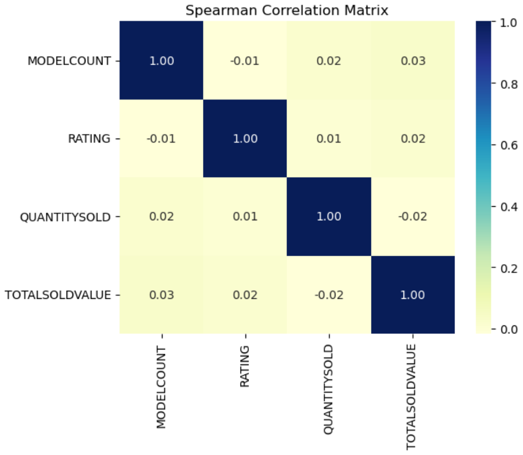


Figure 5. Spearman Correlation Matrix Heat Map

When car category parameters from the dataset are examined using various models for correlation, the results obtained from the applied models' scatter Matrix, Spearman, and Pearson coefficient matrix plots unequivocally demonstrate that no correlation exists.

4.2.1 Time series Analysis evaluation

The time series decomposition of the addictive method outcomes is shown in Figure 6 to assess the time series model of analysing trends and patterns in the data. The observed, trend, seasonality, and residual components of decomposition are provided by this result. Sales growth and decline demonstrate that the trend is erratic. When compared to observed sales, the seasonality component appears more reasonable. With certain data points, the residual component which explains the observation that remains after trend and seasonality seems intriguing.

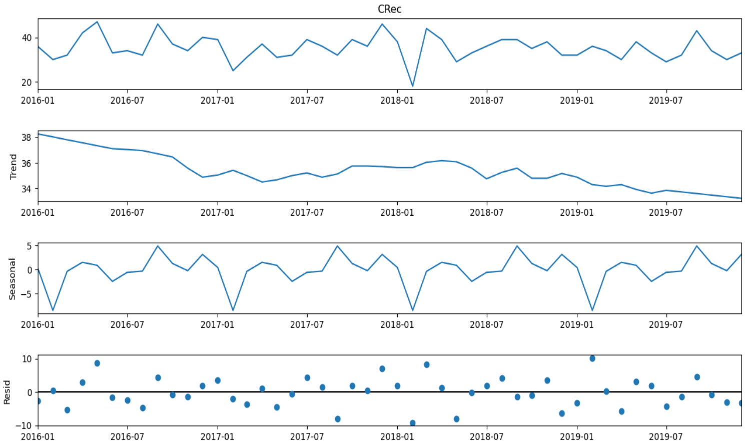


Figure 6. Time series decomposition for monthly sales

The Rolling static stationarity testing plotted as shown in Figure 7 represents the rolling and rolling standard in a linear manner and summing the data is stationary. But to prove with calculations and high confidence the Augmented Dicker Fuller stationarity interpreting method shown in Figure 4.6 is plotted. The evaluation proves to be stationary as the test static value (-7.55) is more negative than the critical values (-3.58, -2.92, -2.60), also the P-value (0) indicates the strong positive evidence of the observed stationarity.

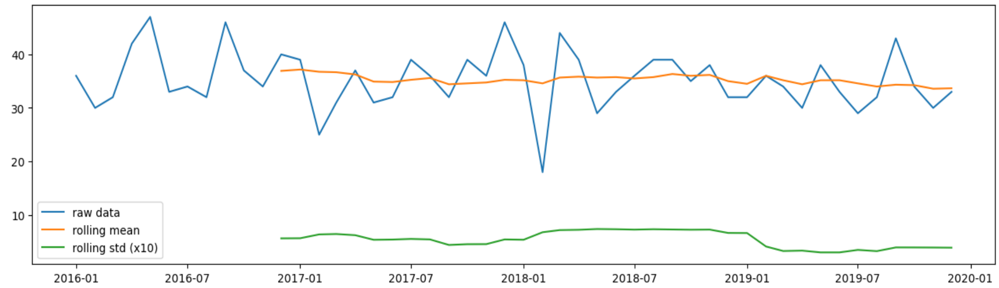
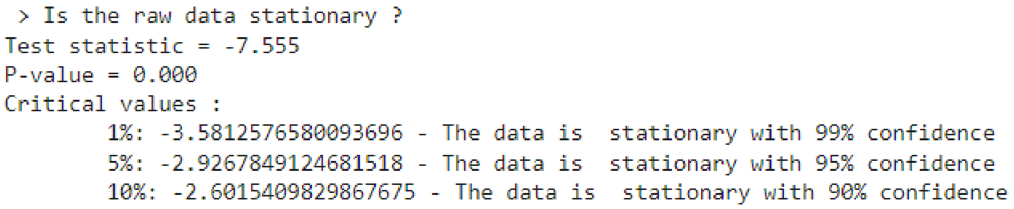


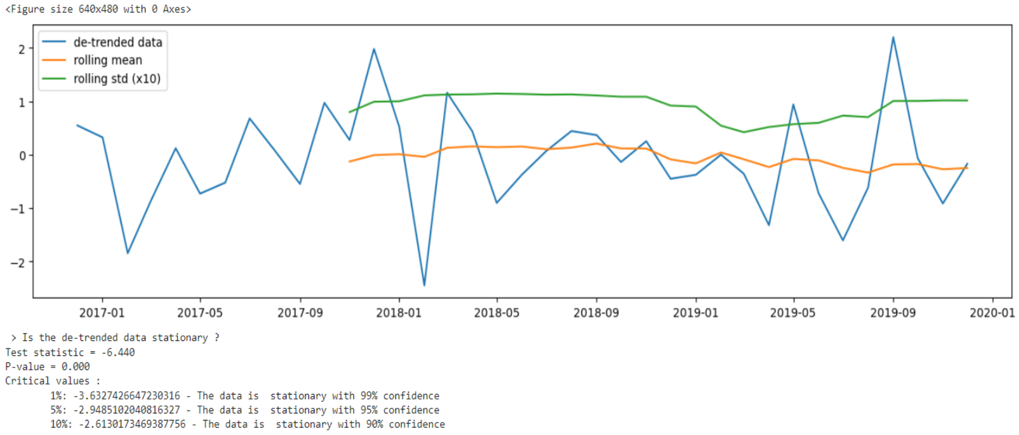
Figure 7. Rolling Statics of raw data



```
 > Is the raw data stationary ?
Test statistic = -7.555
P-value = 0.000
Critical values :
        1%: -3.5812576580093696 - The data is  stationary with 99% confidence
        5%: -2.9267849124681518 - The data is  stationary with 95% confidence
        10%: -2.6015409829867675 - The data is  stationary with 90% confidence
```

Figure 8. Outcomes from ADF stationarity testing

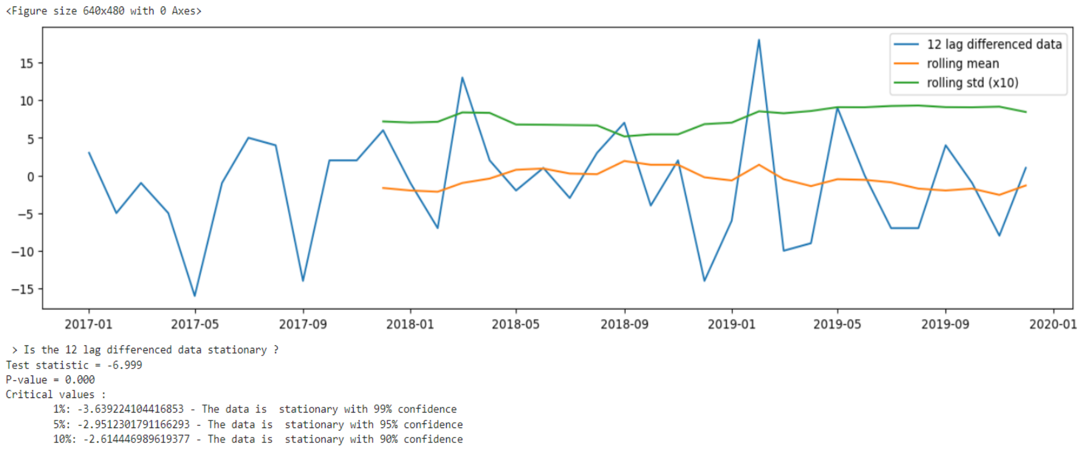
The Differencing, Detrending and the Differencing + Detrending are shown in Figure 9, 10, 11 respectively which clearly shows the data is mostly stationary with high confidence as P-value is ‘0’ and with an average of 95% confidence proved with critical values.



```
<Figure size 640x480 with 0 Axes>
```

```
 > Is the de-trended data stationary ?
Test statistic = -6.440
P-value = 0.000
Critical values :
        1%: -3.6327426647230316 - The data is  stationary with 99% confidence
        5%: -2.9485102040816327 - The data is  stationary with 95% confidence
        10%: -2.6130173469387756 - The data is  stationary with 90% confidence
```

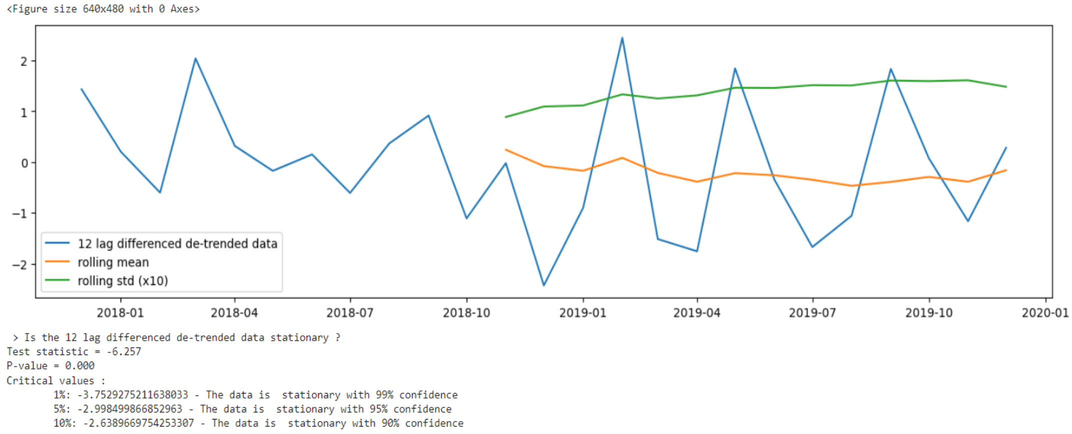
Figure 9. Outcomes of detrended data



```
<Figure size 640x480 with 0 Axes>
```

```
 > Is the 12 lag differenced data stationary ?
Test statistic = -6.999
P-value = 0.000
Critical values :
	1%: -3.639224104416853 - The data is  stationary with 99% confidence
	5%: -2.9512301791166293 - The data is  stationary with 95% confidence
	10%: -2.614446989619377 - The data is  stationary with 90% confidence
```

Figure 10. Outcome of Differencing data with 12 lag difference



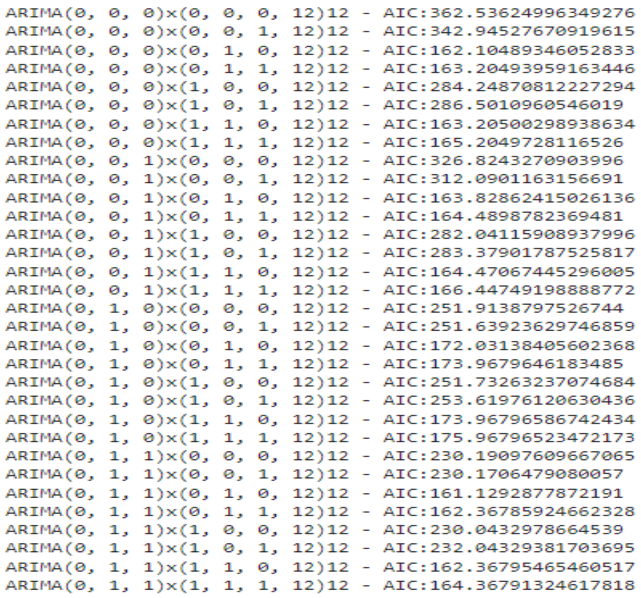
```
<Figure size 640x480 with 0 Axes>
```

```
 > Is the 12 lag differenced de-trended data stationary ?
Test statistic = -6.257
P-value = 0.000
Critical values :
	1%: -3.7529275211638033 - The data is  stationary with 99% confidence
	5%: -2.998499866852963 - The data is  stationary with 95% confidence
	10%: -2.6389669754253307 - The data is  stationary with 90% confidence
```

Figure 11. Outcome of Differencing and Detrending combined

4.2.2 ARIMA Machine Learning Model Evaluation

The grid search was obtained by the SARIMA Machine Learning Model, which was designed with a 12-season period, showing some of the grid parameter values in figure 12.



```
ARIMA(0, 0, 0)x(0, 0, 0, 12)12 - AIC:362.53624996349276
ARIMA(0, 0, 0)x(0, 0, 1, 12)12 - AIC:342.94527670919615
ARIMA(0, 0, 0)x(0, 1, 0, 12)12 - AIC:162.10489346052833
ARIMA(0, 0, 0)x(0, 1, 1, 12)12 - AIC:163.20493959163446
ARIMA(0, 0, 0)x(1, 0, 0, 12)12 - AIC:284.24870812227294
ARIMA(0, 0, 0)x(1, 0, 1, 12)12 - AIC:286.5010960546019
ARIMA(0, 0, 0)x(1, 1, 0, 12)12 - AIC:163.20500298938634
ARIMA(0, 0, 0)x(1, 1, 1, 12)12 - AIC:165.2049728116526
ARIMA(0, 0, 1)x(0, 0, 0, 12)12 - AIC:326.8243270903996
ARIMA(0, 0, 1)x(0, 0, 1, 12)12 - AIC:312.0901163156691
ARIMA(0, 0, 1)x(0, 1, 0, 12)12 - AIC:163.82862415026136
ARIMA(0, 0, 1)x(0, 1, 1, 12)12 - AIC:164.4898782369481
ARIMA(0, 0, 1)x(1, 0, 0, 12)12 - AIC:282.04115908937996
ARIMA(0, 0, 1)x(1, 0, 1, 12)12 - AIC:283.37901787525817
ARIMA(0, 0, 1)x(1, 1, 0, 12)12 - AIC:164.47067445296005
ARIMA(0, 0, 1)x(1, 1, 1, 12)12 - AIC:166.44749198888772
ARIMA(0, 1, 0)x(0, 0, 0, 12)12 - AIC:251.9138797526744
ARIMA(0, 1, 0)x(0, 0, 1, 12)12 - AIC:251.63923629746859
ARIMA(0, 1, 0)x(0, 1, 0, 12)12 - AIC:172.03138405602368
ARIMA(0, 1, 0)x(0, 1, 1, 12)12 - AIC:173.9679646183485
ARIMA(0, 1, 0)x(1, 0, 0, 12)12 - AIC:251.73263237074684
ARIMA(0, 1, 0)x(1, 0, 1, 12)12 - AIC:253.61976120630436
ARIMA(0, 1, 0)x(1, 1, 0, 12)12 - AIC:173.96796586742434
ARIMA(0, 1, 0)x(1, 1, 1, 12)12 - AIC:175.96796523472173
ARIMA(0, 1, 1)x(0, 0, 0, 12)12 - AIC:230.19097609667065
ARIMA(0, 1, 1)x(0, 0, 1, 12)12 - AIC:230.1706479080057
ARIMA(0, 1, 1)x(0, 1, 0, 12)12 - AIC:161.1292877872191
ARIMA(0, 1, 1)x(0, 1, 1, 12)12 - AIC:162.36785924662328
ARIMA(0, 1, 1)x(1, 0, 0, 12)12 - AIC:230.0432978664539
ARIMA(0, 1, 1)x(1, 0, 1, 12)12 - AIC:232.04329381703695
ARIMA(0, 1, 1)x(1, 1, 0, 12)12 - AIC:162.36795465460517
ARIMA(0, 1, 1)x(1, 1, 1, 12)12 - AIC:164.36791324617818
```

Figure 12. SARIMA Grid search parameter combinations

As seen in Figure 13, the trained SRIMA model with the train set of data yields four components: the standard residual, the histogram plus estimation, the normal Q-Q, and the correlogram. It is indicated by the first plot standardized residual for "C," which represents the residual errors, that some data may not be well captured by the model. Second, the model's premise is refuted by the histogram plot, which verifies that the curve's estimated density and standard distribution differ. Because the data are aligned on the line in the Normal Q-Q plot, the model's proper application with the inputs is also demonstrated. This leads to a normal distribution. The outcomes are statistically significant over the 12-month lag applied by the model, as indicated by the correlogram plot.

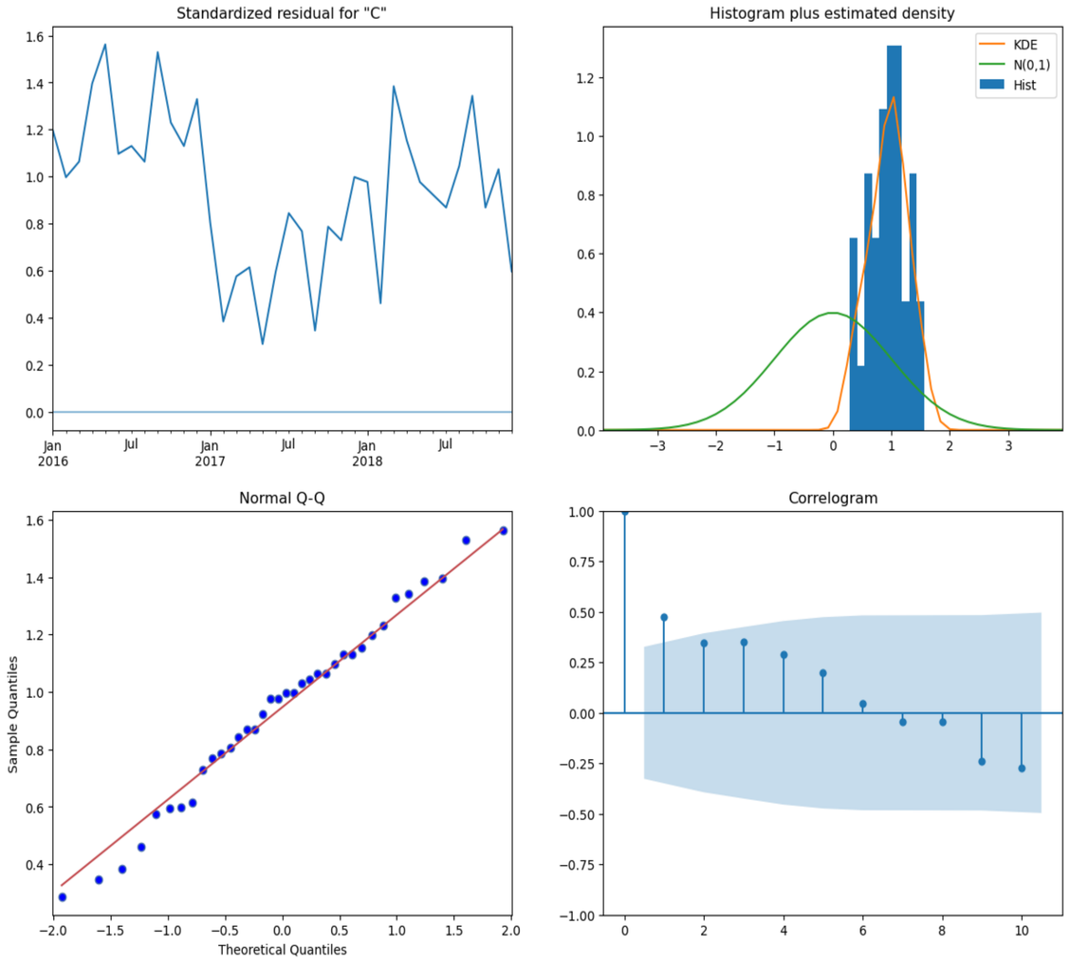


Figure 13. Results of trained SARIMA Model with Grid search parameters

The results of using the trained model to predict car sales over the next eight months are largely significant, as seen in Figure 14 below.

