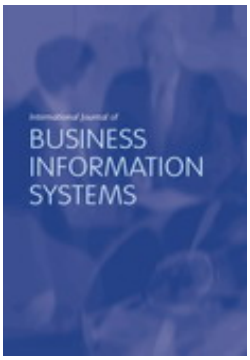
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Business Information Systems

ISSN online: 1746-0980 - ISSN print: 1746-0972
https://www.inderscience.com/ijbis

Beyond the black box: operationalising explicability in artificial intelligence for financial institutions

Sam Solaimani, Phoebe Long

DOI: 10.1504/IJBIS.2025.10071822

Article History:

Received:	20 March 2025
Last revised:	03 April 2025
Accepted:	06 April 2025
Published online:	20 June 2025

Beyond the black box: operationalising explicability in artificial intelligence for financial institutions

Sam Solaimani* and Phoebe Long

Center for Marketing and Supply Chain Management,
Nyenrode Business University,
Straatweg 25, 3620 AC Breukelen, The Netherlands
Email: s.solaimani@nyenrode.nl
Email: phoebelong95@gmail.com
*Corresponding author

Abstract: Artificial intelligence (AI) is transforming the finance sector, driving advancements in fraud detection, risk profiling, and trading strategies. Despite its potential, AI requires robust governance to prevent perpetuating unconscious biases, achievable through the principle of explicability. This study examines explicability in ethical AI governance within finance, focusing on its conceptualisation and operationalisation. Drawing on interdisciplinary literature, the study conceptualises an integrative maturity framework around three core dimensions: transparency, interpretability, and accountability. The framework provides actionable guidance for operationalisation through progressive procedures, tools, and interventions. Empirical validation through expert interviews reveals that explicability should be addressed holistically, operationalised incrementally, and implemented consistently. The proposed explicability maturity framework supports firms in ethically and effectively adopting AI, advancing both academic discourse and industry practices.

Keywords: ethics; explicability; operationalisation; artificial intelligence; financial institutions; maturity model.

Reference to this paper should be made as follows: Solaimani, S. and Long, P. (2025) 'Beyond the black box: operationalising explicability in artificial intelligence for financial institutions', *Int. J. Business Information Systems*, Vol. 49, No. 5, pp.1–38.

Biographical notes: Sam Solaimani is an Associate Professor of Digital Technology, Innovation, & Operations Management at Nyenrode Business University, an Adjunct Professor of Technology & Operations Management at American University in Bulgaria, and a Senior Advisor at Accenture, The Netherlands. He holds a PhD from the Delft University of Technology, focusing on business model innovation in supply networks, graduated Cum Laude in Business Information Systems (MSc) from the University of Amsterdam, and has a BSc in Information Science from Utrecht University. He has published in many academic journals, including *Supply Chain Management: An International Journal*, *European Management Review*, *Journal of Business Research*, *Technological Forecasting and Social Change*, and *Information Systems Frontiers*. His interest revolves around digital transformation and innovation, the non-technical aspects of digital innovation, and the impact of technology on firms' business and operating models.

Phoebe Long is a management consultant at Boer & Croon and an alumna of Nyenrode Business University. She holds a joint MA in Philosophy and Politics from the University of Edinburgh and an MSc in Management from Nyenrode Business University. She specialises in risk management within the finance sector, with a particular focus on understanding risk and ethical dilemmas in emerging technologies.

1 Introduction

The rapid development of artificial intelligence (AI) technology is outpacing societal and legal frameworks designed to uphold accountability (Cunha et al., 2023). In accordance with the social contract as expounded by Hobbes in *Leviathan* in the 17th century, individuals, having submitted themselves to the authority of a sovereign for the sake of order and security, are bound to be held accountable for their decisions and their consequences. However, AI systems, which increasingly make autonomous decisions, challenge these traditional accountability structures, raising critical questions about how responsibility should be assigned in cases of technical failure (Gkeredakis et al., 2021). These challenges demand new frameworks to address accountability in AI-driven decision-making.

In healthcare, AI algorithms are being used for the early detection of brain tumours as well as diagnosing mental health disorders such as anger and anxiety, shifting decision-making roles traditionally held by physicians to AI systems (Gujar et al., 2025; Uddin and Chowdhury, 2024). While these technologies hold promise, they also introduce risks when algorithmic errors occur, potentially jeopardising patient wellbeing (Constantinides et al., 2024). Similarly, the lack of legislative precedents in autonomous vehicles hampers widespread adoption, raising disputes over accountability in insurance and liability claims (Li et al., 2019; Kubica, 2022). In the financial markets, machine learning algorithms and Robo-advisors are increasingly surpassing traditional strategies in prediction and trading (Bouasabah, 2024; Chandani and Bhatia, 2025), but without proper regulatory oversight, their use could increase the risk of systemic disruptions such as flash crashes. These examples illustrate the pressing need for governance mechanisms to address the ethical and practical challenges AI poses. For some innovations, such as autonomous vehicles, the technological possibilities are promising; however, a widespread roll-out is hampered by a lack of legislative and regulatory precedent, which results in insurance disputes and an inability to settle claims (Li et al., 2019; Kubica, 2022).

Discrimination concerns slow some other innovations; for example, in the US judicial system, an algorithm predicting recidivism rates was found to be biased towards white defendants (Belenguer, 2022). As noted by Pereira et al. (2024b), “AI can have biases in decisions when it comes to personal aspects of individuals based on gender, ethnicity, disabilities, and even gross income, being borderline discriminatory towards minorities” (p.439). The complex challenge of unintended AI discrimination and bias is the primary focus of this paper. Unintentional discrimination is when “an apparently neutral rule, criterion or practice would put persons of a racial or ethnic origin at a particular disadvantage compared with other persons” [Council of the European Union, (2000), p.24]. Examples of unintended AI discrimination are in abundance; for instance, in a

recent study conducted by Liang et al. (2023), it was observed that a ChatGPT detection software incorrectly identified non-native English essays as content generated by the GPT model more than 50% of the time. Such misclassification raises concerns about the detection system's accuracy and flags the prevalence of unintended bias in these rapidly developing technologies.

While algorithmic bias poses challenges across sectors, the financial industry faces particularly urgent issues. Bias in AI systems used for credit allocation and insurance pricing disproportionately impacts historically marginalised communities, risking the perpetuation of economic inequities in future generations (Townson, 2020). For example, algorithms have been shown to assign higher credit limits to men than women (Firth, 2021) and impose higher insurance premiums on minority groups, effectively reintroducing prohibited attributes like race through sophisticated proxies (Lehmann, 2021). The concept of technological lock-in (Arthur, 1989) further explains how discriminatory AI models can become entrenched in financial decision-making, making later corrections difficult. These discriminatory outcomes underscore the need for robust governance mechanisms tailored to the financial sector that go beyond technical fixes but that also support "changes in mindset, leadership approaches, and operational practices, integrating digital infrastructure, advanced data analytics, and customer-centric solutions" [Tarigan et al., (2025), p.5].

In the financial sector, biased implications are already coming to fruition; for example, an algorithm for determining credit limits was found to allocate higher credit limits to men than women (Firth, 2021) and concerns have also been raised about privacy, security, and bias in chatbots that are increasingly used in the financial sector (Srivastava et al., 2024). The risks of creating discriminatory AI-enabled solutions are compounded by the potential lock-in effects that new developments can have, and this is the phenomenon of "technologies, once societally successful, [becoming] resistant to change, even if these technologies have adverse effects" [Pesch, (2014), p.926; Stahl et al., 2023]. Mitigating and regulating the potential adverse effects of AI technology is essential (Arora et al., 2023). Indeed, the EU has been crafting AI regulations since 2021 through the AI Act (European Commission, 2021). Several complementary directives have also been developed to support this effort, including the AI Liability Act, which assists victims of AI misuse by lowering the burden of proof to improve the chances of successfully winning liability claims (European Commission, 2023). Further efforts, including the NIS2 directive and the EU Data Governance Act, concerned with improving cybersecurity and data sharing, are also being updated to reflect the new threats and opportunities posed by AI technology (European Commission, 2024; Deloitte Netherlands, 2023). From an industry perspective, the recent surge in the adoption of generative AI tools, notably ChatGPT in 2023, has amplified public interest in AI governance (Korneeva et al., 2023; Ollagnier, 2024). This momentum was fuelled by a collective open letter penned by over 1,000 tech leaders, including Elon Musk, emphasising the perils of an unregulated AI arms race (Metz and Schmidt, 2023; Korneeva et al., 2023).

With AI algorithms becoming increasingly prevalent in the finance industry, it is imperative to minimise discrimination rather than allow it to be embedded into the algorithms. There are two especially pernicious related problems which contribute to this risk. Firstly, the training data, fed to predictive risk algorithms, may be filled with historical discrimination and, therefore, requires some carefully designed corrective interventions (Arora et al., 2023; Townson, 2020). Secondly, even when legally

prohibited characteristics are omitted from training data (e.g., gender and race), the AI algorithms can typically identify alternative, sophisticated patterns which equate to discriminatory attributes (Prince and Schwarcz, 2020). As explained by Prince and Schwarcz (2020), when AI lacks direct data on specific characteristics, it can derive alternative indicators or proxies that are less obvious but still predictive in nature. The result is that the algorithm continues to produce discriminatory predictions despite purporting that such attributes are forbidden from the algorithm. For instance, this problem occurs in life insurance pricing and creditworthiness. Historically, race was commonly used as an underwriting attribute for life insurance (Lehmann, 2021). Despite its current prohibition, the emergence of sophisticated predictive algorithms in the insurance industry leads to concern that the attribute of race could effectively re-emerge, resulting in the potential for minority groups to encounter higher insurance premiums because the data has historically categorised them as a high-risk group (Lehmann, 2021). Similar algorithms are also being used to determine creditworthiness. Against this backdrop, there is expected to be persistent racial discrimination in, for instance, loan pricing, with Latin and African American borrowers paying 7.9 basis points more mortgage interest (Bartlett et al., 2022).

To tackle the risks of unintended discrimination, the 'explicability of AI' is argued as a potential remedy (Van den Berg and Kuiper, 2020; Meske et al., 2022). Explicability integrates transparency, interpretability, and accountability, offering a comprehensive framework for ethical AI governance. However, existing approaches often remain abstract or limited in scope, lacking actionable steps for operationalisation. In that, this study draws on 'ethics as practice' (Stahl, 2012) to argue that explicability must be embedded into organisational practices rather than treated as an abstract principle. By doing so, this study aims to bridge this gap by developing and empirically validating a maturity framework for explicability tailored to the financial sector. It is worth noting that there have been various attempts in terms of principles, frameworks, policies, checklists and more to conceptualise explicability (Krishnan, 2019; Theodorou and Dignum, 2020; HLEG, 2019). However, the extant risk-mitigating approaches are widely dispersed and often only attend to one aspect of explicability, typically proposed by advisors for specific contexts, ranging from individuals and non-profit organisations to local and supranational governments and institutional bodies (Floridi et al., 2018; John-Mathews, 2022). In addition, many recommended principles and policies are limited to self-regulation, and government-level or compulsory legal oversight is mainly lacking (Ada Lovelace Institute et al., 2021). Lastly, many explicability-focused approaches aim to advance our understanding of the concept, remain within a definitional realm, and lack clear directions towards operationalisation. In that, conceptualisation is necessary but not sufficient.

In preserving (and complying with) the ever-changing rules of accountability, an abstract notion of explicability, albeit comprehensive, is mainly useful when operationalised (Floridi et al., 2018). Operationalisation can be understood as making a concept actionable and applicable. This aligns with applied ethics and pragmatism (Dewey, 2008; van de Poel, 2013), emphasising that ethical principle must be tested in real-world contexts and continuously refined. For the same reason, Theodorou and Dignum (2020) argue that to move beyond high-level abstract guidelines, actionable steps in the form of standards and governance, training, and ethics boards need to be adopted. This study synthesises existing knowledge on explicability, proposes an operationalisation maturity framework, and empirically evaluates its validity within the

financial sector. By addressing the fragmented nature of current frameworks, the study contributes to both the theoretical understanding of explicability and its practical application in AI governance. Theoretically speaking, the findings of this study help to give empirical clarity on the concept of AI explicability, i.e., what it means to realise and implement this principle and respond directly to calls for the advancement of explicability frameworks (Floridi et al., 2018; Kuiper et al., 2021) and advancing our understanding on how explicability (and its substituents like explainability) can be used to evaluate, improve, learn and manage AI-driven initiatives within enterprises (Meske et al., 2022). From a practical perspective, this study offers an empirically validated framework for professionals to use for guidance and benchmarking how their organisation applies AI and what associated interventions are required.