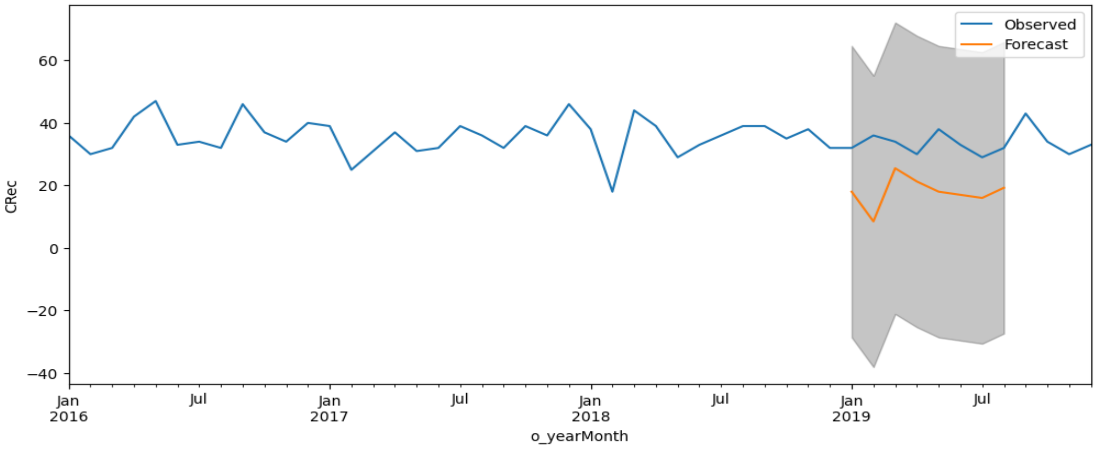
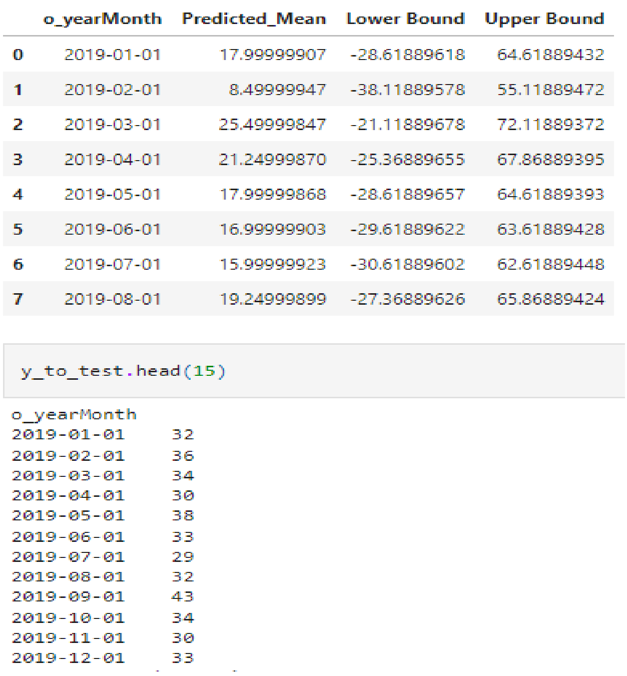


Figure 14. SARIMA model prediction

The test data are compared with the predicted mean along with the values of the lower and upper bounds. As seen in Figure 15, the SARIMA model prediction fits within the bound range but primarily adheres to the slightly higher value of the lower bound.



	o_yearMonth	Predicted_Mean	Lower Bound	Upper Bound
0	2019-01-01	17.99999907	-28.61889618	64.61889432
1	2019-02-01	8.49999947	-38.11889578	55.11889472
2	2019-03-01	25.49999847	-21.11889678	72.11889372
3	2019-04-01	21.24999870	-25.36889655	67.86889395
4	2019-05-01	17.99999868	-28.61889657	64.61889393
5	2019-06-01	16.99999903	-29.61889622	63.61889428
6	2019-07-01	15.99999923	-30.61889602	62.61889448
7	2019-08-01	19.24999899	-27.36889626	65.86889424

```
y_to_test.head(15)
```

```
o_yearMonth
2019-01-01    32
2019-02-01    36
2019-03-01    34
2019-04-01    30
2019-05-01    38
2019-06-01    33
2019-07-01    29
2019-08-01    32
2019-09-01    43
2019-10-01    34
2019-11-01    30
2019-12-01    33
```

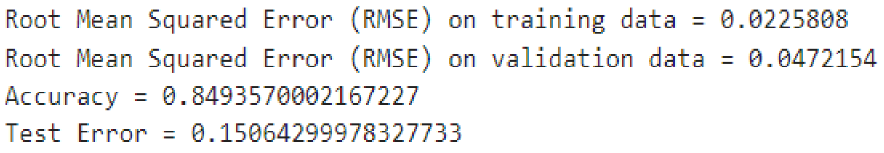
Figure 15. Prediction details with test data

4.2 The outcomes from Machine Learning Models

A variety of metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Accuracy, and Test Error, are used to evaluate machine learning models. Regression mean square error (RMSE) is an immensely popular measure of evaluation for regression-related issues because it not only shows the impact of large errors but also computes the standard deviation between the prediction and the actual value. A lower value signifies a more accurate prediction, whereas a value of 0 means the model fits precisely. The model's accuracy increases with decreasing MAE, which is the mean of the absolute errors between the dataset's predicted and actual values. For the machine learning models in this project, test error is assessed in addition to RMSE and MAE, and the model's accuracy is also computed independently.

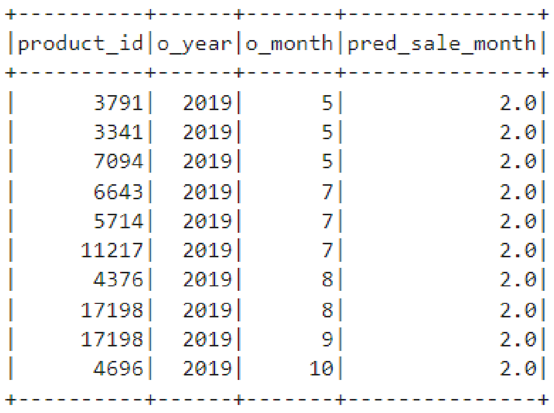
4.2.1 Random Forest Model

When the model is trained on the train data, which consists of a set of 1017 records that make up 64% of the total data, the random Forest machine learning model makes predictions on the test set of data, which consists of 258 records and represents 16% of the total data. As illustrated in figure 4.13 below, the transformed data of the train and test sets are evaluated using the prediction parameter. The results appear to be realistic and satisfactory as per standards, with an 84% accuracy, 0.0225808 RMSE value, and test error of 0.15. Hence, the prediction of top 10 products with predicted sale column shown in figure 16 is usable.



```
Root Mean Squared Error (RMSE) on training data = 0.0225808
Root Mean Squared Error (RMSE) on validation data = 0.0472154
Accuracy = 0.8493570002167227
Test Error = 0.15064299978327733
```

Figure 4.13 Evaluation of Random Forest Model

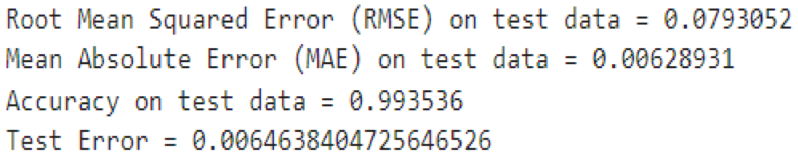


```
+----------+------+-------+--------------+
|product_id|o_year|o_month|pred_sale_month|
+----------+------+-------+--------------+
|      3791|  2019|      5|           2.0|
|      3341|  2019|      5|           2.0|
|      7094|  2019|      5|           2.0|
|      6643|  2019|      7|           2.0|
|      5714|  2019|      7|           2.0|
|     11217|  2019|      7|           2.0|
|      4376|  2019|      8|           2.0|
|     17198|  2019|      8|           2.0|
|     17198|  2019|      9|           2.0|
|      4696|  2019|     10|           2.0|
+----------+------+-------+--------------+
```

Figure 16. Predicted top 10 products by Random Forest Model.

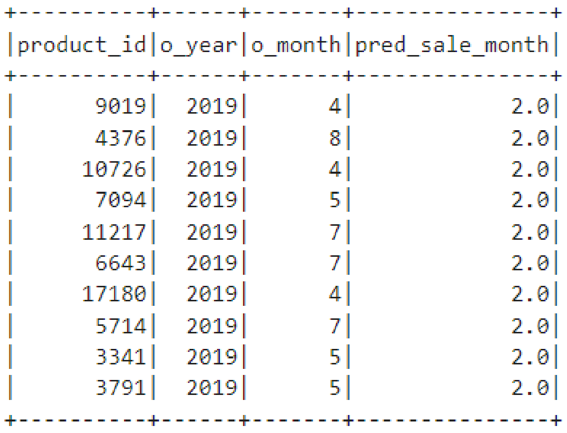
4.2.2 Gradient Boosted Decision Tree Model

20% of the test data comprising of 318 records among overall data, are used to assess the performance of the gradient-boosted decision tree model, which is trained using 80% of the total data having 1283 records. The evaluation metrics as a whole demonstrate that this model fits data more precisely, with RMSE of 0.07, MAE of 0.006, test error of 0.01 demonstrating the lowest values possible, and highest accuracy of 99%, as illustrated in figure 17. As a result, the model's prediction is extremely satisfactory. In comparison to the random forest model, this model turned out to be most effective. Figure 4.16 below displays the top 10 products based on predicted sales from this GBDT model.



```
Root Mean Squared Error (RMSE) on test data = 0.0793052
Mean Absolute Error (MAE) on test data = 0.00628931
Accuracy on test data = 0.993536
Test Error = 0.0064638404725646526
```

Figure 17. Evaluation of GBDT (Gradient Boosted Decision Tree) Model

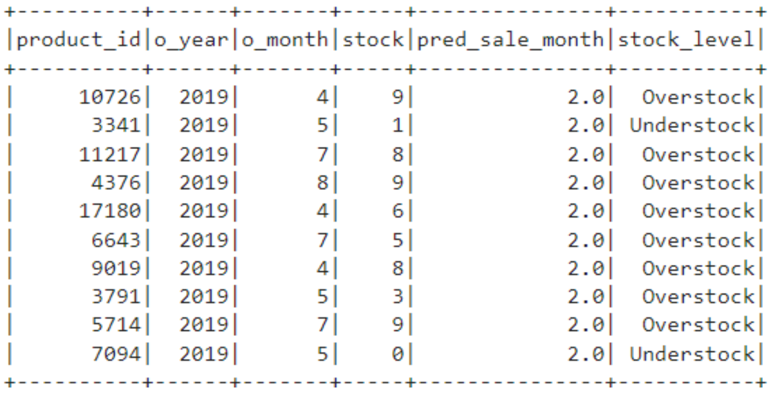


```
+----------+------+-------+---------------+
|product_id|o_year|o_month|pred_sale_month|
+----------+------+-------+---------------+
|      9019|  2019|      4|            2.0|
|      4376|  2019|      8|            2.0|
|     10726|  2019|      4|            2.0|
|      7094|  2019|      5|            2.0|
|     11217|  2019|      7|            2.0|
|      6643|  2019|      7|            2.0|
|     17180|  2019|      4|            2.0|
|      5714|  2019|      7|            2.0|
|      3341|  2019|      5|            2.0|
|      3791|  2019|      5|            2.0|
+----------+------+-------+---------------+
```

Figure 18. Top 10 predicted products from GBDT Model

The impact of OpenAI integration and extracted insights

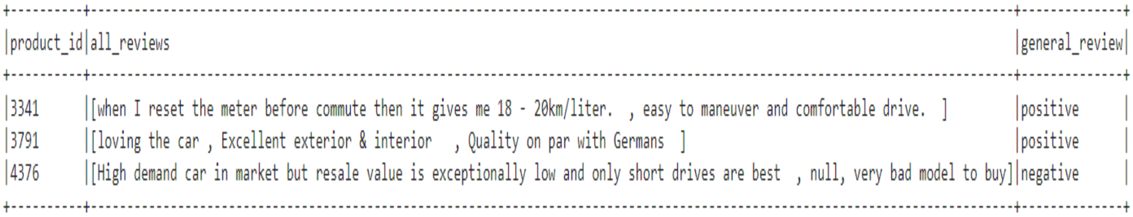
The classification of each product is obtained by considering the best model's output, which is typically the top 10 predicted products from the GBDT model. Figure 19 illustrates how well each product is classified.



```
+----------+------+-------+-----+---------------+-----------+
|product_id|o_year|o_month|stock|pred_sale_month|stock_level|
+----------+------+-------+-----+---------------+-----------+
|     10726|  2019|      4|    9|            2.0|  Overstock|
|      3341|  2019|      5|    1|            2.0| Understock|
|     11217|  2019|      7|    8|            2.0|  Overstock|
|      4376|  2019|      8|    9|            2.0|  Overstock|
|     17180|  2019|      4|    6|            2.0|  Overstock|
|      6643|  2019|      7|    5|            2.0|  Overstock|
|      9019|  2019|      4|    8|            2.0|  Overstock|
|      3791|  2019|      5|    3|            2.0|  Overstock|
|      5714|  2019|      7|    9|            2.0|  Overstock|
|      7094|  2019|      5|    0|            2.0| Understock|
+----------+------+-------+-----+---------------+-----------+
```

Figure 19. Stock level Classification of Products

The products from the above output are limited to three due to the limitation of OpenAI. By passing all the reviews of each product separated by "," in the request, results of three products extracted from ChatGPT, as shown in figure 20. As requested, ChatGPT provided the exact sentiment in a single word. While most of the reviews for products 3341 and 3791 are good, the response is positive. Whereas results for product 4376, which has null, negative, and neutral reviews, response is negative. This indicates that ChatGPT has provided more precise responses for this application and are realistic regardless of the number of reviews submitted as a request.

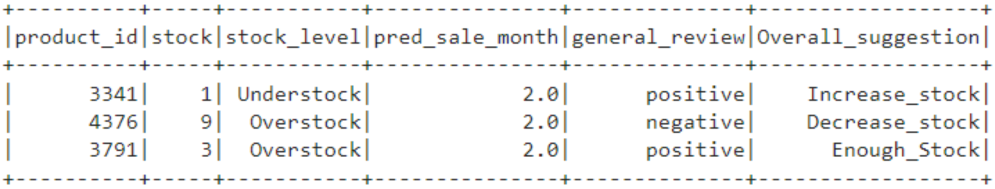


```
+----------+-------------------------------------------------------------------------------------------------------------------------------+--------------+
|product_id|all_reviews                                                                                                                    |general_review|
+----------+-------------------------------------------------------------------------------------------------------------------------------+--------------+
|3341      |[when I reset the meter before commute then it gives me 18 - 20km/liter.  , easy to maneuver and comfortable drive.  ]         |positive      |
|3791      |[loving the car , Excellent exterior & interior   , Quality on par with Germans  ]                                             |positive      |
|4376      |[High demand car in market but resale value is exceptionally low and only short drives are best  , null, very bad model to buy]|negative      |
+----------+-------------------------------------------------------------------------------------------------------------------------------+--------------+
```

Figure 20. Sentiment results from ChatGPT

A comparative assessment using conventional approaches

Figure 21 illustrates the application's final output, which offers ideas for decision-making based on stock level, sales forecast, and feedback from customers.



```
+----------+-----+-----------+---------------+--------------+------------------+
|product_id|stock|stock_level|pred_sale_month|general_review|Overall_suggestion|
+----------+-----+-----------+---------------+--------------+------------------+
|      3341|    1| Understock|            2.0|      positive|    Increase_stock|
|      4376|    9|  Overstock|            2.0|      negative|    Decrease_stock|
|      3791|    3|  Overstock|            2.0|      positive|      Enough_Stock|
+----------+-----+-----------+---------------+--------------+------------------+
```

Figure 21. Final application solution

The conventional sentiment analysis methods have a large process flow of analysing the customer reviews using various algorithms, but this application works with simple steps and is more precise. This innovative application implemented aids in the optimization of inventory in terms of customer sentiments and decision making, making it easier for executives in charge of inventory to adapt to market changes.

5. Conclusions

This study demonstrates the feasibility and potential of integrating ChatGPT, Random Forest, and Gradient Boosted Decision Tree (GBDT) ensemble models on the Microsoft Azure cloud platform for optimizing inventory planning. Both models achieved high accuracy, with GBDT outperforming at 99% accuracy. Regardless of the volume of reviews processed, ChatGPT provided insightful responses that supported decision-making. As an advanced NLP model from OpenAI, ChatGPT leverages an extensive repository of phrases and word patterns to generate coherent and contextually appropriate text across various domains. This makes it particularly useful for conversational interfaces, enhancing its ability to interpret customer feedback and accurately respond in natural language. Deploying ChatGPT on the Azure cloud platform offered a scalable and flexible solution, addressing real-world inventory management challenges faced by business executives. This research lays the foundation for further advancements in intelligent inventory planning, including broader utilization of cloud computing resources, advanced sales forecasting techniques, and deep learning-based machine learning models.

References

Ahmed, A. A. A., Agarwal, S., Kurniawan, I. G. A., Anantadjaya, S. P., and Krishnan, C. 2022. "Business Boosting through Sentiment Analysis Using Artificial Intelligence Approach," *International Journal of System Assurance Engineering and Management* (13:Suppl 1), pp. 699-709.

Alon, I., Qi, M., and Sadowski, R. J. 2001. "Forecasting Aggregate Retail Sales:: A Comparison of Artificial Neural Networks and Traditional Methods," *Journal of retailing and consumer services* (8:3), pp. 147-156.

Ariyo, A. A., Adewumi, A. O., and Ayo, C. K. 2014. "Stock Price Prediction Using the Arima Model," *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pp. 106-112.

Borodavko, B., Illés, B., and Bányai, Á. 2021. "Role of Artificial Intelligence in Supply Chain," *Academic Journal of Manufacturing Engineering* (19:1), pp. 75-79.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. 2020. "Language Models Are Few-Shot Learners," *Advances in neural information processing systems* (33), pp. 1877-1901.

Cheriyan, S., Ibrahim, S., Mohanan, S., and Treesa, S. 2018. "Intelligent Sales Prediction Using Machine Learning Techniques," *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*: IEEE, pp. 53-58.

Cortez, C. T., Saydam, S., Coulton, J., and Sammut, C. 2018. "Alternative Techniques for Forecasting Mineral Commodity Prices," *International Journal of Mining Science and Technology* (28:2), pp. 309-322.

Decker, R., and Gnibba-Yukawa, K. 2010. "Sales Forecasting in High-Technology Markets: A Utility-Based Approach," *Journal of product innovation management* (27:1), pp. 115-129.

Desai, V. P., and Oza, K. S. 2021. "Fine Tuning Modeling through Open Ai," *Progression in Science, Technology and Smart Computing, PRARUP*).

Hendriksen, C. 2023. "Artificial Intelligence for Supply Chain Management: Disruptive Innovation or Innovative Disruption?," *Journal of Supply Chain Management* (59:3), pp. 65-76.

Hornik, K., Stinchcombe, M., and White, H. 1989. "Multilayer Feedforward Networks Are Universal Approximators," *Neural networks* (2:5), pp. 359-366.

Korkmaz, A., Aktürk, C., and Talan, T. 2023. "Analyzing the User's Sentiments of Chatgpt Using Twitter Data," *Iraqi Journal for Computer Science and Mathematics* (4:2), pp. 202-214.

Koumanakos, D. P. 2008. "The Effect of Inventory Management on Firm Performance," *International journal of productivity and performance management* (57:5), pp. 355-369.

Kourentzes, N., Trapero, J. R., and Barrow, D. K. 2020. "Optimising Forecasting Models for Inventory Planning," *International Journal of Production Economics* (225), p. 107597.

Liu, Y. 2006. "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of marketing* (70:3), pp. 74-89.

McCarthy, T. M., Davis, D. F., Golicic, S. L., and Mentzer, J. T. 2006. "The Evolution of Sales Forecasting Management: A 20-Year Longitudinal Study of Forecasting Practices," *Journal of Forecasting* (25:5), pp. 303-324.

Mentzer, J. T., and Bienstock, C. C. 1998. "Sales Forecasting Management," *(No Title)*).

Muller, M. 2019. *Essentials of Inventory Management*. HarperCollins Leadership.

Prabowo, R., and Thelwall, M. 2009. "Sentiment Analysis: A Combined Approach," *Journal of Informetrics* (3:2), pp. 143-157.

Ryu, K., and Han, H. 2010. "Influence of the Quality of Food, Service, and Physical Environment on Customer Satisfaction and Behavioral Intention in Quick-Casual Restaurants: Moderating Role of Perceived Price," *Journal of Hospitality & Tourism Research* (34:3), pp. 310-329.

Salloum, S., Dautov, R., Chen, X., Peng, P. X., and Huang, J. Z. 2016. "Big Data Analytics on Apache Spark," *International Journal of Data Science and Analytics* (1:3), pp. 145-164.

Shetty, S. K., and Buktar, R. 2022. "A Comparative Study of Automobile Sales Forecasting with Arima, Sarima and Deep Learning Lstm Model," *International Journal of Advanced Operations Management* (14:4), pp. 366-387.

Tsou, C.-S. 2008. "Multi-Objective Inventory Planning Using Mopso and Topsis," *Expert systems with applications* (35:1-2), pp. 136-142.

Ugbebor, F., Adeteye, M., and Ugbebor, J. 2024. "Automated Inventory Management Systems with Iot Integration to Optimize Stock Levels and Reduce Carrying Costs for Smes: A Comprehensive Review," *Journal of Artificial Intelligence General Science (JAIGS) ISSN: 3006-4023* (6:1), pp. 306-340.

Voulgaridou, D., Kirytopoulos, K., and Leopoulos, V. 2009. "An Analytic Network Process Approach for Sales Forecasting," *Operational Research* (9), pp. 35-53.

Wood, L., Reiners, T., and Srivistava, H. 2013. "Expanding Sales and Operations Planning Using Sentiment Analysis: Demand and Sales Clarity from Social Media," *ANZAM Conference Proceedings*: N/A, pp. 1-17.