The paper begins with a review of explicability theories and frameworks, leading to a conceptual definition and operationalisation proposal. It then describes the research method, presents the findings, and reflects on key patterns against existing theory. The study concludes with implications, limitations, and directions for future research.

2 Literature review

AI encompasses a range of hardware and software-based technologies, widely adopted across industries (Ratten, 2024). The European Commission defines AI as "systems that display intelligent behaviour by analysing their environment and taking actions—with some degree of autonomy—to achieve specific goals" [HLEG, (2018), p.1]. Belenguer (2022) highlights the autonomy of AI, describing it as an intelligent agent capable of independent reasoning, collecting and processing data, and learning to navigate unfamiliar environments. Historically, AI has progressed from narrow applications, such as rule-based systems, to machine learning (ML) approaches, including supervised, unsupervised, and reinforcement learning (Murphy, 2012). More recently, deep learning techniques, particularly in Generative AI (GenAI), have gained prominence for their ability to create novel content like text, images, and audio (Martineau, 2023). In finance, AI applications have transformed critical areas such as market forecasting, audit, credit scoring, loan allocations, risk management, know your customer (KYC) and know your business (KYB) (Ahmadi and Solaimani, 2021; Colmenarejo et al., 2022; Chen, 2020; Pereira et al., 2024b; Willems and Hafermalz, 2021).

The rapid adoption of AI technology and algorithms is outpacing the human capacity to understand it, leading to the potential for undetected discrimination of various kinds when bias unintentionally becomes ingrained in the algorithms (Monod et al., 2024; Stahl et al., 2023). By outsourcing the skills of human reasoning and decision-making we also sacrifice the "clarity, explainability, predictability, teachability and auditability of human actions and replace them with ambiguity" (Nanda and Kumar, 2024). As pointed out by Ntoutsi et al. (2020), AI algorithms are developed by humans and are fed with data generated by humans; hence, "whatever biases exist in humans enter our system and even worse, they are amplified" (p.3). While we are privy to the inputs and outputs of algorithms, several layers of complexity are added in between, such that it is becoming difficult to understand how the original data is being manipulated to produce predictions (Castelvecchi, 2016). Various incidents of facial recognition software highlight the problem, including Google Photo's app categorising Black people as 'gorillas', Nikon's algorithms consistently mistaking Asian faces for blinking, and an Amazon HR AI

algorithm being abandoned due to gender discrimination (Ananny and Crawford, 2018; Dastin, 2018; Wade, 2010). More related to the financial domain, facial recognition software is often used for authentication in personal online banking. However, it has been shown to consistently fail when used by people with darker skin tones, partly because the data set used to train the model overrepresents lighter skin tones (Wehrli et al., 2021). The solution, therefore, seems to lie in improving the diversity of the dataset. However, Ananny and Crawford (2018) argue that this is only part of the solution; the increasing complexity of algorithms and the inherent difficulty in explaining the operations are a much more challenging problem, which hints at why adopting the principle of explicability in AI design and application is necessary.

Notwithstanding the instrumental role of explicability in fostering a sense of meaningful human control over algorithms (Robbins, 2019), the concept is multifaceted and definitionally fluid. The principle of explicability[1], as introduced by Floridi et al. (2018), can be seen as complementary to the traditional principles of bioethics and positioned as the synthesis of "Explicability both in the epistemological sense of 'intelligibility' (as an answer to the question 'how does it work?') and in the ethical sense of 'accountability' (as an answer to the question: 'who is responsible for the way it works?'), is, therefore, the crucial missing piece of the jigsaw when we seek to apply the framework of bioethics to the ethics of AI" (p.700). Such positioning implies that dissecting and delving into its constituent parts of explicability is imperative to understand the principle. Notwithstanding the scattered literature on explicability, as will be elaborated on in the following sections, several scholars underscore the notions of transparency, interpretability, and accountability in explaining explicability (e.g., Bankins and Formosa, 2023; Glavina, 2024; Hermann, 2022; Van den Berg and Kuiper, 2020).

2.1 Transparency

Transparency has emerged as a response to financial scandals and gained prominence in the late 20th and early 21st centuries to counter corruption (Larsson and Heintz, 2020; Gita and Krishnakumar, 2024). Transparency serves as the antidote to the issue of information asymmetries, where one party holds more pertinent information than the other during a transaction. Financial services commonly employ it to counteract the resultant power imbalance caused by such asymmetry. This is achieved by ensuring both parties access the same information, thus fostering a fairer and more balanced transaction environment. To understand transparency, it is also worth acknowledging its metaphorical etymology. Metaphorically, the concept of knowing as seeing, exemplified by expressions like "I see", underscores our understanding of cognitive processes, including the idea of transparency. Transparency, with its positive connotations, is associated with the mental frame of knowing and understanding, while contrasting metaphors with negative implications, such as being in the dark or the 'black-boxed' performance, emphasise the lack of transparency [Larsson and Heintz, (2020), p.7]. The notion of a black-boxed performance refers to (algorithmic) design choices inscribed in software and applications that, on the one hand, are made unavailable for public scrutiny to protect intellectual property (e.g., sophisticated search engines, social media feeds or recommender systems), and on the other hand, are highly opaque as they are not relying on pre-specified, rule-based instructions but are learning based on evolving weights and refined network connections (Faraj et al., 2018). For instance, the use of AI algorithms for personalisation raises ethical concerns regarding data transparency and user

autonomy. Saura (2024) highlights how smart personalisation can create a privacy paradox, where users trade personal data for convenience without fully understanding the implications. However, transparency exists on a spectrum, as even a simple AI application may be perceived as non-transparent if users are unaware of its use (Van den Berg and Kuiper, 2020).

The EU High-level expert group on AI classifies transparency into three categories, i.e., traceability, explainability, and communication, where

1 traceability entails the ability to track the data sets and processes employed

2 explainability involves providing explanations regarding the extent to which an organisation and its decision-making processes are influenced by AI technology, along with accompanying justifications

3 communication entails effectively conveying the capabilities and limitations of the technology to relevant stakeholders, including ensuring users are aware when they are interacting with an AI system and identifying the responsible individuals (HLEG 2019).

Dignum et al. (2018) provide a more comprehensive definition of transparency, "transparency refers to the need to describe, inspect and reproduce the mechanisms through which an AI system makes decisions and learns to adapt to its environment, and to the governance of the data used and created" (p.62). While both definitions underscore the importance of systems being open to inspection, Dignum et al. (2018) emphasise the ability to reproduce the decision-making mechanisms of AI systems, and it is this idea of replication of algorithms outcomes and clarity to regulators and users that makes it one of the more robust conceptions of transparency in the literature. Transparency in AI systems not only addresses information asymmetries but also helps mitigate concerns about privacy intrusion and fairness, as observed in the use of AI for facial recognition systems (Cunha et al., 2023).

2.2 Interpretability

The second constituent part of explicability is interpretability, also called intelligibility, which is concerned with answering the epistemological question of how AI algorithms work [Floridi et al. (2018), p.700]. Biran and Cotton (2017) describe interpretability as the condition in which systems and their operations can be understood by a human, either through introspection or a produced explanation. In Adadi and Berrada's (2018, p.52141) words, an interpretable system is "a system where a user cannot only see but also study and understand how inputs are mathematically mapped to outputs". Although the concept of interpretability in AI elicits extensive debate, it generally refers to a system being sufficiently open and logical for humans to comprehend its functioning to a considerable extent relative to the level of understanding needed for the end-user (John-Mathews, 2022).

In contrast to transparency, discussions surrounding interpretability in AI primarily centre on technological and practical aspects. The level of interpretability needed varies among AI systems, distinguishing between simpler models like linear regression, clustering algorithms, Bayesian models, and more complex approaches such as multi-layer neural networks (Van den Berg and Kuiper, 2020). Simple models can be interpreted by examining the distribution of feature weights, whereas achieving

interpretability for models like linear models with polynomial features and multi-layer neural networks often requires post-hoc[2] analysis by developers with techniques including reviewing feature relevance, explanations by simplification, and visualisation explanations (Van den Berg and Kuiper, 2020). What is evidenced here is that interpretability is a realisable principle, and techniques are being actively developed, and despite academic debates on defining the concept, there is undoubtedly some degree of consensus on how interpretability can be technically approximated.

2.3 Accountability

The third constituent element of explicability is accountability. According to Floridi et al. (2018), accountability aims to identify who is responsible for how the technology works. It should be ensured that "the technology— or, more accurately, the people and organisations developing and deploying it—are held accountable in the event of a negative outcome" [Floridi et al., (2018), p.700]. The high-level expert on AI also heightens accountability as a vital element for developing trustworthy AI (HLEG, 2018). Their conception of accountability typically includes four elements, namely, auditability, minimisation and reporting of adverse impacts, trade-offs, and adequate redress (HLEG, 2019). Auditability refers to the requirement of comprehensive documentation for review purposes, including evaluation reports and impact assessments.

Similarly, minimising and reporting adverse impacts means ensuring safeguarding for whistle-blowers, NGOs, and trade unions reporting legitimate concerns. Furthermore, implementing these requirements may necessitate trade-offs, such as finding the balance between collecting and utilising customer data for innovation while also respecting their privacy (Glavina, 2024). Ultimately, decision-makers must be accountable for these decisions (Gegenhuber et al., 2023). Finally, adequate redress refers to the need for mechanisms to be in place to protect vulnerable persons or groups when adverse outcomes occur (HLEG, 2019). Dignum et al. (2018, p.62) define accountability as "the ability of a system to explain and justify decisions and actions to partners, users, and other stakeholders while incorporating societal norms and moral values". While both the definitions from the HLEG and Dignam et al. (2018) align with the principles of transparency and interpretability, they can be furthered with a more explicit notion of the attribution of accountability, especially in cases of adverse impacts.

The attribution of accountability is closely related to the concept of responsibility. While some literature uses accountability and responsibility interchangeably, Dignum et al. (2018) distinguish them, highlighting that responsibility encompasses both the capabilities of AI systems and the role of human actors. Santoni de Sio and Mecacci (2021) provide a taxonomy of responsibility, including dimensions such as culpability, moral accountability, public accountability, and active responsibility. Culpability refers to taking responsibility for wrongful actions based on intention, knowledge, or control. Moral accountability involves the duty of individuals to explain their reasoning and actions to others in certain circumstances. Public accountability entails officials' obligation to explain their actions to the general public. Active responsibility encompasses the duty to work towards achieving specific societal goals and values (Santoni de Sio and Mecacci, 2021). Santoni de Sio and Mecacci (2021) argue that AI complicates the taxonomy of responsibility, particularly in intricate decision-making, where AI may obscure the rationale behind certain decisions and individual roles. For example, the outcome of these AI complications is increasingly widening accountability

gaps, which erode the public's sense of trust in both the technologies and the institutions that use them (Omrani et al., 2022). A comprehensive approach to explicability must incorporate these nuanced distinctions in accountability gaps to sufficiently address them.

2.4 A synthesised definition for explicability

Explicability in AI serves as a guiding principle that ensures AI systems produce outcomes and decisions that are interpretable, transparent, and supported by an accountability framework. It is essential for addressing ethical concerns, aligning with regulatory requirements, and fostering trust in AI applications. Based on the synthesis of explicability's core elements and their interpretations, the following definition is proposed:

"Explicability in AI refers to the overarching principle that stresses the need for AI decision-making processes and outcomes to be interpretable, transparent, and accountable, where (i) transparency involves openness, traceability, and explainability to address information asymmetries and power imbalances, emphasising clarity, reproducibility, and the ability to describe, inspect, and audit AI systems throughout their lifecycle, (ii) interpretability focuses on the technical understanding of how AI algorithms function, ensuring decision-making processes can be comprehended by humans through techniques such as post hoc analysis and model simplification, and (iii) accountability assigns responsibility for AI outcomes, ensuring mechanisms to address negative impacts, minimise accountability gaps, and foster trust in AI technologies."