Speaking my mind, but to who? The Ethics of AI Chatbots Supporting Mental Health

Laurence Brooks[1] **and Jayant Shinde**[2]
Information School, University of Sheffield, UK
[1]l.brooks@sheffield.ac.uk; [2]jayants99@gmail.com

Completed Research

Abstract

AI chatbots offer a potential solution to increase accessibility and efficiency in mental health care. However, their implementation raises significant ethical concerns, including data privacy, lack of human empathy, and potential misuse. This study uses interviews to investigate end-user experiences with AI chatbots in a mental health context. Overall, while users appreciate their convenience and accessibility, they primarily view them as supplementary tools for early interventions, not replacements for human therapists. To maximize the benefits of AI chatbots while mitigating risks, it is crucial to prioritize strong data security, maintain human empathy in care, and clearly define their role as supplementary tools. By addressing these ethical considerations, AI chatbots could contribute positively to mental health care.

Keywords: AI, chatbot, mental health

1.0 Introduction

Poor mental health is an important global public health issue, and people from many demographic and geographic backgrounds are affected. Mental health issues are predicted to affect 29% of people, with anxiety, depression, and post-traumatic stress disorder increasing. Despite increased awareness and efforts to reduce stigma, access to mental health services remains restricted, highlighting the need for accessible and effective care (Steel *et al.*, 2014). Other barriers include financial restrictions and a lack of providers offering specialist mental health services, especially in rural and underserved areas (Torous *et al.*, 2021). The consequences are that conventional mental health care systems can hardly match the emerging demand and lead to long waiting lists, which further exacerbates the situation (Kessler *et al.*, 2005).

Advances in technology within the past couple of decades have urged investigation into new methods of enhancing mental health support. Artificial Intelligence (AI) powered chatbots are being integrated into healthcare, providing information, support, and therapeutic interventions through natural language processing (NLP) algorithms, a method that has gained popularity in recent years. These chatbots can keep all conversations private, providing a non-judgmental space for individuals to express themselves, hence facilitating immediate and confidential mental health support (Torous *et al.*, 2021).

AI-powered chatbots, while having much potential for benefits in support for mental health, generally pose a number of ethical issues in their use. AI technologies raise concerns about data privacy and security, potentially leading to breaches and misuse, and potential biases in algorithms, potentially causing unequal care for marginalized populations (Tschandl *et al.*, 2020). One possibility for bias comes from the training models for AI being drawn from unrepresentative datasets, hence giving skewed or simply wrong answers regarding the inquiries posed to these machines (Floridi *et al.*, 2018).

Another ethical issue is the reliability and accuracy of AI chatbots recognizing and responding to complex emotional states. While AI might support people at a basic level, it does not replace the emotional empathy or comprehension that forms the backbone of mental health care (Castellano *et al.*, 2013). The limitation with AI creates serious issues related to the adequacy of such interactions and, in cases of over-reliance, the potential harm when an individual depends on chatbots for mental health support (Torous *et al.*, 2021). These ethical considerations call for extensive guidelines on ethics and proper security of data in order not to breach users' privacy and well-being (Morley *et al.*, 2020). This leads the paper to the following set of research questions 1. What are the main ethical issues associated with the use of AI-powered chatbots for mental health support? 2. In what ways do users perceive and experience AI-powered chatbots in the context of mental health support, and how do these relate to the ethical issues identified? 3. What strategies can be implemented to address ethical concerns and maximize the benefits of AI-powered chatbots in mental health support? In order to answer these questions, the paper first reviews and integrates relevant literature around the ideas of mental health chatbots benefits and key ethical challenges/concerns. This is followed by a brief description of the methodology used to explore individual perceptions of mental health chatbots by those that have and have not used them. The final sections discuss the findings and conclude with…

2.0 Literature Review

This review seeks to discuss the key ethical implications of using AI chatbots in providing mental healthcare in the key areas of privacy, efficacy, and therapist–patient relationships. This review will thus seek to look at existing literature and empirical data to give insights on ethical challenges and opportunities presented (see Figure 1) with the intention of developing recommendations on responsible designing, implementation, and regulation with a view to assuring user well-being.

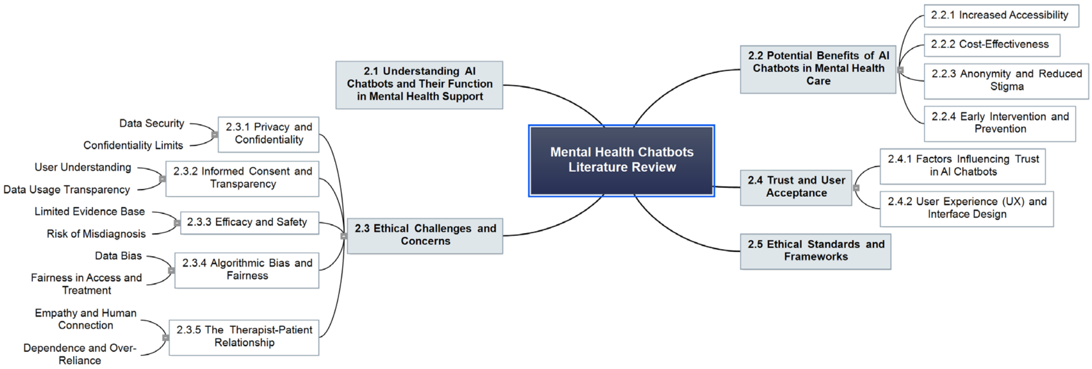

Mental Health Chatbots Literature Review
2.1 Understanding AI Chatbots and Their Function in Mental Health Support
2.2 Potential Benefits of AI Chatbots in Mental Health Care
2.2.1 Increased Accessibility
2.2.2 Cost-Effectiveness
2.2.3 Anonymity and Reduced Stigma
2.2.4 Early Intervention and Prevention
2.3 Ethical Challenges and Concerns
2.3.1 Privacy and Confidentiality
Data Security
Confidentiality Limits
2.3.2 Informed Consent and Transparency
User Understanding
Data Usage Transparency
2.3.3 Efficacy and Safety
Limited Evidence Base
Risk of Misdiagnosis
2.3.4 Algorithmic Bias and Fairness
Data Bias
Fairness in Access and Treatment
2.3.5 The Therapist-Patient Relationship
Empathy and Human Connection
Dependence and Over-Reliance
2.4 Trust and User Acceptance
2.4.1 Factors Influencing Trust in AI Chatbots
2.4.2 User Experience (UX) and Interface Design
2.5 Ethical Standards and Frameworks

Figure 1 Overview of issues relating to mental health chatbots

Mental health problems have grown to be more prevalent in the entire world, constituting major outstanding public health challenges. Steel et al., (2014) estimated that mental health disorders are likely to affect up to 29% of people at some point in their lifetime, also showing the high prevalence of the conditions. Although there is an increasing awareness and attempts to reduce stigma on mental illness, poor access to adequate mental health care remains a big problem. Different factors, including social stigma, financial constraints, and gaps in service provision, especially in rural areas, make access by individuals to necessary support challenging (Smalley, Warren and Rainer, 2012). Traditional mental health services, most of the time, do not keep pace with the demand and many times have long waiting lists and hence cannot provide the necessary support to the people seeking it (Patel *et al.*, 2018).

2.1 Understanding AI Chatbots and Their Function in Mental Health Support

AI chatbots are an innovative development within the mental health landscape, also designed to deliver timely support, information, and basic therapeutic intervention through Natural Language Programming (NLP) algorithms (Fitzpatrick, Darcy and Vierhile, 2017). The interaction of AI-driven tools with users in real time creates a non-judgmental and confidential environment where individuals feel uninhibited about their self-expression. By making mental health support services available 24 hours a day, 7 days a week, AI chatbots have the potential to help address shortages in mental health services and increase the availability of support for underserved populations (Vaidyam *et al.*, 2019).

2.2 Potential Benefits of AI Chatbots in Mental Health Care

2.2.1 Increased Accessibility

Accessibility to mental health care can be enhanced with the inclusion of AI chatbots, which are made available 24/7. Traditional mental health services are usually constrained by office hours and appointment availability. In times of crisis, there is often no avenue of immediate help available through more traditional services. Chatbots operate 24/7 and provide timely support and interventions at the right moments (Fulmer *et al.*, 2018). It is especially useful in the case of an acute or immediate mental health episode outside the normal office hours, as help can be provided immediately without having to wait.

Moreover, AI chatbots can also reach out to the underserved populations who otherwise might face barriers in seeking mental health services. These AI chatbots can also be with populations who have systemic barriers to approaching traditional mental health services, eg. individuals with physical disabilities or those in low-income brackets who cannot afford regular sessions of therapy.

2.2.2 Cost-Effectiveness

AI chatbots in mental health care also can minimise the budgetary impact of conventional therapy. Traditional services for mental health, like individual therapy sessions, can be very expensive and usually beyond the reach of people with low income/less health insurance. In turn, these chatbots are relatively more affordable and can provide basic support and interventions at much lower costs as opposed to traditional methods (Inkster, Sarda and Subramanian, 2018).

Scalability allows for the deployment of chatbots in the broadest manner, which can be done at relatively low incremental costs. Once a chatbot has been developed, it can easily interact with many users simultaneously without requiring extra resources, unlike human therapists, who obviously have only limited capacity often seeing patients one-on-one (Fitzpatrick, Darcy and Vierhile, 2017). This makes the use of chatbots economically viable to expand mental health support services, especially in resource-poor settings where financial barriers are considered major obstacles in seeking care.

2.2.3 Anonymity and Reduced Stigma

Anonymity interacting with AI chatbots can make a difference for those who might be afraid to seek conventional help for fear of stigma and privacy compromise. Stigma pertaining to mental health continues to remain one of the major challenges in seeking care; most people have a fear of judgement or persecution if they share their struggle (Clement *et al.*, 2015). The chatbot provides a private, non-judgmental space where users can discuss issues without the fear of stigma.

It is also likely that anonymity may be particularly attractive to individuals who have never sought mental health services and will serve as a comfortable entry point into services. Privacy with the chatbot interaction will let the user feel more secure in their self-expression, which might allow users to feel more open and honest within disclosures (Bendig *et al.*, 2022). This may better the effectiveness of support provided through increased user engagement in interventions of the chatbot.

2.2.4 Early Intervention and Prevention

AI chatbots hold great potential for early intervention and prevention in the very early stages of developing mental health problems. In concert with continuous engagement, chatbots monitor responses to help in unearthing early signs of mental health problems for timely interventions (Fitzpatrick, Darcy and Vierhile, 2017). Early detection is the essence of successful mental health care, which indicates an end to the progression of the respective condition and a lower burden on afflicted individuals.

Chatbots can provide prevention and learning strategies, helping users develop coping mechanisms and resilience in response to potential severe mental health threats. For instance, a chatbot should be able to provide tips on effective stress management, mindfulness practices, and healthy lifestyle choices that help take care of one's mind (Vaidyam *et al.*, 2019). By offering knowledge about health and self-care tools, chatbots support proactive health management and therefore may reduce cases of more severe states of mental health.

In summary, AI chatbots bring potentials such as increased access to services, cost savings, anonymous therapies, and the possibility of early prevention and intervention. Such strengths identify chatbots as a new hope in the mental health service continuum and a means to address some of the key issues that are problematic in traditional services.

2.3 Ethical Challenges and Concerns

However, integrating AI chatbots into mental health treatment is not without significant ethical concerns and pitfalls. These include issues of data privacy, security, and algorithms, as sensitive personal information collected may be used to breach user confidentiality (Bendig *et al.*, 2022). Also, their reliability is questionable with the potential for misdiagnosis or their dispensing inappropriate treatments (Inkster, Sarda and Subramanian, 2018). In addition, the human touch or empathy and subtle comprehension may be beyond the capacity of a chatbot interaction in addressing complex emotional states (Luxton, 2014) which again brings under scrutiny the viability of such treatments.

2.3.1 Privacy and Confidentiality

Data Security

This inclusion of AI into mental healthcare brings up huge cause-for-concern issues on data security. Chatbots, by design, gather and process sensitive personal information, therefore becoming an attractive target for cyberattacks (Li, 2023). Data breaches create conditions whereby unauthorized access to confidential information among the patients probably leads to identity theft, blackmail, or other forms of exploitation (Looi *et al.*, 2024). Therefore it becomes very important that such sensitive information is secured with robust security measures like encryption, secure authentication protocols, and periodic security audits (Kommisetty and Dileep, 2024). Strong practices in data security will minimize the risks of breaches and ensure that user information remains confidential.

Confidentiality Limits

In developing chatbot performance, many developers, researchers, or even third- party vendors may have access to the data, which makes confidentiality hard to maintain (Rezaeikhonakdar, 2023). For example, developers may be granted access to enhance the performance of the chatbot, while anonymised data can be used by researchers. Such involvement opens potential vulnerabilities in how the data are processed and stored to compromise patient confidentiality (Surani and Das, 2022). Access control and data anonymisation are crucial for patient privacy and confidentiality, and clearly defined access and usage policies and agreements are essential to prevent confidentiality breaches.

2.3.2 Informed Consent and Transparency

User Understanding

Informed consent by users requires a complete understanding of the capability and limitation of an AI chatbot. They need to be aware that it is not with a human therapist but with an AI, which will result in major differences in expectations and trust (Rothstein, 2023). This leads to better preparation and decision-making on the part of users when the true nature of the chatbot is communicated as AI-natured interactions, no human empathy, and limitations in dealing with complex psychological issues (Miner, Milstein and Hancock, 2017). It is always important to consider transparency regarding the manner of operation of the chatbot and what it can plausibly provide as a way of maintaining ethical standards by informing users of the services about the nature of the interaction that they are entering (Codecademy Team, 2023).

Data Usage Transparency

Data usage transparency is critical for the trust of users and for ethical practices (Waseem *et al.*, 2024). The user should be clearly informed by the chatbot when their data is being collected, stored, and used, including data retention policies, the purpose for the collection, and security measures against breaches (Sebastian, 2023). Clear and transparent privacy policies and explicit consent before data collection make up the cardinal best practices that ensure users understand how their information will be used, (Hildt and Laas, 2022). This would minimize the chances of data misuse and ensure ethical dealings in digital mental health interventions.

2.3.3 Efficacy and Safety

Limited Evidence Base

The efficiency and safety of AI chatbots for complex mental health conditions remains under- researched. While these technologies continue to be increasingly adopted, there is a lack of longer-term, empirically based studies demonstrating their efficacy over an extended period (Abd-Alrazaq *et al.*, 2020). Assessments generally tend to focus on short-term results and may fail to consider wider ramifications of longer use. This gap in evidence raises concerns about the reliability of chatbots as standalone interventions for serious mental health conditions (Casu *et al.*, 2024).

Risk of Misdiagnosis

Chatbots using AI can be helpful for support during the initial stages, but there is a risk of mental health condition misdiagnosis due to the unsophistication of algorithms (H. Li *et al.*, 2023). Another drawback is that these systems can misinterpret the symptoms or generalize responses based on set algorithms which might bring out certain inappropriate recommendations or interventions (You *et al.*, 2023). The inability of chatbots to comprehend fully complex emotional states and multi-dimensional psychological states increases the risk of diagnosis going wrong and could lead to suggestions of unsatisfactory treatments that could be harmful to the patient (Miner, Milstein and Hancock, 2017). Ensuring that users understand these limitations is crucial to avoid over- reliance on chatbots for serious mental health issues.

2.3.4 Algorithmic Bias and Fairness

Data Bias

Any biases in training datasets can lead to discriminatory outcomes, further engraining existing health disparities. These biases usually emerge when some datasets are small and under representative or carry imbalances that show up as historical inequalities (Bernhardt, Jones and Glocker, 2022). For instance, if most of the information a chatbot receives comes from one ethnic group, it might have less relevant or less accurate recommendations to make for other groups (Hofmann *et al.*, 2024). Biases like such can widen the gap in mental health care by providing less-than- adequate services to already vulnerable populations and further solidify systemic inequities (Timmons *et al.*, 2023).

Fairness in Access and Treatment

Fairness in access to, and the reduction of algorithmic biases in AI chatbot technology is key in the pursuit of equal health care in mental health. Algorithms need to be developed with representation in mind so that people are not treated in a way that would be deemed unfair (Jones *et al.*, 2014). This would include appropriate representative data at the training stage and monitoring continuously to detect and minimize biases (Chaudhuri and Mohanty, 2024). In addition, a regulatory framework is needed, one that makes sure that chatbot technology is equally accessible to all and that, when possible, the design of such technology does not perpetuate inequality (Weerts, 2025).

2.3.5 The Therapist-Patient Relationship

Empathy and Human Connection

Empathy and emotional support in conventional therapy are rather difficult to replicate with AI chatbots. Human therapists provide emotional support through empathetic understanding and personal connection, which is challenging for AI to authentically replicate (Wang, Sharma and Kumar, 2023). Even though chatbots can provide overall simulated empathetic responses because of their programmed algorithms, they lack the genuine emotional insight and human warmth in making personal contacts that contribute to effective therapeutic relationships (Meadows and Hine, 2024). This points to a serious limitation in the effectiveness of chatbots to provide deep, personalized support that many individuals need for meaningful mental health care.

Dependence and Over-Reliance

A major risk concern is overdependence on chatbots for mental health support. This can make individuals neglect human interaction and professional therapy, which are crucial elements of comprehensive mental health care (Khawaja and Bélisle-Pipon, 2023). This would entail one missing out on complex emotional support and subtle understandings which human therapists could offer, hence incomplete or ineffective therapy (Espejo, Reiner and Wenzinger, 2023). A balance in the use of chatbots, therefore, with traditional therapy will go a long way to ensure holistic and effective support for mental health (Molli, 2022).

2.4 Trust and User Acceptance

2.4.1 Factors Influencing Trust in AI Chatbots

Trust in AI chatbots is largely based on transparency and understanding of their capabilities and limitations. Users trust chatbots when they understand they can handle sensitive information and work appropriately under any circumstances, ensuring transparency and understanding (J. Li *et al.*, 2023). Clear communication could help in the demystification of operations of the chatbot and reduce concerns about data privacy and security (Hasal *et al.*, 2021).

Effective communication is instrumental because it builds trust. Chatbots need to communicate what and how they do things to not be jargoned, and so the user can be sure they are talking to an AI and not a human (Ltifi, 2023). This helps manage realistic expectations and minimize chances for misunderstandings that can possibly lead to trust depletion. As the user gets enlightened by the technology, their confidence in the support system of the chatbot increases, which raises their trust and readiness to collaborate with the system (Kohli, 2024).

2.4.2 User Experience (UX) and Interface Design

UX and interface design help determine user acceptance and increase satisfaction with the AI chatbots. Hence, a user-friendly and empathetic interface design can significantly impact how users interact with a chatbot and generally their level of engagement. An intuitive interface that is easy to use enhances user satisfaction due to reduced frustration and increased access to the various features of the chatbot (Casheekar *et al.*, 2024).

This could be further extended to empathetic interfaces that would not only respond but also be sensitive to the emotions of the users and give them supportive feedback. For example, natural language processing chatbots providing soothing responses and detecting emotional cues make the interaction supportive and relatable (Nallur and Finlay, 2023). Users are made to feel valued and understood through this design, thereby increasing their likelihood of continuing to use the chatbot on a regular basis as part of their mental health support routine. In other words, the transparency of communication and considerate UX design play a crucial role in trust and positive user experiences with AI chatbots. These factors can be attended to in order to further improve user acceptance and overall effectiveness of AI-driven mental health interventions.

2.5 Ethical Standards and Frameworks

Ethical standards play a critical role in guiding the responsible development and deployment of AI technologies. The establishment of ethical frameworks for AI is essential to ensure that these technologies operate in ways that respect human rights and promote fairness (Jobin, Ienca and Vayena, 2019). For AI chatbots used in mental health, ethical considerations include maintaining user privacy, ensuring transparency in data use, and avoiding algorithmic biases (Mittelstadt, 2019). The recommendation for regulatory frameworks is to adopt a human-centred approach, prioritizing user well-being and ethical principles, and requiring developers to provide clear information about AI systems(Floridi *et al.*, 2018). Additionally, establishing independent oversight bodies to monitor compliance with ethical standards and address grievances related to AI use in healthcare can enhance accountability and trust (Winfield and Jirotka, 2018).

2.6 Summary

This literature review has discussed the multifaceted nature of AI chatbots in mental health care, highlighting both their potential benefits and the significant ethical challenges they pose. These tools can bridge gaps in mental health care, particularly for underserved populations and those who might be reluctant to seek traditional help. However, the review also underscores critical ethical concerns. Privacy and confidentiality are paramount issues, with risks of data breaches and challenges in maintaining confidentiality due to multiple parties having access to sensitive data

(Saeidnia *et al.*, 2024). The effectiveness and safety of chatbots remain under scrutiny, as the limited evidence base and potential for misdiagnosis pose significant concerns (Abd-Alrazaq *et al.*, 2020). Algorithmic biases and fairness issues further complicate the landscape, potentially leading to discriminatory outcomes and unequal access (Kim *et al.*, 2023).

Balancing the benefits of AI chatbots with these ethical challenges requires ongoing research and the development of comprehensive ethical frameworks. Future studies should focus on long-term impacts, inclusivity across diverse populations, and the creation of robust guidelines to ensure ethical integration (Floridi et al., 2018). Continued dialogue among researchers, developers, and policymakers is essential to navigate these complexities and harness the potential of AI chatbots while safeguarding user well-being and maintaining ethical standards.