These three pillars, i.e., transparency, interpretability, and accountability, are closely interrelated and complementary. Understanding explicability as the synthesis of these elements creates a robust conceptual foundation that acknowledges the nuances of each term while adopting a collective perspective. This integrated approach ensures that explicability remains a practical and comprehensive framework for guiding the ethical design, implementation, and governance of AI systems.

2.5 Operationalisation of explicability

While a definition is a necessary departure point, more focus on operationalisation with actionable properties, shifting from 'what' to 'how', is needed. Various policy frameworks, ethical guidelines, government regulations, audit standards, impact assessments, and suchlike have been proposed to guide practitioners towards ethical AI (European Commission, 2021; Ada Lovelace Institute et al., 2021; Stahl et al., 2023). However, these guidelines are highly scattered, and operationalisation is often implicit, if not completely overlooked (Floridi et al. 2018; John-Mathews et al., 2022).

2.5.1 Operationalisation of transparency

To operationalise transparency, firms should begin by developing an ad hoc understanding of transparency specific to their context and assess its alignment with their organisational goals. Additionally, the literature stipulates a minimum requirement of notifying end-users that they are interacting with AI systems (Van den Berg and Kuiper, 2020). Once these foundational steps are established, firms can enhance their internal traceability procedures. This includes formally informing all company members about the utilisation of AI systems. Additionally, internal AI impact assessment reports can be

conducted and made accessible to all employees (Ada Lovelace Institute et al., 2021). Impact assessments are tools that can be utilised to examine how AI technologies work in context and are specifically interested in "seeking to better understand, categorise and respond to the potential harms or risks posed by the use of these systems" [Ada Lovelace Institute et al., (2021), p.21]. For companies at a higher AI maturity level, traceability is made externally possible, and methods, tools, and processes are in place to keep internal and external stakeholders, especially users, informed about various stages of data collection and algorithmic processing (Mora-Cantallops et al., 2021). To ensure a representative and diverse data set, samples of the data should be examined, with careful consideration given to the selection process to accurately reflect the diversity and complexity of the real-world scenarios they aim to model and predict (Glavina, 2024). A similar initiative encourages publishing source code and operating logic on government or interest group registries (Ada Lovelace Institute et al., 2021). At more advanced maturity levels, firms might proactively seek out private accreditation opportunities, such as AI risk management audits, to gain the trust of stakeholders, therefore turning transparency into business value (Martin, 2019; Omrani et al., 2022).

2.5.2 Operationalisation of interpretability

When operationalising interpretability, low-maturity firms might have informal discussions of what interpretability means, but it likely needs to be clearly defined for all employees (Krishnan, 2019). Firms prioritising interpretability should establish robust governance structures, which clarify by whom, what, and how AI systems are governed (Scheinder et al., 2023). Governance structures are complemented with clear guidelines (e.g., regarding integrating AI tools like GPT) to ensure managers receive and disseminate trustworthy AI-extracted information and use it effectively in their decision-making process (Rana, 2023; Van den Berg and Kuiper, 2020). For even more mature firms, it is vital to incorporate interdisciplinary perspectives (for example, from law, philosophy, and psychology) within the development team to help recognise potential bias risks early (Theodorou and Dignum, 2020). Multiple degrees of expertise are required to operationalise interpretability; therefore, well-designed and targeted training should be seen as valuable and actionable. Training can be viewed from two perspectives. Firstly, computer science experts must be "trained and perhaps licensed in the safety and societal implications of their designs and implementations, just like those of other disciplines" [Theodorou and Dignum, (2020), p.2]. Secondly, executive management needs adequate data and technological literacy to make informed decisions. At higher maturity levels, interpretability processes are formalised, interdisciplinary perspectives are incorporated, and the technology is monitored across its service lifecycle. This includes the company using highly specialised developers who can conduct post hoc analysis, and the governance structure clearly outlines how their findings are distributed throughout the company (Van den Berg and Kuiper, 2020). Tools such as local interpretable model-agnostic explanations (LIME), Shapley additive explanations (SHAP), and the ELI5 toolkit can enhance interpretability by providing visualisations that clarify why certain predictions are made. LIME approximates the model locally to show input-output relationships, SHAP assigns importance values to features, and ELI5 offers methods to inspect and explain model performance (Vishwarupe et al., 2022).

2.5.3 Operationalisation of accountability

When operationalising accountability, low-maturity firms might create some internal standards and policies; typically, they are aware of a need to comply with external legislation (e.g., the EU's AI liability directive), but more formal processes are required. Furthermore, the bare minimum requirement of employing an AI officer is met; this role entails overseeing the deployment and monitoring of AI systems, ensuring compliance with relevant regulations, and managing the ethical implications of AI technologies within the organisation (Theodorou and Dignum, 2020). For slightly more advanced firms, formal protocols for adequate redress are defined in case AI outcomes are unjustly impacted; for example, responsible individuals are consulted, and procedures are examined. As firms continue to develop in their AI maturity, the various stages of the AI lifecycle (initiate, build and train, deploy, manage and operate) must be clearly distinguished (Sullivan and Wamba, 2022). The steps to mitigate adverse outcomes are clear and easily implemented. At this level, accountability is understood from a reactive perspective (Sandler and Basl, 2019).

At a higher maturity firm, and as an overlapping point with interpretability, interdisciplinary perspectives on accountability within the management team are implemented and monitored, ensuring that decisions consider various concerns and trade-offs, including ethical, legal, and technological. In addition, management is well-trained to provide technological literacy and have a meaningful understanding of the developers' work (Theodorou and Dignum, 2020). For the most mature firms, the approach to accountability is formalised through protocols and by including interdisciplinary perspectives. Furthermore, they have a proactive stance to implementing accountability by creating an ethics committee who are active in resolving disputes, monitoring and ensuring compliance with all relevant policies and seeking extra opportunities to go beyond compliance, for example, partnering with ethical AI communities such as Partnership on AI (Sandler and Basl, 2019). Existing literature suggests that AI ethics boards should resemble those in universities and hospitals, possessing the authority to veto projects that do not adhere to ethical guidelines (Theodorou and Dignum, 2020).