3.0 Methodology

This study employs qualitative research to investigate complex ethical issues surrounding AI powered chatbot services in mental health support. The focus is on human experience, behaviour, and socio-ecological aspects, ensuring a more nuanced understanding (Creswell and Poth, 2018). Considering the exploratory nature of this study, which seeks to understand the perspectives of mental health practitioners, AI developers and users, a qualitative methodology serves to capture depth and diversity of participant experiences.

This study used semi-structured interviews to examine ethical issues related to AI chatbots in counselling services. It involves three categories: a) mental health practitioners, represented by students and professors from a university psychology department, b) adults who have used AI chatbots for mental health issues and c) adults who haven't used these technologies. The aim is to gather diverse views and experiences from these participants to ensure a comprehensive understanding of AI chatbots in mental health contexts. In total 14 participants responded and agreed to take part in a semi-structured interview (see Table 1).

More specifically, purposive sampling is used in order to enlist participants suited to generate meaningful data for the research purpose (Palinkas *et al.*, 2015). The sample size is defined in terms of data saturation whereby data collection is ceased once there is no emergence of new themes in the data retrieved (Guest, Bunce and Johnson, 2006). Recruitment of participants was conducted through emails to academics, via social media and University partnership. This type of recruitment is effective as it enables diversity and variation among the participants in terms of experience as well as their opinions. All participants gave informed consent to participate in the study. Interviews were conducted via personal meetings or video calls using various software applications as the respondents' availability and their choices allowed. For this reason, each interview was recorded with permission and transcribed without any names to protect confidentiality (Braun and Clarke, 2006). The data was analysed via manual thematic analysis which helped identify key ethical aspects that are commonly faced when using AI chatbots in mental health care.

Table 1: Participants and their experience with mental health chatbots

Code	Role and use of mental health chatbots
5	Adult - not used
12	Mental health practitioner
1	Student – not used
3	Student – not used
4	Student – not used
7	Student – not used
11	Student – not used
13	Student – not used
15	Student – not used
14	Student – not used/practitioner
8	Student – used chatbots
9	Student – used chatbots
10	Student – used chatbots
6	Student – used chatbots a bit

4.0 Results

4.1 Overview

In the analysis of interviews with three sets of participants, the focus was mainly placed on perspective and experience in the context of AI in mental health care: data privacy and security, emotional connection and empathy, trust and scepticism, the role of AI and its effectiveness, ethical and legal considerations, customization/personalization.

The results showed that while people generally like the ease of access and convenience of AI chatbots, data privacy concerns, ethical issues, and the lack of emotional understanding that can be provided by AI are still very real. Each group presented various concerns and levels of acceptance, reflecting the nuances within trust, perceived efficacy, and ethical use of AI in mental health. This paper, therefore, looks to integrate those insights from the participant groups as a way of drawing a more rounded understanding of the benefits, limitations and ethical issues with AI-powered mental health interventions.

Key Findings:

- AI chatbots provide accessibility and convenience but raise huge concern regarding privacy and security of data.
- The availability of genuine emotional support and empathy that AI can extend is not clear.
- There exists a common feeling of distrust due to concerns about the reliability of AI and manipulation.
- AI is generally seen as an adjunct tool, not a total replacement for human therapists.
- The demand for ethical AI deployment guidelines and practitioner training is increasing, aiming to ensure responsible mental health care applications.
- Users would love more personalized and emotionally intelligent interactions from AI.

4.2 Perspectives of Mental Health Practitioners

The views of nascent mental health practitioners, as represented by a set of University psychology students, were fairly neutral. A strong theme for this sample was Data Privacy and Security, largely identifying concerns over data privacy violations and security which are not strongly implemented. For example, P3 underscored, "Data privacy is a big concern. Without proper security, this sensitive information may be exposed, leading to various consequences." This point was also made by P4, who, in a similar viewpoint, said, "users should know how their data is being handled: transparency in everything builds trust".

Another critical topic was Emotional Connection and Empathy. The psychology students particularly noted strongly that AI could not offer genuine empathy, one of the most vital elements of effective mental health support. P2 added further that "AI lacks the human touch needed in therapy. It can't truly understand the emotional depth of someone's struggles." This view aligns with concerns about the Role and Limitations of AI in Mental Health, where AI was seen to be a competitor but not a substitute to human therapists. P3 said, "AI might help with non-critical issues, but it should totally not be relied on for severe cases. There's the danger of a misdiagnosis."

Another theme identified related to Trust and Credibility of AI. The question of trust took the upper hand, as the participants were highly reticent regarding AI's possibilities to provide reliable and ethical care. One of them framed it as, "There is always the risk of manipulation. AI may be programmed commercially rather than therapeutically". In addition, the students pointed to the necessity to make Ethical and Legal Considerations. The respondents stated their clear interest in developing ethical guidelines that provide for the cautious use of AI. P4 explained this by saying, "The bottom line is its ethical use; we need to make sure that AI will be harnessed responsibly and that practitioners will be aware of, and hence be trained, to handle these kinds of tools."

Accessibility and Convenience were seen by some as an advantage, particularly for initial consultations and destigmatization of mental health issues. According to P4, "AI could serve as a great entry point for those hesitant to seek traditional therapy. It makes mental health support more accessible." The ultimate statement from psychology students is that it can play a supportive role, but AI should not replace human interaction and oversight in mental health care.

Key Findings:

- Data privacy and security are significant concerns among mental health practitioners.
- There is a perceived lack of emotional connection and empathy in AI chatbots.
- Trust issues are prevalent, with fears of potential manipulation and misuse of AI.
- AI is viewed as a supportive tool, not a replacement for human therapists.
- Ethical guidelines and practitioner training are seen as essential for responsible AI implementation.

4.3 Experiences of Adults Using AI Chatbots for Mental Health Aid

The Adults that had sought the use of AI chatbots for their mental health needs were represented both negatively and positively. Consequently, the strong themes that came forth included Data Privacy and Security. Consequently, many respondents had this to say regarding misuse and breaches of data, “I'm always worried about my data being hacked. It is very difficult for people to be able to simply trust them.” (P7) P5 also captured the same idea, emphasizing encryption and security of the information: “There need to be stronger safeguards. Knowing my data is secure gives me peace of mind.”

Another underlying theme was Emotional Connection and Empathy. The users expressed that because of the lack of emotional intelligence, AI chatbots were not able to provide as promising a therapy experience as compared to the human therapists. As shared by P6, “AI just can't understand emotions like a human therapist. Responses feel scripted and without any genuine empathy.” P10 added, “It is obvious that the chatbot does not actually understand what I am going through, it feels more like talking to a machine rather than a person.”

Despite all these concerns, users valued the ease of accessibility and convenience provided by AI chatbots. Many valued the 24/7 availability: “There was support provided during non-traditional hours.” P5 mentioned, “Being able to talk to a chatbot at any time is a big advantage. Comforting in itself it is to know support is available." This accessibility was particularly useful for initial assessments and guidance. For P8 it was easy to use: "It's straightforward to use. I can quickly get advice without the hassle of scheduling an appointment."

Users reflected trust and scepticism in equal measures. While a few trusted the AI for minor problems, the overwhelming feeling was that AI was meant to supplement and not supplant traditional therapy. As P7, said, "AI can't be a substitute for real therapy...It's good for quick help but not for deep, emotional issues." This view also relates to the theme of Role and Effectiveness of AI in Mental Health, where the role of AI was agreed to be supportive. As P9 summed up: "AI chatbots are helpful for initial support but should be integrated with traditional methods for comprehensive care.

Finally, Customization and Personalization was also an important theme in enhancing user satisfaction. Participants valued AI tools offering tailored advice and insight. As one participant noted, "I like it when the chatbot gives personalized advice. It feels more relevant to my situation." However, a great deal of personalization is still needed or sought in this area for the 'therapeutic feeling'.

Key Findings:

- AI lacking emotional intelligence results in a distant therapeutic experience.
- AI chatbots are valued by users because it is always there.
- AI is viewed as more of a complementary tool rather than a substitute for traditional therapy.
- There is a desire for better data security and more tailored interactions with AI.

4.4 Views of Adults Who Have Not Used AI Chatbots

Adults who never used an AI chatbot for mental health treatment revealed concerns and interest in those technologies, highlighted by key themes that shape their views. Some of the major concerns were Data Privacy and Security. Indeed, many participants were extremely sceptical when it came to the safety of their data, eg. P12 said, "I'm worried my data will be exposed.". The privacy of personal information is, therefore, a critical concern, especially in discussing sensitive mental health issues. This was further combined with the expectation of transparency in handling data, where, as P13 described, “I want to know exactly how my data is used, and who has access to it.”

Another underlying theme was Emotional Connection and Empathy, whereby the nonusers felt AI could not offer the emotional support needed. As expressed by P14, “The missing ingredient in AI is the human touch…it cannot replace the empathy and understanding of the human therapist.” This again shows a preference for more human-like interaction and increasing emotional intelligence in AI, as shown by P11, “If AI understood emotions better, then I would consider using it, but now it isn't personal enough.”

Most participants expressed a mix of lack of trust and scepticism, which seemed to be a significant challenge for the adoption of AI. Elements of scepticism include whether AI could pinpoint and provide accurate mental health diagnoses and treatments surfaced throughout the statements of several participants, eg. P12 remarked, “I'm sceptical about how accurate AI can be in diagnosing and treating mental health conditions.” There was an appreciation of the need for trust, which it was claimed, could be obtained with the validation and certification of the tools. P13 remarked, “AI tools need to be accredited, validated by mental health experts to make sure the standards of the tools reach certain standards.”

The role of AI in mental health was perceived to be supplementary and not primary. Participants viewed AI as a potential aid, but not a replacement for human therapists, “I think AI could support traditional therapy, but it shouldn't replace it. It may be helpful for individuals who are too bashful to approach someone in person for their problems” (P11). This view aligns with the theme of Role and Effectiveness of AI in Mental Health Support, which sees AI's role as one of early intervention and as support for human therapists.

Ethical and Legal Considerations: This was another area of concern. The ethical guidelines and practices that will make for responsible use of AI in mental health were called for. P15 identified, “There really need to be clear guidelines on ethics regarding the use of AI in mental health. It is about being responsible with sensitive information." This was reinforced by P14, emphasizing training for practitioners, "Mental health professionals need to be trained in using AI tools effectively and ethically."

Key Findings:

- Data privacy and security are primary concerns for non-users of AI chatbots.
- There is scepticism about AI's ability to provide genuine emotional support.
- Trust issues hinder the adoption of AI in mental health care.
- AI is seen as a potential supplement but not a replacement for human therapists.
- Ethical guidelines and transparency in data handling are crucial to build trust.

4.5 Thematic Analysis of Interviews

Thematic analysis of the interviews reveals a complex landscape regarding the perspectives on AI in mental health care. Data privacy and security, emotional connection and empathy, trust and scepticism, the role and effectiveness of AI, ethical and legal considerations, and customization and personalization are among the top themes allowed for across all participant groups. The themes are consistent, but the emphasis and the cautions differ among the groups.

Psychology students are mainly concerned with ethical implications and practitioner training. Adults having experienced an AI chatbot appreciate the ease and convenience but emphasise that emotional intelligence needs to be developed and enhanced, as does data security. Adults that have never tried an AI chatbot for mental health remain highly sceptical of the reliability of the tool with human emotions and introduce the aspect of human contact and strict ethics guidelines.

4.6 Summary of Key Findings

The findings from this study suggest that although AI chatbots have promoted greater accessibility and convenience for mental health support, data privacy concerns, emotional connection, and trust continue to pose significant barriers to their widespread adoption.

Summary of Key Findings:

- Data Privacy and Security: This is a key concern in all groups, indicating the development and implementation of stringent data protection measures and a need for greater transparency of practices.
- Emotional Connection and Empathy: Some believe that AI lacks emotional understanding, hence it requires, in most cases, human supervision and more enhancement in the emotional intelligence of AI itself.
- Trust and Scepticism: The issues of trusting the use of AI are recurrent, questioning the element of reliability and manipulation and ethical commitment involved in deploying AI for mental health care.
- Role and Effectiveness of AI: Most believe AI to be a complement to, not a substitute for, human therapists. They work best only in the first levels of support and non-critical matters.
- Ethical and Legal Considerations: The huge demand for informed ethical guidance, practitioner training, and ethical certifications on responsible deployment of AI is increasing and is quite urgent.
- Customization and Personalization: More users wish for AI interactions that are personalized in such a way that they can help optimise both the therapeutic experience and relevance of the support offered.

These findings imply that while implementing AI, this area needs to balance technological innovation with ethical responsibility; secondly, this has to be associated with educating the users, while continuously upgrading its AI capabilities towards meeting the users' emotional and security needs.

5.0 Discussion and Conclusion

This study findings raise serious questions about data privacy and security, emotional connection and empathy, trust and scepticism, and the role and effectiveness of AI in supporting mental health. These themes should be looked at as a point of difficulty and challenge using AI in mental health contexts while underlining the need for ethical guidelines, transparency, and human oversight. This discussion develops an understanding of the status quo of AI within current mental health care through comparing these results with the existing literature, and goes further to indicate some recommendations that could help practitioners, policymakers, and developers improve ethical deployment and effectiveness of such technologies (Floridi *et al.*, 2018).

5.1 Data Privacy and Security Concerns

Among the participant groups, the concern for data privacy and security was one of the strongest themes and thus a major barrier to the adoption of AI chatbots powering mental health care. Across the board, the participants expressed fears of possible breaches, disclosures, and misuse of sensitive personal data besides a general lack of clarity around the protection of sensitive personal data. This also aligns with literature to date, highlighting that the application of digital health tools increases concerns on data privacy considerably (Bommu, 2022). Undoubtedly, such fear is amplified when the process of collecting and managing data is not transparent, which then results in the lack of trust in AI systems. Fears, such as unauthorized access into sensitive mental health information or use of data for other purposes rather than those consented, go a long way in eroding confidence in AI tools. Respondents referred to regulatory measures such as the General Data Protection Regulation (GDPR) and mentioned that the said regulations are essentially required frameworks that guarantee the security of data (Floridi *et al.*, 2018). The conformity with GDPR is crucial when considering ethics in the deployment of AI, and GDPR demands transparency, data minimization, and consent from users. The findings would thus indicate that there is a perceived lack of adequate safeguards, and therefore, a need for more stringent enforcement and communication of data protection practices are in place through regulation.

5.2 Emotional Connection and Empathy

One of the key themes that have come out through the contributions of all groups of participants is a lack of emotional understanding and empathy by AI chatbots. This was already very well remarked upon in the literature, as the ability of technology to truly emulate emotional responses in the same way as humans is quite complex (Rubin *et al.*, 2024). Not having this might be linked to a lack of personalized care and less effective support if that is necessary in some critical situations. The psychological implications of a lack of empathy in mental health care are very serious. Empathy is important in developing trust and therapeutic alliances that form part of the bedrock of effective mental health care. Elliott et al. (2011) indicate that a lack of empathy might make users feel alienated and dissatisfied, which could eventually discourage them from seeking help. These findings highlight the importance of human involvement in AI interactions, since this allows for more emotional support and understanding that can improve overall effectiveness with such tools.

5.3 Trust and Scepticism

The scepticism and concerns with regard to reliability and intent were fairly common, and many participants seemed not to trust AI chatbots at all. This goes in line with the prevailing literature on trust in AI, mentioning transparency, reliability, and perceived competence to be some of the most critical factors in determining trust in AI systems (Klossner, 2022). Some issues were surrounding appropriate advice over sensitive issues and bias-free functioning. Trust issues can have a major impact on the level of user engagement and satisfaction with any AI tool. If the user doesn't trust the AI regarding the secure handling of their personal information or the reliability of its support in mental health issues, they are unlikely to interact meaningfully with the technology. Transparency in data use, clarity in the communication of limitations versus the possibilities of AI, together with the assurance that tools based on AI are certified and validated by mental health practitioners, presents important issues around building trust (Zhang and Dafoe, 2019). Trust-building strategies are developed around ethical certification of AI, promoting professional validation through professional accreditation, and fostering ongoing user education into AI capabilities and limitations.

5.4 Role and Effectiveness of AI in Mental Health Support

Participants generally perceived AI as augmenting rather than replacing human therapists. This perception is supported by findings that AI has more potential as a first-line support provider, for non-severe problems, and to improve access to mental health resources (Fitzpatrick, Darcy and Vierhile, 2017). Participants pointed out that AI has the advantages of being available 24/7 and able to respond immediately, which might be highly useful for those users who cannot avail themselves of any other means of treatment. However, limitations concerning intricate mental health issues and situations requiring immediate intervention were discussed. AI is not capable of offering elaborate, personalized care but instead relies on responses pre-programmed in the software; it is in no way designed to attend to grave mental disorders. This suggests that even as AI can accelerate access to mental health support, it is not a replacement for professional care (Jiang *et al.*, 2017). It is here that human oversight by professionals is extremely essential in ascertaining that appropriate care is accorded to users, especially in those cases which require an in-depth understanding of emotional and psychological complexities.

5.5 Limitations of the Study

There are several limitations to the research that could potentially impact the interpretation and generalizability of the findings. The sample was relatively small and focused on the specific demographic groups that may not represent the broader population. Because participants in this study comprised of psychology students and adults with and without experience in the AI chatbot, generalization outside this study may be limited to other groups or other settings (Creswell, 2014).

Secondly, if the interviews themselves are semi-structured, then the eventual reliance upon such a method may introduce interviewer effects and jeopardise the standardisation of responses. The subjective nature of personal experiences and perceptions also suggests that findings can be underpinned by participants preconceptions or social desirability bias (King, Horrocks and Brooks, 2019). In addition, the qualitative data obtained from this study cannot present detailed quantitative data related to AI chatbot efficacy or user behaviour, which limits the thorough understanding of the impact of these tools.

Finally, the limitation is on time constraints, as it affected the depth at which the data collection and analysis could go. Since AI-powered chatbots are a new technology, the findings are subject to rapid changes in the advancement of AI technology. Another challenge is when looking across global ethical standards and regulations, since the ethical and legal place of one region might not be applicable in another one (Floridi *et al.*, 2018).

5.6 Implications for Practice and Policy

Several practical implications of these findings are given for mental health practitioners, policy makers, and AI developers. For the *mental health professions*, the integration of AI into practice should be cautiously approached, making certain that such technologies are used only to complement and never replace human contact. *Practitioners* should be trained in the capabilities and limitations of AI tools to properly guide patients and provide oversight where necessary. The role that *policy makers* play is crucial for ensuring that ethics are deployed with AI into mental health care. It also calls for strict regulations on data privacy and security that protect the sensitive information of users. A clear ethical guideline and standard on privacy laws and ethical standards with respect to AI systems should be audited regularly. Transparency in data handling and consent processes with users should be done to build trust and ensure safe usage when it comes to the implementation of ethical AI frameworks. The *developers* should emphasise enhancing emotional intelligence for improving user satisfaction and experience. That would involve the development of algorithms capable of recognizing emotional cues much more clearly and responsively and providing empathetic support accordingly.

Finally, cultural sensitivity and inclusion should be a part of developers' attention for AI tools to be used for diverse populations with different mental health needs. This entails ongoing development and improvement in the field of AI empathy and adherence to ethical standards. Recruiting expertise from mental health professionals in the development and validation of AI technologies will inform how this would fit with both therapeutic principles and ethical practices (Jiang *et al.*, 2017). This needs effective and responsible collaboration between technologists, clinicians, and policymakers.

References

Abd-Alrazaq, A.A. *et al.* (2020) ‘Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis’, *Journal of Medical Internet Research*, 22(7), p. e16021. Available at: https://doi.org/10.2196/16021.

Bendig, E. *et al.* (2022) ‘The Next Generation: Chatbots in Clinical Psychology and Psychotherapy to Foster Mental Health – A Scoping Review’, *Verhaltenstherapie*, 32(Suppl. 1), pp. 64–76. Available at: https://doi.org/10.1159/000501812.

Bernhardt, M., Jones, C. and Glocker, B. (2022) ‘Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms’, *Nature Medicine*, 28(6), pp. 1157–1158. Available at: https://doi.org/10.1038/s41591-022-01846-8.

Bommu, R. (2022) ‘Ethical Considerations in the Development and Deployment of AI-powered Medical Device Software: Balancing Innovation with Patient Welfare’, *Journal of Innovative Technologies*, 5(1), pp. 1–7.

Braun, V. and Clarke, V. (2006) ‘Using thematic analysis in psychology’, *Qualitative Research in Psychology*, 3(2), pp. 77–101. Available at: https://doi.org/10.1191/1478088706qp063oa.

Casheekar, A. *et al.* (2024) ‘A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions’, *Computer Science Review*, 52, p. 100632. Available at: https://doi.org/10.1016/j.cosrev.2024.100632.

Castella, G. *et al.* (2013) ‘MULTIMODAL AFFECT MODELING AND RECOGNITION FOR EMPATHIC ROBOT COMPANIONS’, *International Journal of Humanoid Robotics*, 10(01), p. 1350010. Available at: https://doi.org/10.1142/S0219843613500102.

Casu, M. *et al.* (2024) ‘AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications’, *Applied Sciences*, 14(13), p. 5889. Available at: https://doi.org/10.3390/app14135889.

Chaudhuri, S. and Mohanty, I. (2024) ‘The Importance of Bias Mitigation in AI: Strategies for Fair, Ethical AI Systems’, 24 July. Available at: https://www.uxmatters.com/mt/archives/2023/07/the-importance-of-bias-mitigation-in-ai-strategies-for-fair-ethical-ai-systems.php#comments (Accessed: 13 February 2025).

Clement, S. *et al.* (2015) ‘What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies’, *Psychological Medicine*, 45(1), pp. 11–27. Available at: https://doi.org/10.1017/S0033291714000129.

Codecademy Team (2023) *Ethics of Chatbots*. Available at: https://www.codecademy.com/article/ethics-of-chatbots (Accessed: 11 February 2025).

Creswell, J.W. (2014) *Research design : qualitative, quantitative, and mixed method approaches*. Fourth edition. Thousand Oaks: Thousand Oaks : SAGA Pub., 2014.

Creswell, J.W. and Poth, C.N. (2018) *Qualitative inquiry & research design : choosing among five approaches.* Fourth edition, International student edition / John W. Creswell, Cheryl N. Poth. Los Angeles: Los Angeles : SAGE, 2018 (Qualitative inquiry and research design : choosing among five approaches).

Elliott, R. *et al.* (2011) ‘Empathy.’, *Psychotherapy*, 48(1), pp. 43–49. Available at: https://doi.org/10.1037/a0022187.

Espejo, G., Reiner, W. and Wenzinger, M. (2023) ‘Exploring the Role of Artificial Intelligence in Mental Healthcare: Progress, Pitfalls, and Promises’, *Cureus* [Preprint]. Available at: https://doi.org/10.7759/cureus.44748.

Fitzpatrick, K.K., Darcy, A. and Vierhile, M. (2017) ‘Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial’, *JMIR Mental Health*, 4(2), p. e19. Available at: https://doi.org/10.2196/mental.7785.

Floridi, L. *et al.* (2018) ‘AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations’, *Minds and Machines*, 28(4), pp. 689–707. Available at: https://doi.org/10.1007/s11023-018-9482-5.

Fulmer, R. *et al.* (2018) 'Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial', *JMIR Mental Health*, 5(4), p. e64. Available at: https://doi.org/10.2196/mental.9782.

Guest, G., Bunce, A. and Johnson, L. (2006) 'How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability', *Field Methods*, 18(1), pp. 59–82. Available at: https://doi.org/10.1177/1525822X05279903.

Hasal, M. *et al.* (2021) 'Chatbots: Security, privacy, data protection, and social aspects', *Concurrency and Computation: Practice and Experience*, 33(19), p. e6426. Available at: https://doi.org/10.1002/cpe.6426.

Hildt, E. and Laas, K. (2022) 'Informed Consent in Digital Data Management', in K. Laas, M. Davis, and E. Hildt (eds) *Codes of Ethics and Ethical Guidelines*. Cham: Springer International Publishing (The International Library of Ethics, Law and Technology), pp. 55–81. Available at: https://doi.org/10.1007/978-3-030-86201-5_4.

Hofmann, V. *et al.* (2024) 'AI generates covertly racist decisions about people based on their dialect', *Nature*, 633(8028), pp. 147–154. Available at: https://doi.org/10.1038/s41586-024-07856-5.

Inkster, B., Sarda, S. and Subramanian, V. (2018) 'An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study', *JMIR mHealth and uHealth*, 6(11), p. e12106. Available at: https://doi.org/10.2196/12106.

Jiang, F. *et al.* (2017) 'Artificial intelligence in healthcare: past, present and future', *Stroke and Vascular Neurology*, 2(4), pp. 230–243. Available at: https://doi.org/10.1136/svn-2017-000101.

Jobin, A., Ienca, M. and Vayena, E. (2019) 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, 1(9), pp. 389–399. Available at: https://doi.org/10.1038/s42256-019-0088-2.

Jones, S.P. *et al.* (2014) 'How Google's "Ten Things We Know To Be True" Could Guide The Development Of Mental Health Mobile Apps', *Health Affairs*, 33(9), pp. 1603–1611. Available at: https://doi.org/10.1377/hlthaff.2014.0380.

Kessler, R.C. *et al.* (2005) 'Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication', *Archives of General Psychiatry*, 62(6), p. 593. Available at: https://doi.org/10.1001/archpsyc.62.6.593.

Khawaja, Z. and Bélisle-Pipon, J.-C. (2023) 'Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots', *Frontiers in Digital Health*, 5, p. 1278186. Available at: https://doi.org/10.3389/fdgth.2023.1278186.

Kim, J. *et al.* (2023) 'Assessing Biases in Medical Decisions via Clinician and AI Chatbot Responses to Patient Vignettes', *JAMA Network Open*, 6(10), p. e2338050. Available at: https://doi.org/10.1001/jamanetworkopen.2023.38050.

King, N., Horrocks, C. and Brooks, J.M. (2019) *Interviews in qualitative research*. Second edition. Los Angeles: Los Angeles : SAGE, 2019.

Klossner, S. (2022) 'AI powered m-health apps empowering smart city citizens to live a healthier life–The role of trust and privacy concerns'.

Kohli, S. (2024) 'Building User Trust in Conversational AI: The Role of Explainable AI in Chatbot Transparency', *International Journal of Computer Engineering and Technology (IJCET)*, 15(5), pp. 406–413. Available at: https://doi.org/10.5281/ZENODO.13833413.

Kommisetty, P.D.N.K. and Dileep, V. (2024) ‘Robust Cybersecurity Measures: Strategies for Safeguarding Organizational Assets and Sensitive Information’, *IJARCCE*, 13(8). Available at: https://doi.org/10.17148/IJARCCE.2024.13832.

Li, H. *et al.* (2023) ‘Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being’, *npj Digital Medicine*, 6(1), p. 236. Available at: https://doi.org/10.1038/s41746-023-00979-5.

Li, J. *et al.* (2023) ‘Determinants Affecting Consumer Trust in Communication With AI Chatbots: The Moderating Effect of Privacy Concerns’, *Journal of Organizational and End User Computing*, 35(1), pp. 1–24. Available at: https://doi.org/10.4018/JOEUC.328089.

Li, J. (2023) ‘Security Implications of AI Chatbots in Health Care’, *Journal of Medical Internet Research*, 25, p. e47551. Available at: https://doi.org/10.2196/47551.

Looi, J.C. *et al.* (2024) ‘Psychiatric electronic health records in the era of data breaches – What are the ramifications for patients, psychiatrists and healthcare systems?’, *Australasian Psychiatry*, 32(2), pp. 121–124. Available at: https://doi.org/10.1177/10398562241230816.

Ltifi, M. (2023) ‘Trust in the chatbot: a semi-human relationship’, *Future Business Journal*, 9(1), p. 109. Available at: https://doi.org/10.1186/s43093-023-00288-z.

Luxton, D.D. (2014) ‘Recommendations for the ethical use and design of artificial intelligent care providers’, *Artificial Intelligence in Medicine*, 62(1), pp. 1–10. Available at: https://doi.org/10.1016/j.artmed.2014.06.004.

Meadows, R. and Hine, C. (2024) ‘Entanglements of Technologies, Agency and Selfhood: Exploring the Complexity in Attitudes Toward Mental Health Chatbots’, *Culture, Medicine, and Psychiatry*, 48(4), pp. 840–857. Available at: https://doi.org/10.1007/s11013-024-09876-2.

Miner, A.S., Milstein, A. and Hancock, J.T. (2017) ‘Talking to Machines About Personal Mental Health Problems’, *JAMA : the journal of the American Medical Association*, 318(13), pp. 1217–1218.

Mittelstadt, B. (2019) ‘Principles alone cannot guarantee ethical AI’, *Nature Machine Intelligence*, 1(11), pp. 501–507. Available at: https://doi.org/10.1038/s42256-019-0114-4.

Molli, V.L.P. (2022) ‘Effectiveness of AI-Based Chatbots in Mental Health Support: A Systematic Review’, *Journal of Healthcare AI and ML*, 9(9), pp. 1–11.

Morley, J. *et al.* (2020) ‘The ethics of AI in health care: A mapping review’, *Social Science & Medicine*, 260, p. 113172. Available at: https://doi.org/10.1016/j.socscimed.2020.113172.

Nallur, V. and Finlay, G. (2023) ‘Empathetic AI for ethics-in-the-small’, *AI & SOCIETY*, 38(2), pp. 973–974. Available at: https://doi.org/10.1007/s00146-022-01466-3.

Palinkas, L.A. *et al.* (2015) ‘Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research’, *Administration and Policy in Mental Health and Mental Health Services Research*, 42(5), pp. 533–544. Available at: https://doi.org/10.1007/s10488-013-0528-y.

Patel, V. *et al.* (2018) ‘The Lancet Commission on global mental health and sustainable development’, *The Lancet*, 392(10157), pp. 1553–1598. Available at: https://doi.org/10.1016/S0140-6736(18)31612-X.

Rezaeikhonakdar, D. (2023) ‘AI Chatbots and Challenges of HIPAA Compliance for AI Developers and Vendors’, *Journal of Law, Medicine & Ethics*, 51(4), pp. 988–995. Available at: https://doi.org/10.1017/jme.2024.15.

Rothstein, M.A. (2023) ‘Should Chatbots Be Used to Obtain Informed Consent for Research?’, *Ethics & Human Research*, 45(6), pp. 46–50. Available at: https://doi.org/10.1002/eahr.500190.

Rubin, M. *et al.* (2024) ‘The Value of Perceiving a Human Response: Comparing Perceived Human versus AI-Generated Empathy’. Open Science Framework. Available at: https://doi.org/10.31219/osf.io/ng97s.

Saeidnia, H.R. *et al.* (2024) ‘Ethical Considerations in Artificial Intelligence Interventions for Mental Health and Well-Being: Ensuring Responsible Implementation and Impact’, *Social Sciences*, 13(7), p. 381. Available at: https://doi.org/10.3390/socsci13070381.

Sebastian, G. (2023) ‘Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information’, *International Journal of Security and Privacy in Pervasive Computing*, 15(1), pp. 1–14. Available at: https://doi.org/10.4018/IJSPPC.325475.

Smalley, K.B., Warren, J.C. and Rainer, J.P. (2012) *Rural mental health : issues, policies, and best practices*. 1st ed. New York, NY: New York, NY : Springer, c2012.

Steel, Z. *et al.* (2014) ‘The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013’, *International Journal of Epidemiology*, 43(2), pp. 476–493. Available at: https://doi.org/10.1093/ije/dyu038.

Surani, A. and Das, S. (2022) ‘Understanding Privacy and Security Postures of Healthcare Chatbots’, in. Available at: https://api.semanticscholar.org/CorpusID:253105981.

Timmons, A.C. *et al.* (2023) ‘A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health’, *Perspectives on Psychological Science*, 18(5), pp. 1062–1096. Available at: https://doi.org/10.1177/17456916221134490.

Torous, J. *et al.* (2021) ‘The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality’, *World Psychiatry*, 20(3), pp. 318–335.

Tschandl, P. *et al.* (2020) ‘Human–computer collaboration for skin cancer recognition’, *Nature Medicine*, 26(8), pp. 1229–1234. Available at: https://doi.org/10.1038/s41591-020-0942-0.

Vaidyam, A.N. *et al.* (2019) ‘Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape’, *The Canadian Journal of Psychiatry*, 64(7), pp. 456–464. Available at: https://doi.org/10.1177/0706743719828977.

Wang, X., Sharma, D. and Kumar, D. (2023) ‘A Review on AI-based Modeling of Empathetic Conversational Response Generation’, in *2023 Asia Conference on Cognitive Engineering and Intelligent Interaction (CEII). 2023 Asia Conference on Cognitive Engineering and Intelligent Interaction (CEII)*, Hong Kong, Hong Kong: IEEE, pp. 102–109. Available at: https://doi.org/10.1109/CEII60565.2023.00026.

Waseem, D. *et al.* (2024) ‘Consumer vulnerability: understanding transparency and control in the online environment’, *Internet Research*, 34(6), pp. 1992–2030. Available at: https://doi.org/10.1108/INTR-01-2023-0056.

Weerts, S. (2025) 'Generative AI in public administration in light of the regulatory awakening in the US and EU', *Cambridge Forum on AI: Law and Governance*, 1, p. e3. Available at: https://doi.org/10.1017/cfl.2024.10.

Winfield, A.F.T. and Jirotka, M. (2018) 'Ethical governance is essential to building trust in robotics and artificial intelligence systems', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), p. 20180085. Available at: https://doi.org/10.1098/rsta.2018.0085.

You, Y. *et al.* (2023) 'Beyond Self-diagnosis: How a Chatbot-based Symptom Checker Should Respond', *ACM Transactions on Computer-Human Interaction*, 30(4), pp. 1–44. Available at: https://doi.org/10.1145/3589959.

Zhang, B. and Dafoe, A. (2019) 'Artificial Intelligence: American Attitudes and Trends', *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.3312874.

Understanding the Socio-Technical Impacts of AI on Intelligence Analysis

Kathleen Vogel*, Katina Michael*, Hussein Abbass^
**Arizona State University, ^UNSW Canberra*

Abstract

Almost every profession in existence today has been impacted by artificial intelligence (AI)/ generative AI (GenAI). The field of intelligence analysis has especially changed as a result of AI-based emerging technologies. In this paper we present research-in-progress that aims to investigate how AI will impact intra-organizational intelligence boundaries but also across organizational and ally boundaries (i.e., inter-organizational and extra-organizational). In doing so the socio-technical impacts can be studied with respect to organizational culture, human performance, integrity of information, information sharing among partners, and public trust. The proposed study will include a literature review across six domains, conduct interviews with key informants and deliberative workshops, in addition to an experiment simulated in the Metaverse providing the ability to study a multi-perspectival (micro, meso and macro) phenomenon with interconnected and interdependent stakeholders. While AI/GenAI provide new capabilities, they are likely to be an aid and augmentation to humans and not a replacement.

Keywords: AI, intelligence analysis, socio-technical impacts, business transformation

1.0 Introduction

Over the past few years, there have been a series of public and government concerns expressed about the potential impacts of AI on the defence and security communities (Cohen, 2023; Jenkins, 2023; Morgan et al., 2020; Johnson, 2019). To date, most scholarly work in the social sciences has focused on the impacts of AI on military troops (Rashid et al., 2023; Morgan et al., 2020; David, 2019; Lele, 2019), with few studies focused on how AI might affect intelligence communities and the knowledge they produce for decision-makers (Vogel et al., 2021). Disruptive technologies, such as AI, could offer an efficiency dividend to intelligence organizations. Still, the full impact of such disruption is far from trivial and could cascade not just within intra-organizational boundaries but also across organizational and ally boundaries (i.e., inter-organizational and extra-organizational), including impacts on organizational culture, human performance, integrity of information, information sharing among partners, and public trust. The larger and more complex the autonomy in the workflow system, the greater the latency in discovering malicious interferences and vulnerabilities. Complex controls could reduce gains, not to mention hide more vulnerabilities. Balancing human involvement and autonomy in intelligence analysis is a delicate decision that can only be made when the chain of negative reactions is mapped and assessed properly. In this paper, we pose the following research question: what is the impact of AI on the effectiveness and efficiency of an intelligence organisation? To answer this question, we explore the interaction of the micro, meso and macro levels.

2.0 Literature Review

There have been some studies published on the use of wargames and table-top exercises to generate insights for understanding how a variety of factors can shape military action (Dorton et al., 2023). There are few empirical studies that have focused on understanding how new technologies like artificial intelligence can shape intelligence analysis at various levels: individual, organizational, and national (government agency/ societal) (Dorton et al., 2022; Dorton and Harper, 2022a; Dorton and Harper, 2022b; Vogel et al., 2021). However, there are no published studies to date that examine the non-linear effects of AI on intelligence analysis from a whole-of-system micro-level (analyst), meso-level (organization), and macro-level (socio-political-economic ecosystem) context. We need to better understand how the use of AI in intelligence can lead to beneficial and harmful disruptions in the defence intelligence workflow.

Furthermore, swarm drone systems bring new challenges to tactical intelligence, including challenges in distributed sensor deployment and management, guidance of the swarm with least resources, adversarial swarming effects, and counter-swarming (Foster and Petty, 2021). First, these swarm systems accelerate and complicate the time taken in an intelligence cycle. That cycle commences from guided data acquisition by many drones, and integration of this information/knowledge, to the production of actionable intelligence, and is well below the time needed for effective operational responses. These systems will bring challenges for the human analyst in the loop: a human must be in control of hundreds of drones needing real-time guidance for intelligence, surveillance, and reconnaissance functions. This new environment needs more advanced AI technologies to be sitting between the analyst and the swarm. Second, the majority of the analysis needs to happen at a speed that can only be achieved through automated decision-making (ADM). The human analyst needs to be able to understand the collective intelligence gathered to be able to direct the swarm towards areas with the highest information gain (IG). Third, the human analyst needs to collaborate with multiple agencies to enable timely situational risk assessment and how this will affect authorization of command and response. This environment is likely to shake up the practices, procedures, and policies within intelligence organizations and lead to transformations in these organizations on multiple fronts. Moreover, longstanding socio-technical and cultural barriers within intelligence communities can also impede effective use of these technologies (Vogel et al., 2021; Treverton, 2016; Turnley and McNamara, 2016; Johnston, 2005). This new AI-enabled environment can also affect national and international security raising important ethical, legal and social issues within and across allied countries, such as the Five Eyes intelligence alliance, regarding the accelerated intelligence collection and analysis process, and the need to incorporate important societal feedback loops into this process (U.S. Congress, 2024). Before any solution can be proposed, the problem space needs to be better understood.

2.1 Conceptual Framework

This collaborative research project between stakeholders in the USA and Australia will take a co-design approach to analysing the socio-technical dynamics of the use of AI in intelligence analysis that considers the interplay between macro, meso and micro layers taking a design thinking approach which is inherently iterative. The micro-level relies on a constructive experimental swarm-metaverse environment (Nguyen et al., 2023), whereby human-AI robot scenarios are designed to assess design options for intelligence gathering, situational awareness, effectiveness, efficiency, and other metrics of performance of intelligence analysis under different levels of autonomy. The meso-level will explore the organizational processes and procedures in play with these technologies within an intelligence organization. The macro-level will examine the governance, controls and information sharing across the Five Eyes and potential ethical, legal, and social, implications of this environment. We will also explore the interplay between the micro, meso, and macro layers and throughout the system to diagnose risks and design mitigation strategies for the future deployment of these technologies in intelligence analysis. A representation of the conceptual framework of this study is show in Figure 1. While the stand-alone layers (micro-meso-macro) are depicted with their anticipated bounds, it is the learning cycles, stakeholder interconnections, and entity interdependencies between the layers that will shed the greatest light on the impact of AI on the practice of intelligence analysis.

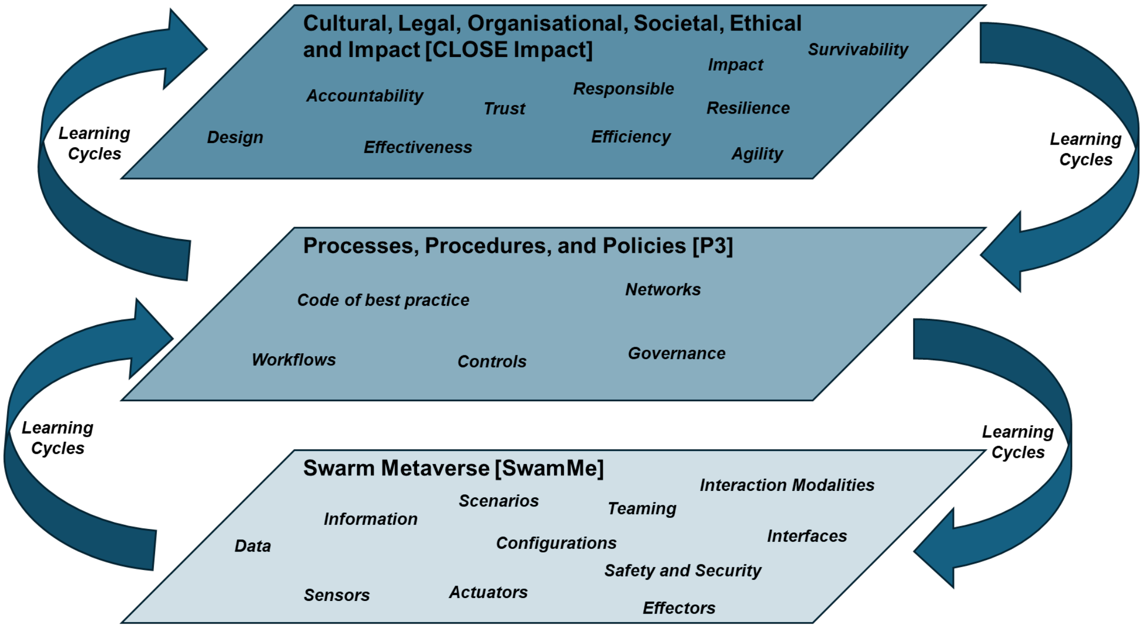


Figure 1. AIS2INT Multi-Level Perspective: Micro (SwamMe), Meso (P3) and Macro (CLOSE Impact)

3.0 Methodology

The research project will use four research methodologies across three years: (1) desk research; (2) interviews with intelligence practitioners; (3) metaverse experiment; and (4) deliberative workshops.