2.6 A maturity model for explicability

The techniques, tools, and interventions extracted from existing guidelines and discussed in the previous section provide a basis for an integrative maturity framework for explicability. Generally, maturity models are recognised as an effective way for organisations to measure their progress against established benchmarks (Oruthotaarachchi and Wijayanayake, 2023). The merits of a maturity model for explicability include assisting firms by showing

1 what the available tools, techniques and interventions are

2 facilitating firms with their particular and changing needs for explicability

3 providing a baseline for benchmarking so that firms can identify their AI maturity (Adekunle et al., 2022).

There exist several commonly used maturity models in the IT sector, including the capability maturity model integration (CMMI SM 2002), the COBIT framework (Lainhart et al., 2019), and the Risk Management Framework for Information Systems

and Organisations (NIST, 2018). What unites these maturity models is that they all take a stage theory approach (Kazanjian and Drazin, 1990). Although these models have many relevant elements, this study proposes conceptualising a maturity model based on the American Institute of Certified Public Accountants/Canadian Institute of Chartered Accountants (AICPA/CICA) privacy maturity model because it deals with large quantities of data, which is also a vital issue in the case of AI. The shared requirement for data risk management unites these two topics and makes the AICPA/CICA a solid foundation for building the proposed framework. Furthermore, the AICPA/CICA clearly distinguishes between maturity stages and is transparent in its requirements, which contrasts with other frameworks that require a large degree of subjective interpretation (Simonsson et al., 2010). While in this study, the building blocks of such a maturity framework can be derived from the extant dispersed body of knowledge (see previous sections), the maturity stages were adopted from the AICPA/CICA privacy maturity model with five states, i.e.,

1 'ad hoc' implying "procedures or processes are generally informal, incomplete, and inconsistently applied"[3]

2 'repeatable' suggesting "procedures or processes exist; however, they are not fully documented and do not cover all relevant aspects"

3 'defined' implying "procedures and processes are fully documented and implemented, and cover all relevant aspects"

4 'managed' implying "reviews are conducted to assess the effectiveness of the controls in place"

5 'optimised' indicating "regular review and feedback is used to ensure continuous improvement towards optimisation of the given process" [AICPA/CICA Privacy Maturity Model, (2011), p. 2].

An overview of the discussed body of knowledge on the operationalisation of explicability organised in the format of the AICPA/CICA maturity model is presented in Table 1.

Table 1 Proposed conceptualised framework for operationalising explicability

	Explicability levels of maturity			
	Repeatable	*Defined*	*Managed*	*Optimised*
Transparency	Transparency is ad hoc, with informal practices that offer limited instrumental value. AI usage is minimally disclosed to stakeholders.	Transparency practices are formalised, with AI information accessible and traceable. Impact on organisational processes is systematically documented. Stakeholders are notified of AI interactions.	Transparency extends to external audits and stakeholder engagement. Full lifecycle traceability is in place, supported by algorithmic documentation and periodic reviews.	Transparency achieves industry-leading standards through proactive accreditation and certifications, turning it into a competitive business advantage. Stakeholders have real-time access to explainable AI decisions.

Table 1 Proposed conceptualised framework for operationalising explicability (continued)

	Explicability levels of maturity			
	Repeatable	*Defined*	*Managed*	*Optimised*
Operationalisation of transparency	Initial understanding of transparency is reactive and inconsistent. Users are passively informed about AI interaction.	Internal traceability processes ensure employees can access AI usage and impact assessments. Sampling of datasets and partial algorithm audits are conducted.	External traceability mechanisms ensure stakeholders are consistently updated, including algorithm registries and third-party audits.	Transparency protocols are integrated into the firm's strategic goals, with external accreditation ensuring public trust. Transparency becomes a benchmark for industry collaboration and innovation.
Interpretability	Basic understanding of AI models exists within a limited group of employees. Definitions of interpretability lack consensus.	Interpretability is structured with governance frameworks guiding the dissemination of AI outputs across teams. Training programs ensure developers understand and explain AI decisions to relevant stakeholders.	Stakeholders from multiple disciplines (e.g., legal, ethical, and technical) actively collaborate to refine interpretability practices. Bias risks are assessed systematically.	Interpretability is embedded into organisational culture, with continuous learning mechanisms and partnerships with regulators to align AI decision-making practices with evolving standards.
Operationalisation of interpretability	Interpretability discussions are siloed and lack formal structure. Developers analyse outputs post-hoc using basic techniques.	Governance structures and guidelines ensure that AI decision-making outputs are accessible and explainable to all relevant stakeholders. Training incorporates interdisciplinary insights to minimise bias.	Interdisciplinary review mechanisms are established to regularly assess and improve AI model interpretations. Advanced explainability tools like SHAP, LIME, and ELI5 are systematically applied.	Interpretability is integrated into organisational workflows, with cross-departmental collaboration ensuring decisions are communicated effectively. Predictive monitoring tools are implemented to ensure interpretability across the AI lifecycle.

Table 1 Proposed conceptualised framework for operationalising explicability (continued)

	Explicability levels of maturity			
	Repeatable	*Defined*	*Managed*	*Optimised*
Accountability	Accountability is reactive and limited to post-implementation assessments. Roles and responsibilities in AI development are vaguely defined.	Accountability standards are developed internally, with compliance to external regulations. AI officer roles are formalised, and initial redress protocols for unjust outcomes are defined.	Auditability processes enable clear role identification, and responsibility for AI outcomes is disseminated across organisational levels. Management ensures regular training in technological literacy.	Accountability is a proactive, firm-wide commitment led by an ethics committee. Multidisciplinary teams ensure compliance with regulations while fostering innovation. Accountability mechanisms are continuously refined to exceed industry standards.
Operationalisation of accountability	Internal processes are informal, with limited awareness of external compliance. Accountability is addressed sporadically.	Defined accountability frameworks assign clear roles for AI outcomes. Lifecycle stages of AI systems are distinguished, with mitigation strategies for adverse effects in place.	Justification protocols for AI decisions are established, ensuring stakeholders have access to comprehensive explanations. Proactive governance integrates diverse perspectives into accountability discussions.	Accountability is institutionalised, with management adopting a proactive stance. Formalised ethics committees ensure that accountability is embedded into strategic decision-making, driving ethical innovation.