3.1 Phase One: Desk Research

For the desk research component of the project we aim to conduct literature reviews of distinct and diverse disciplinary perspectives, using multiple sources of evidence, relevant to the project: (1) swarm systems from an engineering perspective; (2) human-centred AI design approaches; (3) socio-technical systems literature with an emphasis on multi-level analysis and successful transitions; (4) human-machine teaming literature from a human-systems engineering perspective; (5) knowledge management literature with respect to business processes; and (6) cognitive and organizational psychology literature. These are distinct sets of literatures that have not yet been brought into conversation with respect to the impact of AI on intelligence analysis from a multi-level perspective. In this integrated literature review we will explore: (1) the impact of changing relationships to knowledge and skill development in relation to information processed with minimum input from humans regarding AI in intelligence analysis; (2) the impact of emerging tech such as AI on the nature of intelligence analysis; (3) the relationship between emerging tech (AI and swarm in our case) and broader ecosystem relationships (e.g., national, Five Eyes).

3.2 Phase Two: Interviews with Intelligence Practitioners

The second component of the project, entails the conduct of interviews with approximately 40 intelligence practitioners who have experience in defence intelligence analysis and can provide real-world perspectives on the socio-technical issues that they have faced in their careers and what challenges regarding AI in intelligence might exist in the future. Australian and US-based intelligence communities will be the target of the interviews providing for a comparative analysis of the socio-technical issues faced by intelligence analysts. The national intelligence community in Australia is made up of ten agencies that protect and enhance Australia's security and sovereignty, while the U.S. landscape is larger and more complex, made up of 18 organisations, and elements within organisations (see Table 1).

Australian Intelligence Agencies	**United States Intelligence Organisations**
Australian Criminal Intelligence Commission (ACIC)	Director of National Intelligence (ODNI)
Australian Federal Police (AFP)	Central Intelligence Agency (CIA)
Australian Geospatial-Intelligence Organisation (AG-IO)	Defense Intelligence Agency (DIA)
Australian Secret Intelligence Service (ASIS)	National Security Agency (NSA)
Australian Security Intelligence Organisation (ASIO)	National Geospatial- Intelligence Agency (NGA)
Australian Signals Directorate (ASD)	National Reconnaissance Office (NRO)
Australian Transaction Reports and Analysis Centre	Intelligence elements of the five DoD services; the Army, Navy, Marine Corps, Air Force, and Space Force
Defence Intelligence Organisation (DIO)	Department of Energy's Office of Intelligence and Counter-Intelligence
Department of Home Affairs (DHA)	Department of Homeland Security's Office of Intelligence and Analysis
Office of National Intelligence (ONI)	U.S. Coast Guard Intelligence

Australian Intelligence Agencies	United States Intelligence Organisations
	Department of Justice's Federal Bureau of Investigation (FBI)
	Drug Enforcement Administration's Office of National Security Intelligence
	Department of State's Bureau of Intelligence and Research
	Department of the Treasury's Office of Intelligence and Analysis

Table 1. Interview source agencies in the Australian and U.S. intelligence communities. Sources: https://www.intelligence.gov.au/agencies, and https://www.dni.gov/index.php/what-we-do/members-of-the-ic.

3.3 Phase Three: Metaverse Experiment

From the desk research and interviews, we propose constructing and executing a Metaverse experiment (Nguyen et al., 2023) involving 20 experts and intelligence practitioners centred on the use of AI in intelligence analysis involving a swarm system and a threat scenario involving Five Eyes partners. We will collect and analyse data from the experiment, and also conduct post-interaction interviews with the participants after the experiment, to explore the micro, meso, and macro level impacts on defence intelligence. For example, through this experiment we will explore: (1) How to measure the impact of AI on the intelligence community?; (2) How to measure impact on multiple levels (macro, meso and micro)? (3) How to connect the measures across levels?; (4) How to measure impact in a non-stationary environment, where disruptive technologies are not fixed and changing over time; (5) How to what extent will AI change the nature of analytic work in the intelligence community?

We propose a Metaverse experiment, because these kinds of assessments need to occur in a psychologically safe-environment—where consequences of decisions do not have severe impact on real systems, distant from the real-environment, while the experimental design needs to ensure results are transferable. Such an interaction will also facilitate multi-stakeholder consultation through the development of a policy sandbox (i.e., within a closed environment), where insights can be garnered on anticipatory governance mechanisms for future AI enabled intelligence deployments, thus better addressing the traditional challenge of the policy-practice pacing problem (Michael et al. 2024). Before the Metaverse experiment is launched we will hold a deliberative workshop with a select group of approximately 15 experts to test the proposed experimental setup and obtain feedback on further refinement of the experimental design and its contents. Once refinements have been completed, an additional deliberative workshop will be held to obtain final recommendations.

The socio-cultural implications of national and international security will also be studied, with an emphasis on raising allied partner awareness and education, which is also an understudied and undertheorized area. Thus, the experiment will map out the policy implications of this national and international security environment using PESTLE dimensions (Political, Economic, Sociological, Technological, Legal and Environmental). PESTLE dimensions allow for the examination of external factors in the global market, sometimes known as megatrends, that have the capacity to introduce opportunities or threats to an entity (Rastogi & Trivedi, 2016). An experimental design that incorporates these implications in the Metaverse can inform the design and development of real-world adaptation of AI by intelligence communities. These PESTLE factors will act as a global set of levers that will orient the context (e.g., with respect to the level of regulation or societal acceptance) from the macro right through to the micro, impacting decisions.

4.0 Preliminary Results

AI-enabled business transformation is a critical aspect of any organisation; this is equally true of governmental agencies, defence, and intelligence organisations that act to leverage new technologies for the conduct of intelligence analysis toward global security. While government agencies may have been slow to adopt traditionally, players in the intelligence ecosystem require greater agility to meet expectations given the increasing complexity in the threat landscape, requiring intra-/ inter-/ and extra-organisational flows of information and communication. While much of this change has occurred due to the way intelligence analysis functioned in the past (Lefebvre, 2004; Marrin 2007; Odom 2008; National Research Council 2011), the introduction of decision science and subsequently big data analytics (Dhami et al. 2015), as well as artificial intelligence (Barnea 2020; Vogel et al. 2021), has markedly brought into question the impact of AI on intelligence analysis.

While the process mapping of what an intelligence analyst does has been well documented in the literature (e.g., Mitchell et al. 2019), early surveys of the intelligence community noted that: “the main features that the users wanted to have in the intelligent user agent were: having the agent look over the analyst’s shoulder to make sure that he/she wasn’t omitting useful facts and hypotheses in solving a case; making sure that the hypotheses and resulting conclusion seem consistent and reasonable; helping the analyst step through the analysis phase and key strokes in the Wisdom Builder product; having the agent help the analyst in the thinking and reasoning processes involved in making intelligence judgments” (Phillips et al. 2001, p. 60). “Wisdom Builder” relied on case-based reasoning (CBR), while today, AI has completely challenged many aspects of an intelligent analyst’s work tasks (Ghioni et al., 2024; Hepenstal et al. 2019). AI capabilities today are allowing intelligent analysts to rethink the way they do business, and creating considerations around speculative changes from the traditional *present mode of operation* to *future mode of operation*. While these are fraught with ethical considerations both for the individual analyst “as employee”, they have vast socio-technical impacts for organisations and implications for society at large (Blanchard and Taddeo, 2023, pp. 11-12).

While AI will never replace the role of the intelligence analyst, AI tools will, and are already, augmenting the way tasks are done in intelligence organisations (Regens, 2019). These capabilities considered as standard practice of *future modes of operation* can be clustered into the 4 following categories: (1) summarisation and translation (e.g., Sufi, 2024); (2) data processing and synthesis (e.g., Ganor, 2021); (3) predictive analysis (e.g., Dahl and Strachan-Morris, 2024); and (4) report validation, accuracy and bias (e.g., Turner, 2024). In summarisation and translation are the ability to use GenAI to condense large numbers of documents (quantitative or qualitative), conduct open-source combined with *other sources of evidence* to conduct sentiment analysis using large language models (LLMs) with available datasets dating back to the previous year. The sentiment analysis could be related to organisations, events, or actors, allowing for comparative analysis and the extrapolation of shared patterns that may aid in cases (Reddy et al., 2024). Additionally, data processing and synthesis allows for the recombination of structured and unstructured data from multiple sources. Analysts with this capability can perform sorting on large datasets (e.g., based on date/time). Images and video can be labelled and patterns recognised in addition to trend identification uncovering hidden insights. Predictive analysis uses identified patterns to predict potential threats. Entrenched in this practice are the adoption of scenarios, vignettes, and the forecasting of future events and socio-technical impacts. Finally, report validation allows for reports to be written in an unbiased manner, challenging the intelligence analyst with key outcomes and judgments.

There is a great deal of training that is required to bring intelligence organisations to a point where decisions can be made to adopt or reject certain tools and techniques, but the transformation must be done through awareness, education and an understanding of the benefits, risks and shortcomings that AI-enabled intelligence analysis may herald. Business transformation in any organisation also requires cultural transformation to ensure the successful deployment of any emerging technology, and this begins with empathy (Michael et al., 2024).

5.0 Potential Implications for National Defence Intelligence

Our project aims to better understand the impact of artificial intelligence (AI) on intelligence analysis in defence intelligence organizations. The research will encourage interaction between stakeholders at multiple perspectives (macro, meso, micro). It will shed light on the strategic, operational, and tactical use of AI in a variety of scenario contexts within the intelligence community. The study's contribution will be in shifting the focus from using AI as a siloed mechanism (e.g., a given operational decision to deploy troops on the battlefield), to developing a Metaverse environment where stakeholder perspectives can be shared and a more holistic process of decision-making take place, acknowledging the role of both humans and AI.

This is an important interdisciplinary area to research, given the U.S., UK, and Australia joint cooperation in signals intelligence formalised in the multilateral UK-USA Agreement known as the Five Eyes intelligence sharing alliance, and with advancing the U.S., UK, and Australia AUKUS partnership. This project will glean important understandings on the implications of AI for the intelligence analyst. This research can help defence intelligence organizations better plan for the implications of AI technologies into intelligence work, which are priorities indicated in the U.S. Office of the Director of Intelligence 2019 Augmenting Intelligence through Machines (AIM) Initiative, the 2023 Executive Order on the Safe, Secure, and

Trustworthy Development and Use of Artificial Intelligence, and the 2023 U.S. Department of Defence Data, Analytics, and Artificial Intelligence Adoption Strategy, and have informed guidance by the Australian Signals Directorate's Australian Cyber Security Centre (ASD's ACSC). This work is essential and vital to achieve situational awareness and operational superiority in today's environment.

6.0 Conclusion

This project will help advance basic social science research on human-centred AI, human-machine teaming, knowledge management, and cognitive/organizational psychology as it relates to defence and security knowledge-workers, as well as create novel insights from the integration of these fields. This work will also advance thinking on the ethical, social, and legal implications of AI and emerging technologies related to defence applications with allied partners. The research has been supported by a seed grant of the PLuS Alliance's Security & Defence+ unit, and more information about the study can be found here: https://securityanddefenceplus.plusalliance.org/ (Michael et al., 2025).

References

Abbass, Hussein A., Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust, *Cognitive Computation*, 11(2), (2019), 159-171.

Abbass, Hussein A., Eleni Petraki, Kathryn Merrick, John Harvey, and Michael Barlow, Trusted autonomy and cognitive cyber symbiosis: Open challenges, *Cognitive Computation* 8, (2016), 385-408.

Barnea, A., How will AI change intelligence and decision-making? *Journal of Intelligence Studies in Business*, 10(1), (2020), 75-80.

Blanchard, A., Taddeo, M. The Ethics of Artificial Intelligence for Intelligence Analysis: a Review of the Key Challenges with Recommendations, *Digital Society*, 2, 12 (2023), https://doi.org/10.1007/s44206-023-00036-4

Cohen, Charles, AI in Defense: Navigating Concerns, Seizing Opportunities, *National Defense* (July 25, 2023), https://www.nationaldefensemagazine.org/articles/2023/7/25/defense-department-needs-a data-centric-digital-security-organization.

Dahl, E. J., & Strachan-Morris, D., 'Predictive intelligence for tomorrow's threats': is predictive intelligence possible?, *Journal of Policing, Intelligence and Counter Terrorism*, 19(4), (2024), 423-435.

Davis, Z., Artificial intelligence on the battlefield, *Prism*, 8(2), (2019), 114-131.

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E., Improving intelligence analysis with decision science, *Perspectives on Psychological Science*, 10(6), (2015), 753-757.

Dorton, Stephen, Theresa Fersch, Emily Barrett, Andrew Langone, Mark Seip, Shane Bilsborough, Curtis B. Hudson, Paul Ward, and Kelly Neville, The Value of Wargames and Tabletop Exercises as Naturalistic Tools, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67(1), (2023), 2454-2459. https://doi.org/10.1177/21695067231192617.

Dorton, Stephen, Samantha B. Harper, and Kelly Neville, A Naturalistic Investigation of Trust, AI, and Intelligence Work, *Journal of Cognitive Engineering and Decision Making*, 16(2), (May 2022), https://doi.org/10.1177/15553434221103718.

Dorton, Stephen, Samantha B. Harper, and Kelly Neville, Adaptations to Trust Incidents with Artificial Intelligence, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 66(1), (October 2022), 95-99, https://doi.org/10.1177/1071181322661146.

Dorton, Stephen and Samatha B. Harper, Self-Repairing and/or Buoyant Trust in Artificial Intelligence, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 66(1), (October 2022), 162-166, https://doi.org/10.1177/1071181322661098.

Foster, Foster Kevin L., Petty Mikel D., Estimating the tactical impact of robot swarms using a semi-automated forces system and design of experiments methods, *The Journal of Defense Modeling and Simulation*, 18(3), (2021), 247-269, https://doi.org/10.1177/15485129211008532.

Ganor, Boaz, Artificial or human: A new era of counterterrorism intelligence?, *Studies in Conflict & Terrorism*, 44(7), (2021), 605-624.

Ghioni, R., Taddeo, M., & Floridi, L., Open source intelligence and AI: a systematic review of the GELSI literature, *AI & Society*, 39(4), (2024), 1827-1842.

Hepenstal, S., Zhang, L., & Wong, B. W., An analysis of expertise in intelligence analysis to support the design of Human-Centered Artificial Intelligence, *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2021, October), 107-112, IEEE.

Jenkins, Michael P., The impact and associated risks of AI on future military operations, *Federal News Network*, (October 18, 2023), https://federalnewsnetwork.com/commentary/2023/10/the-impact-and associated-risks-of-ai-on-future-military-operations/.

Johnson, James, Artificial intelligence & future warfare: implications for international security, *Defense & Security Analysis*, 35(2), (2019), 147-169, https://doi.org/10.1080/14751798.2019.1600800.

Johnston, Rob, Analytic Culture in the U.S. Intelligence Community: An Ethnographic Study (Washington, DC: Center for the Study of Intelligence, 2005).

Lefebvre, S., A look at intelligence analysis, *International Journal of Intelligence and CounterI*, 17(2), (2004), 231-264.

Lele, A., Artificial Intelligence (AI), In: *Disruptive Technologies for the Militaries and Security. Smart Innovation, Systems and Technologies*, 132. Springer, Singapore, https://doi.org/10.1007/978-981-13-3384-2_8.

Marrin, S., Intelligence analysis theory: Explaining and predicting analytic responsibilities. *Intelligence and National Security*, 22(6), (2007), 821-846.

Michael, Katina, Roba Abbas, Jeremy Pitt, Kathleen M. Vogel, and Mariana Zafeirakopoulos, Securitization for Sustainability of People and Place, *IEEE Technology and Society Magazine*, 42(2), (June 2023), 22-28, https://doi.org/10.1109/MTS.2023.3283829.

Michael, Katina, Jordan R. Schoenherr, and Kathleen M. Vogel, Failures in the Loop: Human Leadership in AI-Based Decision-Making, *IEEE Transactions on Technology and Society*, 5(1), (March 2024), 2-13, https://doi.org/10.1109/TTS.2024.3378587.

Michael, Katina, Kathleen N. Vogel, Hussein Abbass, AI and Swarm Systems Disruptive Impacts on Intelligence and Global Security, *Security & Defence PLuS*, (14 January 2025), https://www.youtube.com/watch?v=FPW69FeKD1w.

Michael, K., Belinda Gibbons, Tim Vo, Amanda Winks, Monique Watts, Empowering Employees in Organizations using Empathy: Applying Design Thinking Practices in the Pursuit of Cultural Change Aiding in Successful Innovation and Business Transformation, *Australia & New Zealand Academy of Management*, Workshop, Wollongong, NSW, Australia, (4 December 2024).

Mitchell, K., Mariani, J., Routh, A., Keyal, A., & Mirkow, A., *The future of intelligence analysis,* Deloitte Insights, (2019).

Morgan, Forrest E., Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derik Grossman, *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World* (Santa Monica: RAND, 2020), https://www.rand.org/pubs/research_reports/RR31391.html.

National Research Council, Division of Behavioral, Social Sciences, Board on Behavioral, Sensory Sciences, Committee on Behavioral, & Social Science Research to Improve Intelligence Analysis for National Security, (2011), *Intelligence analysis for tomorrow: Advances from the behavioral and social sciences,* National Academies Press.

Nguyen, Hung, Aya Hussein, Matthew A. Garratt, and Hussein A. Abbass, Swarm metaverse for multi-level autonomy using digital twins, *Sensors*, 23(10), (2023), 4892.

Odom, W. E. (2008). Intelligence analysis, *Intelligence and National Security*, 23(3), 316-332.

Phillips, J., Liebowitz, J., & Kisiel, K., Modeling the intelligence analysis process for intelligent user agent development, *Research and Practice in Human Resource Management*, 9(1), (2001), 59-73.

Rashid, A.B., Kausik, A.K., Sunny, A.A.A.H., Bappy, M.H. Artificial Intelligence in the Military: An Overview of the Capabilities, Applications, and Challenges, *International Journal of Intelligent Systems*, art. 676366, (6 November 2023), 31 pages, https://doi.org/10.1155/2023/8676366.

Rastogi, N. and M.K. Trivedi, PESTLE technique–a tool to identify external risks in construction projects, *International Research Journal of Engineering and Technology (IRJET)*, 3(1), 2016, 384-388.

Reddy, R. G., Lee, D., Fung, Y. R., Nguyen, K. D., Zeng, Q., Li, M., ... & Ji, H. *SmartBook: AI-Assisted Situation Report Generation for Intelligence Analysts*, (2023), arXiv preprint arXiv:2303.14337.

Regens, J. L., Augmenting human cognition to enhance strategic, operational, and tactical intelligence, *Intelligence and National Security*, 34(5), (2019), 673-687.

Roy, Denny, The U.S. Military Has a New Strategy to Fight China in a Taiwan War, *The National Interest* (March 2, 2024), https://nationalinterest.org/blog/buzz/us-military-has-new-strategy-fight-chinataiwan-war-209816.

Sufi, F., An innovative GPT-based open-source intelligence using historical cyber incident reports. *Natural Language Processing Journal*, 7, (2024), 100074.

Treverton, Gregory F., *New Tools for Collaboration: The Experience of the U.S. Intelligence Community* (January 2016), https://csis-website-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/publication/160111_Treverton_NewTools_Web.pdf.

Turner, B., Artificial intelligence analysis: What the rise of AI means for human intelligence analysts, *International Journal of Contemporary Intelligence Issues*, 1(1), (2024), 52-65.

Turnley, J. G., and McNamara, L. A., An Ethnographic Study of Culture and Collaborative Technology in the Intelligence Community, *Sandia Technical Report*, SAND 2007-2593-J (January 2016).

U.S. Congress, H.R.6425 - *To direct the Secretary of Defense to establish a working group to develop and coordinate an artificial intelligence initiative among the Five Eyes countries, and for other purposes*, 118th Congress (2023-2024), https://www.congress.gov/bill/118th-congress/house-bill/6425.

Vogel, Kathleen M., Gwendolynne Reid, Christopher Kampe, and Paul Jones, The impact of AI on intelligence analysis: Tackling issues of collaboration, algorithmic transparency, accountability, and management, *Intelligence and National Security*, 36(6), July 2021, 827-848, https://doi.org/10.1080/02684527.2021.1946952.

Convergence in the New Security Environment: Risk Intelligence and the Potential of Generative AI

Katina Michael
Newcastle University, UK

Abstract

For some time there has been a movement away from the traditional view of security as a purely functional activity which occurs within a single department of an agency or enterprise, to security being understood as a value-added capability serving the overall mission of an organization. Enterprise risk management (ERM) is a process that is conducted by private companies for the purpose of due diligence informing key decision makers like chief information officers (CIOs). In the same light, the intelligence cycle is conducted by government organizations for the purpose of maintaining national security and informing policy makers like heads of state, ministers and other agencies tasked with security such as the military. The new security paradigm has spurred on the development of enabling business processes that have not only an enterprise-wide view of risk but an interdependent organization-to-organization view of risk. Entities interconnected in the intelligence community (IC) must now share their information to ensure robustness in their decision-making capabilities. In changing the way things have been done, entities in the new security environment are undergoing the trend of convergence on a number of levels including information, products and services, platforms (i.e., through standardization), and organizations. Of importance in this paper, is the convergence and integration occurring between the risk management and intelligence cycles which has born about the emerging concept of risk intelligence (RI). Generative AI (GenAI) has more recently acted to bring together data, analytics, and AI-related goals, however its potential in the intelligence cycle is yet to be fully realised.

Keywords: Security Environment, Security Convergence, Enterprise Risk Management, Intelligence Cycle, Intelligence Community, Risk Intelligence, Generative AI

1.0 Introduction

This paper argues that convergence is occurring within the security environment, notably in the fields of risk management and intelligence. Commentators are unanimous in their assessment that the security environment is undergoing a steep rate of change in the way the intelligence community functions, some stating that the change is so dramatic that it can even be considered revolutionary. The trend of convergence is prevalent at multiple levels, causing a cultural shift away from a silo and stovepipe mentality towards transparent information sharing (U.S. Government, 1998, p. 28). The paper begins by defining convergence in the new security environment, and broadly outlines the different types of convergence that have been defined in the literature over the last 30 years since the widespread adoption of Internet Protocol (IP)-based technology. A normative description of risk management and intelligence is then provided, showing the basic steps carried out for each by enterprise and government organizations.

The contribution of this paper is in identifying how risk management and intelligence cycles can be integrated through business processes and the benefits ensuing from this integration. Beyond integration, it is predicted in this paper that the risk management and intelligence processes will soon be referred to interchangeably and universally in the literature. The emerging concept of *risk intelligence*, explicitly merging together the domains of *risk management* and *intelligence* is then discussed prior to concluding remarks restating the importance of the trend of convergence in the new security environment. Generative artificial intelligence (GenAI) has acted to propel this trend forward through rapid business transformation focused on various types of information. A discussion is presented on the potential of GenAI to play a catalytic role in risk intelligence but not withstanding some major challenges and risks to predictive capabilities (Dahl and Strachan-Morris, 2024). The research questions being addressed in this paper include: (1) the characterisation of security convergence in all its forms; (2) the definition of risk intelligence as a business process within a national security context; (3) the waves of technological innovations that have acted to impact intelligence communities over the last 20 years; and (4) the potential benefits and shortcomings of adoptive generative artificial intelligence techniques focused more on products than processes.

2.0 Security Convergence

The term *convergence* has its roots in mathematics and the natural sciences dating back to the late sixteenth century (Borodzicz, 2005, p. 13). In its modern interpretation convergence has to do with the evolutionary trends in technological development. The term is therefore now linked to the idea of symbiosis occurring between products or between processes. At an enterprise level, convergence can be observed as individual business units coming together to enhance security for the purpose of creating competitive advantage (Booz, 2005, p. 6). At a state level, convergence can be understood within the context of national security, as agencies that start looking more and more alike come together to engage in collaborative efforts to meet performance criteria, and to ultimately reduce costs by removing duplication and redundancy. ASIS International defines *security convergence* as "the identification of security risks and interdependencies between business functions and processes within the enterprise and the development of managed business process solutions to address those risks and interdependencies" (Booz, 2005, p. 6).

Dating back to 2008, Johnson and Spivey (p. 31) related security with respect to enterprise security risk management (ESRM) and emphasized the combined management of physical and logical security as a "holistic risk management process that aligns organizational drivers affecting strategy, processes, people, technology and knowledge to protect key assets in accordance with governance, risk, and compliance (GRC) requirements". The authors describe that ESRM will only work effectively if multiple management disciplines come together through cross-functional collaboration such as security and safety, legal and risk management, and business continuity (Johnson and Spivey, 2008, p. 31).

2.1 Types of Security Convergence

In discussing convergence, this paper engages the reader at four different levels. Convergence of *security organizations* at the national and enterprise level (Pathak, 2005, p. 569). This type of convergence includes companies that are coming together to offer solutions to the intelligence community, as well as convergence of government agencies that would work more effectively together than as stand-alone organizations. Convergence of *security processes* (i.e. standards/ platforms). This involves the identification "of security risks and interdependencies between business functions and processes within the enterprise and the development of managed business process solutions to address those risks and interdependencies" (Peterson, 2006, p. 3). Convergence of *security products and services*, fundamentally involving "different companies' people and IT systems working together to deliver a convergent product or service" (Layton, 2008, p. 2). Finally, convergence of *information*, i.e., sources and quality content that is used to inform business processes, including technical, human, open source intelligence (Peterson, 2006, p. 3).

The model of convergence has been said to be ideal for "managing uncorrelated… risk through a systematic, coordinated process" (Williams, 1999, p. 14). However, the complexity of convergence in reality- transforming dozens of agencies “that can plan manage and carry out operations effectively”- should not be understated (Wrightson and Caldwell, 2005, p. 8). Taking policies and processes that were once created in silos and trying to make some collective sense out of them to institute change, is multifaceted and complicated (Podowitz and Tretick, 2008, p. 1). While at the enterprise level convergence is driven by compliance, government agencies did not act toward convergence, until they were subject to major intelligence failures.

2.2 Security as a Value-Add Capability

The premise for the convergence phenomenon sweeping the global security industry has been a shift in mindset that sees security as a “value add” to the overall mission of businesses and government agencies alike (Booz, 2005, p. 4). It is the realization that security cannot be achieved alone, but requires a meshed network of stakeholders and entities to work together towards a common goal. More than any other event in recent U.S. history, September 11 (2001) showed the failure of intelligence agencies in sharing information regarding possible terrorist targets. For instance, an inquiry into the actions of the Federal Bureau of Investigation (FBI) some twenty years ago, concluded that the main problems were: severely inadequate information and communication technology (ICT) systems, an inability to bridge together human intelligence (HUMINT) and technical intelligence (TECHINT) to conduct all source analysis, and problems related to the recruitment and training of analysts (Gill, 2004, p. 475).

Apart from asymmetric terrorist strikes that have caused significant loss of life post September 11, imperatives towards convergence in the security environment have come from enabling high technologies that have blurred traditional functional boundaries, new compliance and regulatory regimes, and the emphasis today on information-based assets (i.e. as opposed to physical items) (Booz, 2005, p. 8). Security convergence has meant change in the context of:

- people and their respective roles and responsibilities;
- processes in terms of standards to follow and regulations; and
- technology in terms of enabling tools and applications.

Despite the trend toward security convergence among the intelligence community, nothing has served to propel change as much as GenAI, that has impacted the role of the individual intelligence analyst. As Usher et al., (2024) reflected: “the [Large Language Models] LLMs available within the next three years will probably far surpass the capabilities of systems we use today and will be able to solve complex problems, take action to collect and sort data, and deliver well-reasoned assessments at scale and at speed.”

2.3 The End-to-End Security Lifecycle

The motivation behind convergence in the security environment is one that espouses a whole-of-life, holistic, highly collaborative exchange between organizations and agencies in order to “better make decisions to protect themselves from risk” (Banham, 1995, p. 22). It is a movement away from the silo functional organizational security view which treated the areas of prevention, detection, response and recovery separately, toward a view which espouses the entire end-to-end security lifecycle as a super-system (Michael et al., 2025). The challenge with such a system is getting organizations and agencies who have thought and acted a particular way for decades, to change their ways and to begin working closer together in order to solve problems (Booz, 2005, p. 6).

The new security environment is characterized by strategic changes, changes to processes, and changes to the roles and responsibilities of people in security organizations. Security in the 1980s predominantly had to do with a simplistic conception of physical access control and the movement of assets. The function and application of the security industry began to shift with the rise of the digital era, which immediately incorporated a much larger range of risks (Borodzicz, 2005, p. 68). The nine traditional operating levers can be adapted to help organizations perform better in the new converged security environment. The levers that can be applied with respect to internal and external drivers include: risk management, governance, budget processes, standards and guidelines, integration, business case, roles and responsibilities, leadership and knowledge of business.

3.0 The Risk Management Process

This section focuses on the relationship between security and risk management. The narrow definition of security is challenged, away from a single organisational context to a multi-tiered / multi-stakeholder context of risk (Clarke and Michael, 2024).

3.1 Security = Risk Management

Till now this paper has focused on the notion of convergence. In this section the risk management domain is explored within the context of the new security environment. Merkow and Breithaupt (2006, p. 27) explicated that "security equals risk management". They agree with Borodzicz (2005, p. 50) who wrote: "security can be seen as risk management in practice". To begin with risk is defined, as a unified language (from historical, psychological, sociological, functionalist, management, normative, structural, and descriptive perspectives), that is presently missing from the domain (Borodzicz, 2005, pp. 52-55). This is vitally important as often different fields of study claim to be the "owners" of risk management (e.g. information technology and insurance) when quite oppositely, risk is enterprise-wide and industry agnostic (Dhillon, 2007, p. 157; Miccolis, 1996, p. 46). Where there are security issues of any type, then risk management practices should be instituted. Traditionally risk was only considered to be about physical assets- "the potential that a given threat will exploit vulnerabilities of an asset or group of assets and thereby cause harm, to the organization" (ENISA, 2008, p. 4). Today, however, the business of risk has changed (Menzies et al., 2024). Risk management is now more about the organization's strategic-level initiatives which encompass both physical and logical assets (Slay and Koronios, 2006, p. 2; Institute of Risk Management, 2002, p. 2). For this reason, enterprise risk management (ERM) is about "bringing business functions (e.g. finance, line management, R&D, human resources) closer together to build a common risk-based framework for better decision making…" (Miccolis, 1996, p. 48).

3.2 The Process of Risk Management

No matter what risk analysis process is used the standard method remains the same (Peltier, 2001, p. 5). Risk management is composed of three main parts: risk assessment, risk mitigation, and risk evaluation (Dhillon, 2007, pp. 155-170). Will Ozier (2001, p. 224) defines risk management as the process "of identifying risks, risk-mitigating measures, the budgetary effect of implementing decisions related to the acceptance, avoidance, or transfer of risk… [it also] includes the process of assigning priority to, budgeting, implementing, and maintaining appropriate risk-mitigating measures in a continuous or periodic cycle of … management.” While many versions of the risk management cycle are available from diverse sources-international bodies like the OECD (Organization for Economic Co-operation and Development) and the ISO (International Standards Organization), national standards bodies, government agencies, industry-specific bodies and even single organizations - the cycles all encompass the same broad steps (Clarke and Michael, 2024). According to Peltier (2001, pp. 17-19), these steps are:

1. Assess risk and determine needs;
2. Implement appropriate policies and related controls;
3. Promote awareness; and
4. Monitor and evaluate policy and control effectiveness.

The latest international standard in risk management is ISO 31000:2018 (ISO, 2018).

The heart of any risk management process is a risk assessment (Clarke and Michael, 2024). Typically, a risk assessment begins with identifying risks. Risks are usually categorized into different types to make assessment more meaningful. A method is then formulated to prioritize risks which typically include both quantitative and qualitative data, and may take the form of a risk score and/ or mapping exercise. A critical risk analysis is then conducted to evaluate risk-loss/risk-return values modelled against performance indicators. The risk model is then implemented and strategies are recommended to mitigate losses (Foley & Lardner, 2007, p. 2). It is important to emphasize that risk is everybody’s business. A risk assessment is considered robust if it covers a range of issues- technological, human factors, policies, third party, etc. Dhillon (2007, p. 235) claims rightly that “since most systems are interconnected and interdependent, any risk assessment should also consider threats that might originate elsewhere”.

3.3 What does Risk Management have to do with National Security?

It has already been established that risk management and enterprise security go hand-in-hand. But a question that can be legitimately posed is whether or not risk management has any relevancy to national security? While it is typical to think of risk management in areas like insurance and finance, it is atypical to relate risk management to domestic terrorism. And yet, the risk management process was embraced by the U.S. Congress, post September 11, in order to strengthen against future terrorist strikes. In the context of national security then, risk management can be defined as a "strategy for helping policymakers make decisions about assessing risks, allocating resources, and taking actions under conditions of uncertainty" (Wrightson and Caldwell, 2005, p. 8). In the following section we investigate first the intelligence cycle, and then the likeness of the intelligence cycle to the risk management process. It is thus proposed that the intelligence cycle and risk management are converging domains and that before too long, the processes will be used interchangeably.

4.0 Intelligence Cycle

The common misconception is made, that the intelligence cycle is strictly something conducted by tactical military analysts. However, it is well-known that practitioners in industry widely practice intelligence-related activities for a variety of reasons, including for the purpose of competitive business intelligence (BI). Unlike the risk management process which has undergone a great deal of standardization due to compliance and other globalization factors, the intelligence cycle has remained a fairly generic framework that organizations can choose to follow completely or partially. On the national security and defence side, however, intelligence as a process is continually being improved upon, especially to combat future asymmetric attacks.

4.1 Defining the Intelligence Cycle

The intelligence cycle (Johnson, 1986) can be defined as "the process by which information and data is collected, evaluated, stored, analyzed, and then produced or placed in some form for dissemination to the intelligence consumer for use. The cycle consists of: consumer, collector, evaluation, analysis, production, dissemination, consumption, consumer" (United States Government Accountability Office, 1998, p. 27). The New Zealand Qualifications Authority (NZQA, 2003) define intelligence as the collective "functions, activities, and/or organizations which are involved in the process of planning, gathering and analyzing information of potential value to decision makers, and to the production of intelligence". The goal of intelligence is to "produce guidance based on available information within a time frame that allows for purposeful action" (Willis, 2007, p. 3). The main phases carried out in a typical intelligence cycle; the distinct phases have remained relatively unchanged in modern times, save for the addition of the initial "requirements" phase, enabling policy makers to make a request for information (RFI) (Directorate of Intelligence, 2008). This phase helps analysts to plan and better direct the intelligence effort. Data is then collected, processed, analyzed and disseminated to the appropriate stakeholders. The U.S. military have developed a sophisticated "Intelligence Process Model" (IPM) that helps analysts to work through RFIs and also for decision-makers to track the status of their request(s) (Miller, 2008, p. 5).

4.2 The Phases of the Intelligence Cycle

The request for information is where information needs are identified by policy makers (Canadian Intelligence Security Service, 2004, p. 2). In the planning and direction phase, resources are identified and care is taken to balance the level of intrusiveness of the request with what is legally permissible (Canadian Intelligence Security Service, 2004, p. 2). The collection phase follows and is where the raw information is gathered. Information comes from varied sources- it may be public, foreign or illegally intercepted via satellite or other communication technology. Examples of collection assets include satellites, surveillance equipment, even consumer edge devices, providing time and date stamps, and increasingly location coordinates (Abbas et al., 2011). According to Miller (2008, p. 4), the "means and methods of collection are highly dependent on the source of the information and these sources are generally categorized into various intelligence disciplines." These sources are combined with open source intelligence (OSINT) including newspapers, periodicals, foreign and domestic broadcasts (e.g., CNN, BBC, Aljazeera.net) and official documents (e.g., Commonwealth inquiries).

The collected data is then processed and made into a form that is usable by analysts. This is often where the most errors creep into the process, as different sources of data are brought together. Maintaining quality in the data sets being processed is of paramount importance. Some referred to this processing 'melting pot' as the "fusion centre" (United States Government Accountability Office, 1998, p. 27). It can be postulated that the ultimate fusion centre has now been given a front-end interface through Generative Pre-Trained Transformer (GPT) models that rely on large language models. In the traditional processing phase, it was here where linkages were made between the structured and unstructured data that might be the difference between good and bad intelligence. Given the rise of GenAI, it has become increasingly difficult to determine the overall quality of the output, given that sources are not explicitly referenced. In the analysis and production phase, fused data is prepared to make intelligence products which are usually categorized by their primary use, for example, indications and warning and counterintelligence (Miller, 2008, p. 4).

According to NZQA (2003, p. 2): "[a]nalysis refers to a process in the production step of the intelligence cycle in which intelligence information is subjected to systematic examination in order to identify significant facts and derive conclusions. The 'raw intelligence' collected, whether by human or technical means, is frequently fragmentary and at times contradictory. Through analysis a sorting, evaluating and interpreting of the various pieces of data occurs including an interpretation of meaning and associated significance." Common analyses performed in these products include association, temporal and spatial charting; and link, financial, content and correlation analysis (United States Government Accountability Office, 1998, p. 27). These analyses are still relevant in the context of GenAI and can be semi- or fully-automated taking significantly less time to produce the products. The dissemination phase of the intelligence output can happen in two ways: (i) delivered to the consumer who requested it in a push action; (ii) or stored to be pulled at a later date (Miller, 2008, p. 4).

5.0 Integrating Risk Analysis into the Intelligence Cycle

In this section the symbiosis between risk analysis and the intelligence cycle is explored.

5.1 Is Risk Management and the Intelligence Cycle Linear?

Now that a brief overview of the risk management process and the intelligence cycle have been presented, let us examine the premise that both processes are not linear but network-centric, meshed, and highly collaborative. This does not mean that the actual steps or phases are contested in each process- but the manner in which stakeholders interact with one another is brought into question. The move is revolutionary and towards a network-centric collaboration process using a target-centric approach to interlink stakeholders and information (Barger 2005, p. 20; Clark, 2004, pp. 17-18). In the new security environment convergence is acting to bring stakeholders (e.g., collectors, processors, analysts, policy makers) together to communicate through a centralized means to make decentralized decisions (Clark, 2004, pp. 17-18). This does not mean that hierarchy is abandoned altogether in the intelligence community but that stakeholders can make use of technologies which allow for a more agile working environment.

The National Infrastructure Protection Plan (NIPP) in the United States presents a context for information sharing amongst the primary stakeholders. It does not mean that the new environment contains members belonging to “one large happy family”, as each organization still differs in their mission and goals (Lahneman, 2006, p. 3). Importantly, “The NIPP information-sharing approach constitutes a shift from a strictly hierarchical to a networked model, allowing distribution and access to information both vertically and horizontally, as well as the ability to enable decentralized decision making and actions” (Homeland Security, n.d., 2)

5.2 Risk Management Based Intelligence (RMBI)

If real-time collaboration is a result of the new security environment, and private and public members of the intelligence community are sharing data (i.e., contributing and retrieving data), then it follows that processes too can be integrated. Willis (2007, p. 3f) states that risk analysis can be used to sharpen intelligence products and to prioritize resources for gathering intelligence. He goes on to explain that “risk analysis can be a tool that can help intelligence practitioners sharpen their conclusions by providing analytic support for identification of scenarios of greatest concern” (Willis, 2007, p. 15). It must be noted however, while risk analysis enhances intelligence, it still remains mere intelligence and far from fool proof. The same applies in the adoption of GenAI tools that are used to generate intelligence- they are not fool proof (Vogel et. al., 2024).

The integration of the risk management process and the intelligence cycle has been referred to as *risk management based intelligence* (RMBI) (Ylönen and Aven, 2023).

RMBI is defined by the United States Government Accountability Office (1998, p. 28), as: “an approach to intelligence analysis that has as its object the calculation of the risk attributable to a threat source … a means of providing strategic intelligence for planning and policy making especially regarding vulnerabilities and countermeasures designed to prevent criminal acts; a means of providing tactical or operational intelligence in support of operations against a specific threat source, capability or modality; can be quantitative if a proper data base exists to measure likelihood, impact and calculate risk; can be qualitative, subjective and still deliver a reasonably reliable ranking of risk for resource allocation and other decision making in strategic planning and for operations in tactical situations.” This approach is about understanding intelligence in the context of a broad strategic approach and not a response to a single case-based RFI request. It can be denoted from the full-length RMBI description that risk management is clearly integrated in the modern intelligence cycle; from this integration stems an even closer relationship which we can refer to here as symbiosis, that is, the trend of convergence at multiple levels including organisational, process, product, and information (Anderson, 2007, p. 7). It is perhaps the latter convergence trend, that of information convergence, that has propelled the cultural shift in the intelligence community at large; now described as *big data* (Michael and Miller, 2013).

When analysts from different organizations (public or private) begin to rely on the same information sources, and are able themselves to contribute information to such causes as critical infrastructure protection (CIP), then opportunities for convergence in products, processes, and organisations emerge. As Peterson (2006, p. 1) emphasises, “… the world is converging around the value of information, not that information is converging around or into something else. Instead, information is the new central actor, defining the enterprise organization and its business. On one hand, information is power and a competitive weapon. In this sense, information is the chief asset of the business. Yet, on the other hand, information is also the chief risk. It is a legal and security liability... In the end, it is this paradox that is the catalyst for change; change which is transforming the Information-Centric Enterprise.” Robinson (2007, p. 4) described the “… creation of a common data structure for risk and control processes and a common technology architecture supporting this effort. This common ground not only enables the Risk/Control functions to speak a single language, it also fosters communication, greater coordination, and increased understanding." And yet, the proliferation of disinformation and misinformation through deepfake technology has become a significant liability to this end-to-end process; any organisation is only as good as the quality of its data.