3 Research approach

Given the explorative nature of this study, i.e., an empirical evaluation of a conceptualised maturity model for AI explicability, a qualitative, inductive approach is considered most suitable (Glaser and Strauss, 1999). According to Tavory and Timmermans (2014), an inductive approach is "an analytical choreography with an immersion in data and transcend to higher levels of abstraction" while also having the flexibility to return to the literature. As such, space was created to develop a novel framework, which was embraced and offered for feedback rather than tossed away during the fieldwork (Conaty, 2021). Such an explorative approach can help evaluate the proposed framework's relevance, coherence, and comprehensiveness. In doing so, several semi-structured interviews with experts were conducted. Semi-structured interviews blend open and closed questions, encouraging follow-up inquiries while helping the interviewees to feel at ease, leading to more natural conversations (Adams, 2015). A

semi-structured interview also allows flexibility in discussing pre-defined (theory-driven) themes and unravelling and zooming in on unforeseen topics (Adams, 2015). In this study, the interviewees were subject matter experts, and a semi-structured mode of interviews facilitated natural, nuanced discussions and the exploration of topics from and beyond the proposed maturity model.

3.1 Research sample and participants

A total of 16 interviews were conducted. The number of interviews aligns with previous studies on similar topics, e.g., a paper on management perspectives of ethics in AI with nine interviews (Baker-Brunnbauer, 2020) and explainability in AI with 13 (Kuiper et al., 2021). Moreover, it adheres to a recommended baseline of 12 interviews, as Guest et al. (2006) put forward, which finds that data saturation can mainly be achieved within 13 interviews. The sample size of 9–17 is similarly supported by Hennink and Kaiser, who argue that this range represents the ideal balance between not reaching saturation and raising ethical issues, "such as wasting funds, overburdening study participants, and leading to wasted data" (2022, p.8).

Non-probabilistic sampling was used to identify prospective experts, leveraging the alum network and other university business networks. A snowball sampling approach was adopted to ensure access to a vast pool of relevant participants. Interviewees were grouped into three categories: professionals working with AI in the finance sector, academics specialising in explainable AI, and technology experts with engineering expertise. These categories were selected to ensure the study captured a holistic view of AI explicability, addressing both practical and theoretical aspects. This categorisation aligns with prior research on AI maturity model development, highlighting the importance of industry-specific approaches (e.g., Sonntag et al., 2024). The diverse perspectives of the interviewees, comprising seven business professionals, two technology experts, and six academics, significantly contributed to the external validity of the research (Daly et al., 2007). Additionally, including participants from different countries, i.e., the UK, the USA, the Netherlands, and Sweden, broadened the international scope of the study and ensured the findings were applicable across multiple contexts.

An interview protocol was developed[4] to preserve internal validity, outlining critical aspects such as an invitation letter, interview questions, definitions, examples, a-priori codes, time guidance, and informed consent (Brooks and King, 2014). This protocol was carefully aligned with the study's objectives to ensure systematic and relevant data collection. The interview questions were categorised into three themes:

1 defining explicability to evaluate the comprehensiveness of the proposed framework

2 operationalisation of explicability focusing on transparency, interpretability, and accountability

3 feedback on the framework's applicability, coherence, and comprehensiveness (Adekunle et al., 2022).

The protocol underwent iterative refinement following trial interviews, which ensured clarity and engagement, further strengthening the data collection process.

Table 2 Specifications of interviewees

	Interviewee	*Role/function*	*Expertise*	*Industry (department)*	*Experience (years)*	*Location*
Academia	Expert A (AC)	Professor and expert in AI and EA	xAI	University (Computer Science)	5	NL
	Expert B (AC)	Professor and expert in Technology Law	Trustworthy AI	University (Technology & Social Change)	10	SE
	Expert C (AC)	Senior Researcher in Emerging Technologies	Human-centred AI	University (Computer Science)	5	NL
	Expert D (AC)	Professor and expert in Technology Law	Trustworthy AI	University (Computer Science)	12	SE
	Expert E (AC)	Professor of Philosophy	Human-AI interaction	University (Philosophy)	10	US
	Expert F (AC)	Professor of Computer Science	xAI tools	University (Philosophy)	9	UK
Business	Expert A (BU)	Manager	AI Strategy	Financial services (Data & Technology Advisory)	4	UK
	Expert B (BU)	Manager expert in KYC	Data Analytics	Bank (Risk Management)	6	UK
	Expert C (BU)	Partner	AI Strategy	Financial advisory (Responsible AI)	9	NL
	Expert D (BU)	Manager	AI Strategy and architect	Bank (Compliance)	20	NL
	Expert E (BU)	Director	AI strategy	Finance & banking consulting (AI & ethics lead)	6	NL
	Expert F (BU)	Partner	AI strategy	Finance (Technology Consulting)	4	UK
	Expert G (BU)	Partner	Data & AI driven Innovation	FinTech (Data & AI)	15	NL
Technology	Expert A (EN)	AI Engineer	Model development	FinTech (Data & AI)	7	NL
	Expert B (EN)	AI Engineer	Model development	Commercial Banking (Data & AI)	7	NL
	Expert C (EN)	Senior Project Manager	Full-stack engineer	Bank & Insurance	12	NL

Interviewees were first asked to share their unbiased perspectives and experiences in AI projects to ensure a common understanding of concepts. This was followed by presenting a literature-based definition (see Section 2.4) to establish a shared foundation for discussing explicability. To mitigate potential biases, interviews were anonymised during transcription to prevent any contextual influence on coding decisions. Additionally, findings were cross-validated through iterative discussions among the researchers, ensuring the robustness and reliability of the themes identified. The interviews were recorded by Microsoft Teams (version 1.5.00.31156), and the automatic transcription feature was utilised. All transcriptions were subject to post-editing to correct incomprehensible words (caused by obstructed audio) and remove filler words. To ensure that the interview questions are easily understandable, three trial interviews were conducted as suggested by Brinkmann and Kvale (2009), based on which the decision to add examples and time guidance to the interview protocol was made. Ultimately, the average length of each interview was 45 minutes.