5.3 Risk Intelligence as a Business Process

In restating Sherman Kent's classic definition of intelligence as a kind of *knowledge*, then *information convergence* can be considered as enabling business processes between members of the intelligence community (Rathmell, 2002, p. 88f). In the corporate world, the recognition that knowledge equated to power became prevalent in the 1990s. Organizations were quite aware that there was 'too much information, and too little knowledge'. It was at the turn of the millennium that ICT solutions also became available to solve the problem of islands of information through electronic resource planning systems (ERP), many of which contained a business intelligence module to go beyond data warehousing. As Gill (2004, p. 476) pointed out, "the construction of ever-larger databases, data warehousing and data-mining, though of great significance in intelligence, cannot 'solve' intelligence problems without a process of targeting, careful evaluation of information and human analytical skills." Despite the advances in big data, and commensurately the rise of data analytics, the need for human-centred capabilities are necessary to ensure a human is not out-of-the-loop when AI is being utilised (Michael et al. 2024; Schoenherr et al. 2023).

It should be no surprise to us then, that today *risk intelligence* (RI) has emerged as a completely new business process (Azvine et al., 2007, p. 155). Two consulting companies, Deloitte (2025) and KPMG (2025), have already begun to market a RI framework. Robinson (2007, p. 4), wrote: "Many organizations are now looking at convergence models to integrate risk and control processes and create a common framework for assessing and monitoring the organization's risks." Risk intelligence enterprises are those organizations that are characterized by their future vision, ability to bridge silos and speak a common language, conduct impact assessments, weigh up the vulnerabilities, allocate resources appropriately, act with a risk conscious spirit, and even pursue risk for the purposes of higher rewards (Layton, 2008, p. 2). The risk intelligent chief information officer (CIO) is someone who practices risk intelligence (Dittmar and Kobel, 2008, p. 42). And just like any other framework or process, there are differing levels of sophistication that can be attained (Houser and Conlin, 2006, p. 6). GenAI has acted to seemingly grant individual intelligence analysts new-found skills but the attention is quickly shifting from the product to the process the analyst uses through prompts, and the accuracy and validity of outputs through certification (Abbass et al., 2024).

5.4 Problems Associated with the Risk Intelligence Process

A number of problems plague the intelligence community in the new security environment. It does not mean that risk intelligence will not work, but governments need to understand that these challenges are not trivial, and attempt to combat them with longer-term initiatives. Even if we take the naïve view that implementing convergence is 'easy', we still require competent analysts who understand the data and can deal with the increasing complexity of technical products, particularly in the era of AI (Usher et al. 2024). Pre-GenAI, Lahneman's (2006, p. 3) assessment was "…if current practices continue[d], the intelligence community (IC) of 2020 will experience an imbalance between the demand for effective overall intelligence analysis and the outputs of the individually-oriented elements and outlooks of its various analytic communities". It is to say that this "imbalance" has been exacerbated through the introduction of GenAI.

For many, the answer lies in professionalizing the security-risk industry. Training programs for analysts by a single accreditation organization is widely recommended. Providing intelligence in a timely manner is another issue, alongside the capability to simplify the information being gathered so it is meaningful and can be applied into action by decision makers (Azvine, 2007, p. 155). In addition, what kind of data will reside in the intelligence system for the conduct of all-source analysis by organizations should not be forgotten as a key challenge- after all garbage in/garbage out (GIGO): "Good data often leads to visionary and profitable decision making. Poor data quality is often the cause of bad strategic decisions and inaccurate financial and management reporting" (Azvine, 2007, p. 160).

Perhaps the biggest challenge at hand however, is governance- how do you bring the intelligence community together within an integrated culture, break down the barrier of secrecy, and still maintain limits to information accessibility based on RFIs.

Lahneman (2006, p. 10) is scathing in his assessment of the U.S. context: "The U.S. intelligence community is the "Community that Isn't." It is a series of nearly autonomous organizations, each with its own way of doing business. The analytic portion of the IC reflects the fragmentation of the overall intelligence enterprise. Such a fragmented approach is at odds with the need for greater knowledge sharing to enable effective analysis of dispersed threats and other issues." Trust in people and systems, therefore, along with enforceable policies and procedures will be paramount in this emerging environment. While much has changed since Sept 11, and great reforms undergone in the global intelligence community, big data, analytics (Coulthart et al. 2023), and AI (Vogel, 2021), have acted to create a whole new set of questions with respect to risk intelligence.

Coordination between analysts in intelligence agencies and organisations, and coordination with third party providers offering public/open and even private data is paramount. Without agency and organisational policies to govern risk intelligence processes, the pacing problem may well exacerbate accuracy and effectiveness issues (Michael et al., 2024). Rather than *too much information and too little knowledge*, the stage is set for *too many informational products and no confidence in the knowledge outcomes* that are used to make risk-related decision affecting nation states, and the world at large. Moreover, as new technologies and techniques become available cognitive overload and adequate testing may be compromised. Some may argue this is an age-old problem and will breed even greater innovation within the intelligence community, but the efficacy of these activities cannot be guaranteed nor measured in the short term. Additionally, what is different about the current landscape are the capabilities being offered by the new technologies, allowing for some level of autonomy in decision-making based on various AI approaches. In the next section we explore this in relation to generative AI (GenAI).

6.0 GenAI: The Ultimate Technology Convergence Catalyst?

There is no doubt that technology is that catalyst in the context of transforming intelligence communities, predominantly through AI and machine learning (ML), and now very much GentAI (SCSP, 2024). GenAI is placing pressure on IC agencies to adopt the technology at every part of the risk intelligence cycle, and to continue to adopt it in a layered fashion building on each breakthrough made possible (SCSP, 2024). But with this acceleration in adoption, in process development, in creating a talent pool that can harness the power of GenAI, are significant opportunities but also commensurate risks. With every strategic advantage, comes the reality, that competing institutions, and enemy states, are able to replicate the same approach, but even worse, perhaps flood the open source intelligence, and even other sources of intelligence with disinformation to undermine the whole risk intelligence process (SCSP, 2024). This has been described as the AI paradox (Michael et al., 2023).

Currently GenAI is commonly being used for (1) summarisation and translation; (2) data processing and synthesis; (3) predictive analysis; and (4) report validation, accuracy and bias determination. But analysts will soon be utilising GenAI for much more; even creating risk intelligence-based business processes that can conduct automatic decisions-making (ADM) through a variety of information sources, without human intervention. It is at this point that the risk of risk-based information products may far outweigh the value of the insights and the impact they may be having on society at large. In an era of disinformation campaigs by enemy states, relying on data whose quality is unknown, even if the finest analysers are used to detect bot behaviour, is fraught with danger. As systems are newly built with infrastructural implications, layering decisions upon decisions of variable quality will lead to even greater error, despite that more data used in a problem is allegedly supposed to diminish the impacts of dirty data.

Thus, while we can point to AI as that which is bringing a new capability to risk intelligence, convergence has been a centrepiece of intelligence in every domain and field since digitalisation- the integration of digital technologies and big data toward analytics, demanding interoperability. Mueller (2011) described this as a “new paradigm” that required “integration of intelligence and law enforcement capabilities” in addition to “augmenting relationships and information sharing” through the “development of information technology systems” with the IC, toward a “continuous intelligence cycle that drives investigative strategies to ensure resources are targeting the most pressing threats”. More broadly, this capability was known as a “strategy management system.”

Over the last 20 years, six main trends in the intelligence community can be observed:

1. The big data revolution- focused on gathering data and information sharing through all forms, structured and unstructured, quantitative and qualitative (Michael and Miller, 2013; Vogel, 2021);
2. Professionalisation of the role of the intelligence analyst- focused on agents who can "quickly and effectively review, analyze, and disseminate the intelligence collected in the field" (Mueller, 2011);
3. The growth of the field of risk intelligence across organisations (private, public, governmental and in the third sector)- focused on managing the ever-increasing large volume of data available for analysis (text, image, audio, video) stemming from sensor-generated surveillance technology and new ways in which data analytics could assist the sector (US DOD, 2023);
4. A commitment to adopting artificial intelligence in a responsible manner as a strategic advantage- focused on denoting the authenticity, accuracy, validity of data (US DOD, 2023);
5. Acknowledgement that GenAI has a role to play in risk intelligence but is still in its nascent stage of development, particularly when it comes to predictive models and the data that is used to pre-train large language models, e.g., the risk of deepfakes (Bajak, 2024).
6. The potential for human-machine teaming for decision-making- focused on ensuring human oversight in data-driven/GenAI-driven judgements requiring greater governance (SCSP, 2024).

All in all, this will act to advance the data, analytics and AI ecosystem at large which brings together stakeholders from across the information supply chain (Clark, 2023).

7.0 Conclusion

The overarching benefit of convergence in maintaining national security is strategic, i.e., keeping one step ahead of the enemy to prevent terrorist attacks in order to minimize the element of surprise. Convergence has the ability to make a reduction in overhead and duplication and to streamline once separate security groups and organizations. Today convergence is about remaining successful; and more than that it is about giving life to new opportunities and emergent benefits that cannot be achieved individually. At the moment the trend towards a unified security program seems to be about reducing risks and increasing control through quality intelligence. However, one could also be critical of the security industry at large and point out, that the trend towards a 'super' converged system is destined to failure because monolithic systems are subject to singularities, and could create more complications than answers. Some may even say the effort towards convergence is a waste of money, time and energy because anti-terrorism capabilities are a fallacy. Is the technology available today propelling us all toward a future environment that may create even more problems for us as a society? Time will tell.

References

Abbas, R., Michael, K., Michael, M.G., Aloudat, A. (2011) *Emerging forms of covert surveillance using GPS-enabled devices*, Journal of Cases on Information Technology, 13(2) pp. 19-33.

Abbass, H., Michael, K., Vogel, K.M. (2024) *Swarm Metaverse: Understanding Socio-Technical Innovation and Trust Before Certification*, ADSTAR: Australian Defence Science, Technology and Research, 18 September 2024 <https://www.adstarsummit.com.au/ > Accessed: 3 January 2025.

Anderson, K. (2007) *Convergence: A holistic approach to risk management*, Network Security, pp. 4-7.

Azvine, B., Cui, Z., Majeed, B. and Spott, M. (2007) *Operational Risk Management with Real-Time Business Intelligence*, BT Technology Journal, 25(1) pp. 154–167.

Banham, R. (1995) *The convergence of risk*, Risk Management, 42(7), p. 22.

Barger, D.G. (2005) *Toward a Revolution in Intelligence Affairs* <http://www.rand.org/pubs/technical_reports/2005/RAND_TR242.pdf> Accessed: 2 February 2008.

Allen Booz, (8 November 2005) *Convergence of Enterprise Security Organizations*, The Alliance for Enterprise Security Risk Management <www.asisonline.org/newsroom/alliance.pdf> Accessed: 1 May 2008.

Bajak, F. (24 May 2024) *U.S. intelligence agencies' embrace of generative AI is at once wary and urgent*, PBS, <https://www.pbs.org/newshour/world/u-s-intelligence-agencies-embrace-of-generative-ai-is-at-once-wary-and-urgent> Accessed: 3 January 2025.

Borodzicz, E.P. (2005) *Risk, Crisis and Security Management*, New York, Wiley.

Canadian Intelligence Security Service (2004) *Backgrounder No. 3: CSIS and the Security Intelligence Cycle* <http://www.csis-scrs.gc.ca/en/newsroom/backgrounders/backgrounder03.asp> Accessed: 9 March 2008.

Clark, J. (2023) *DOD Releases AI Adoption Strategy*, U.S. Department of Defense, <https://www.defense.gov/News/News-Stories/Article/Article/3578219/dod-releases-ai-adoption-strategy/> Accessed: 3 January 2024.

Clarke, R.A., and Michael, K. (2024) *Multi-Stakeholder Risk Assessment of Socio-Technical Interventions*, Australasian Conference on Information Systems, University of Canberra, ACT, Australia.

Clark, R.M. (2004) *Intelligence Analysis: A Target-centric Approach*, CQ Press.

Coulthart, S., Hossain, M. S., Sumrall, J., Kampe, C., & Vogel, K. M. (2024). *Data-science literacy for future security and intelligence professionals*. In Strategic Minds (pp. 40-60). Routledge.

Dahl, E. J., & Strachan-Morris, D. (2024) *Predictive intelligence for tomorrow's threats': is predictive intelligence possible?* Journal of Policing, Intelligence and Counter Terrorism, 19(4) pp. 423–435 https://doi.org/10.1080/18335330.2024.2404834.

Deloitte (2025) *Risk Intelligence*, <https://www2.deloitte.com/us/en/pages/risk/solutions/risk-intelligence.html> Accessed: 3 January 2025.

Dhillon, G. (2007) *Information Systems Security: Text and Cases*, Prospect Press, North Carolina.

Dittmar, L. and Kobel, B. (2008) *The Risk Intelligent CIO*, Risk Management, 55(3) p. 42.

Directorate of Intelligence (2008) *The Intelligence Cycle*, Federal Bureau of Investigations <http://www.fbi.gov/intelligence/di_cycle.htm> Accessed: 27 April 2008.

Enisa (2008) *Glossary of Risk Management,* ENISA: A European Union Agency, <http://www.enisa.europa.eu/rmra/glossary.html> Accessed: 27 April 2008.

Foley & Lardner LLP (2007) *Enterprise Risk Management - Risk Intelligence and Anti-Fraud Controls*, National Director's Institute, Accessed: 27 April 2008.

Gill, P. (2004) *Intelligence and the Post 9/11 Shift*, Intelligence and National Security, 19(3) pp. 467–489.

Homeland Security (n.d.) *National Infrastructure Protection Plan Information Sharing*, U.S. Homeland Security, https://www.dhs.gov/xlibrary/assets/NIPP_InfoSharing.pdf, Accessed: 23 April 2008.

Houser, N. and Conlin, S. (2006) *Creating Risk Intelligence: A High Level "How To" Guide for Program Managers*, Deloitte <management.energy.gov/06W_RMconHou.ppt> Accessed: 27 April 2008.

ISO (2018), *Risk management — Guidelines*, https://www.iso.org/standard/65694.html Accessed: 20 December 2008.

Johnson, L.K. (1986) *Making the "Intelligence" Cycle Work*, International Journal of Intelligence and Counter-Intelligence, 1(4) pp. 1-23.

Johnson, M.P. and Spivey, J.M. (2008) *ERM and the Security Profession*, Risk Management, 55(1) pp. 30-32, 34-35.

KPMG (2025). KPMG Risk Intelligence: transform your risk management, <https://kpmg.com/us/en/risk-intelligence.html> Accessed: 3 January 2025.

Institute of Risk Management, (2002), *IRM: A Risk Management Standard*, AIRMIC <www.theirm.org/publications/documents/Risk_Management_Standard_030820.pdf> Accessed: 27 April 2008.

Lahneman, W.J. (2006) *The Future of Intelligence Analysis: Volume I, Final Report* Center for International and Security Studies at Maryland <http://www.cissm.umd.edu/papers/files/future_intel_analysis_final_report1.pdf> Accessed: 27 March 2008.

Layton, M. (2008) *Urgent Convergence: Fostering Risk Intelligence in the Technology*, Media & Telecommunications Industries, Deloitte <www.deloittte.com/RiskIntelligence> Available: 27 April 2008.

Menzies, J., Sabert, B., Hassan, R., and Mensah, K. (2024) *Artificial intelligence for international business: Its use, challenges, and suggestions for future research* *and practice*, Thunderbird International Business Review, 66(2) pp. 185-200, https://doi.org/10.1002/tie.22370.

Merkow, M. and Breithaupt, J. (2006) *Information Security Principles and Practice* Sydney, Pearson.

Miccolis, J.A. (1996) *Towards a Universal Language of Risk*, Risk Management, 43(7) p. 46.

Michael, K., and Miller, K. (2013) *Big Data: New Opportunities and New Challenges*, Computer, 46(6) pp. 22-24 doi: 10.1109/MC.2013.196.

Michael, K., Abbas, R., Roussos, G. (2023) *AI in Cybersecurity: The Paradox*, IEEE Transactions on Technology and Society, 4(2) pp. 104-109, June 2023, doi: 10.1109/TTS.2023.3280109.

Michael, K., Schoenherr, J.R., Vogel, K.M. (2024) *Failures in the Loop: Human Leadership in AI-Based Decision-Making*, IEEE Transactions on Technology and Society, 5(1) pp. 2-13, March 2024, doi: 10.1109/TTS.2024.3378587.

Michael, K., Vogel, K.M., Pitt, J., and Zafeirakopoulos, M. (2025) Artificial Intelligence in Cybersecurity: A Socio-Technical Framing, IEEE Transactions on Technology and Society, 6(1) pp. 15-30 doi: 10.1109/TTS.2024.3460740.

Miller, J.O. (2008) *Modeling the U.S. Military Intelligence Process*, Department of Defense <www.dodccrp.org/events/9th_ICCRTS/CD/papers/044.pdf> Accessed: 27 April 2008.

Mueller, R.S. (6 October 2011), Statement Before the House Permanent Select Committee on Intelligence, Federal Bureau of Investigation <https://archives.fbi.gov/archives/news/testimony/the-state-of-intelligence-reform-10-years-after-911> Accessed 3 January 2025.

New Zealand Qualifications Authority (2003) *Intelligence Analysis: Demonstrate knowledge of the intelligence analysis process*, New Zealand Government <www.nzqa.govt.nz/nqfdocs/units/doc/18503.doc> Accessed 23 December 2024.

Ozier, W. (2001) *Risk Assessment and Management* in Thomas R. Peltier (ed), Information Security Risk Analysis, New York, Auerbach Publications.

Pathak, J. (2005) *Risk management, internal controls and organizational vulnerabilities*, Managerial Auditing Journal, 20(6) pp. 569-577.

Peltier, T.R. (2001) *Information Security Risk Analysis*, New York, Auerbach Publications, https://doi.org/10.1201/b12444.

Peterson, M. (2006) *Information Convergence, Transforming the Information-Centric Enterprise*, SNIA Data Management Forum <www.sresearch.com/articles/SRC-DMF-Article_Information-Convergence_20060112.pdf > Accessed: 27 April 2008.

Podowitz, M. and Tretick, B. (8 January 2008) *Compliance, Convergence and How IT Fits*, CIO <http://www.cio.com/article/print/170000> Accessed: 27 April 2008.

Rathmell, A. (2002) *Towards Postmodern Intelligence*, Intelligence and National Security, 17(3) pp. 87-104.

Robinson, J. (2007) *Risk Convergence: Future State*, Ernst & Young Consulting <http://www.ey.com> Accessed: 27 April 2008.

Schoenherr, J.R., Abbas, R., Michael, K., Rivas, P., Anderson, T.D., (2023) *Designing AI Using a Human-Centered Approach: Explainability and Accuracy Toward Trustworthiness*, IEEE Transactions on Technology and Society, 4(1) pp. 9-23, March 2023, doi: 10.1109/TTS.2023.3257627.

SCSP (April 2024) *Intelligence Innovation: Repositioning for Future Technology Competition*, Special Competitive Studies Project, <https://www.scsp.ai/wp-content/uploads/2024/04/Intelligence-Innovation.pdf> Accessed: 3 January 2025.

Slay, J. and Koronios, A. (2006) *Information Technology and Risk Management*, Wiley.

United States Government Accountability Office (1998) *Information Security Management*, U.S. Government, pp. 1-69.

US DOD, (10 August 2023) *DOD Announces Establishment of Generative AI Task Force*, U.S. Department of Defense, <https://www.defense.gov/News/Releases/Release/Article/3489803/dod-announces-establishment-of-generative-ai-task-force/> Accessed: 3 January 2025.

Usher, W., Caples, A., Kurata, K., Balakrishnan, N. (3 September 2024) *The future of intelligence analysis: US-Australia project on AI and human machine teaming*, ASPI: Australian Strategic Policy Institute, <https://www.aspi.org.au/report/future-intelligence-analysis-us-australia-project-ai-and-human-machine-teaming>, Accessed: 23 October 2024.

Vogel, K.M. (2021) *Big Data, AI, Platforms, and the Future of the U.S. Intelligence Workforce: A Research Agenda*, IEEE Technology and Society Magazine, 40(3) pp. 84-92, Sept. 2021, doi: 10.1109/MTS.2021.3104384.

Vogel, K.M., Abbass, H., Michael, K. (2024) *The impact of AI on intelligence analysis: tackling issues of collaboration, algorithmic transparency, accountability, and management*, ADSTAR: Australian Defence Science, Technology and Research, 19 September 2024 <https://www.adstarsummit.com.au/> Accessed: 3 January 2025.

Williams, T.L. (1999) *Convergence*, Risk Management, 46(8) pp. 13-14.

Willis, H.H. (2007) *Using Risk Analysis to Inform Intelligence Analysis*, RAND Corporation <https://www.rand.org/pubs/working_papers/WR464.html> Accessed: 3 January 2025. <http://www.rand.org/pubs/working_papers/2007/RAND_WR464.pdf> Accessed: 7 February 2008.

Wrightson, M.T. and Caldwell, S.L. (2005) *Risk Management*, United States Government Accountability Office.

Ylönen, M., & Aven, T. (2023) *A new perspective for the integration of intelligence and risk management in a customs and border control context*, Journal of Risk Research, 26(4) pp. 433–449 https://doi.org/10.1080/13669877.2023.2176912.

Author Index

Keyword Index