All interviews were conducted with the interviewees' informed consent to the terms of conditions, ensuring anonymity, allowing for potential quotation use, and granting the option to withdraw consent within a week. Anonymisation of organisations in the finance sector was agreed upon, with identifying details promptly removed during transcription and replaced with unique codes stored separately, to be deleted after the study's completion.

3.2 Data analysis approach

The data analysis followed the approach of Miles and Huberman (1994), beginning with familiarisation with the interview transcripts. Transcriptions were initially coded using descriptive codes, which denoted the straightforward meaning of the text, followed by the assignment of pattern codes to uncover deeper relationships and themes (Miles and Huberman 1994). The qualitative analysis software Atlas.ti (Web-22) facilitated the systematic visualisation of patterns, identifying relationships across transcripts, and refining nested codes. This iterative coding process ensured that the analysis remained grounded in the research objectives, allowing the themes to emerge organically while maintaining alignment with the proposed maturity framework. Miles and Huberman (1994) suggested that the coding process started with a provisional list of codes before fieldwork, keeping research questions central. The initial codes were based on the proposed maturity framework, refined and expanded with emerging empirical insights. Accordingly, code labels and structure were adjusted when needed, reflecting an evolving analysis (Miles and Huberman, 1994).

The a priori codes that resonated with the interviewees and the emerging codes extracted from the interviews were converted into category cards, forming a codebook (Miles and Huberman, 1994). The Atlas.ti software aided this process, providing easy access to all codes through a 'code manager' feature, each accompanied by a short definition and a reminder of a relevant moment as evidence. Further refinement is realised and distilled into the final code list (a list is available upon request). More profound thought themes and patterns began to form throughout the coding process. Continuous referencing between data and existing literature was necessary to identify fully explicated and unexplored themes and relationships. Another measure to enhance internal validity and find consensus amongst authors was to have interactive sessions to review and discuss the codes and themes. These sessions resolved disagreements on

assigning codes, and themes became more refined and deemed essential for generating sufficiently reliable conclusions (Krippendorff and Craggs, 2016). Consequently, themes were examined against the proposed framework and categorised as emerging, confirming, or adjusted based on their alignment with the corresponding cell (e.g., Solaimani et al., 2022). Finally, a network data display was created, visually illustrating theme and code interrelationships, facilitating iterative adjustments between data and literature [Miles and Huberman, (1994), p.94]; see Appendix. An elaboration of the findings follows in the next section.

4 Results

The interview data offers a nuanced perspective on the conceptual model presented in Section 2. This section delves into the empirical findings concerning the three dimensions of explicability.

4.1 Transparency

Corroborating with earlier discussion in Section 2, a crucial first step for operationalising transparency is making users actively aware of their interaction with AI: "*If you are talking to someone online, as a user, you should be aware that you are communicating with a chatbot... it should be clear how AI is used, the limitations and potential unknown (or harmful) scenarios while using it*" (Expert A_AC). Another early requirement within the proposed maturity model is determining the appropriate level of transparency required for each company. As discussed in Section 2, transparency plays a crucial role in the finance sector compared to other sectors, namely as an antidote to information asymmetries. However, what the interviewees brought to light is that firms' adherence, even within the financial industry, varies based on their unique characteristics, including their sensitivity to societal and environmental factors – e.g., commitment to environmental, social, and corporate governance (ESG) – firms' ownership and investment structure (e.g., dependency on public funding), firms' dependency on public opinion: "*Some financial firms are intrinsically more transparent than others because of their societal mission and environmental aspirations… some are heavily reliant on their intellectual property, some are directly dealing with end customers and need to be transparent to build trust*" (Expert A_AC). Once this level is determined, firms must be intentional about how they convey their position on transparency to their customers, as one interviewee noted:

> "Firms can assess the comprehensibility of their Code of Conduct (CoC) from the client's viewpoint, which may seem like a basic transparency requirement, yet even leading market players like Facebook have faced challenges in this regard, particularly considering that Facebook isn't handling the most sensitive personal information, such as financial details." (Expert G_BU)

Furthermore, the interviews highlighted a range of complexities regarding how companies use AI, an aspect not fully addressed in earlier sections. For example, corporate banks use unsupervised learning AI lending models with "*risk quantification being used to grant or decline financing*" (Expert B_EN). In contrast, one interviewee who works in personal banking said that although their organisation works with AI

modelling, "*the outputs are not being used to drive customer outcomes yet because they cannot be certain the output is not biased*" (Expert B_BU). For other organisations, "*employees do not even know what they have in AI and analytics*" (Expert E_BU). An essential starting point is, therefore, to determine what kind of AI a firm uses and how complex the models might be because this determines the appropriate level of transparency required for each company.

Once basic transparency measures are in place, firms are advised to look beyond as-is analysis and develop traceability measures. One interviewee considered traceability to involve both technical and organisational aspects, "*both the technical aspects, like developing model analysis methods such SHAP[5], and more governance-based practices such as conducting process flows where knowledge is collected and documented are integral parts of traceability*" (Expert B_AC). Aligned with the earlier discussed principles of traceability in AI, the findings highlight the significance of traceable processes in upholding transparency at higher maturity levels. One of the interviewees (Expert A_AC) put forward the example of a bank that wrongly denied loan applications because of biased credit risk models and emphasised that *"…it is increasingly crucial for a financial service provider to be able to backtrack the automated actions of the model and identify distinct steps within the decision-making process*". At the lower AI ethics capability maturity level, the transparency measures are primarily focused on documenting expected scenarios. At the same time, more mature firms pay equal attention to mechanisms for mitigating and managing unforeseen situations. A case in point is the example given by one of the experts: "*We should be able to follow up if something goes wrong, for example, when we got surprised that a chatbot started using offensive language; in such a situation, you cannot get off scot-free by calling it a 'black box' case*" (Expert B_AC). To ensure a fully-fledged transparency policy, a holistic view across the data lifecycle is posited, i.e., from data generation and collection to consumption and production, is essential: "*Transparency involves the ability for continuous monitoring of the data quality across the lifecycle, often structured along the customer journey, from data generation, collection and storage, to deployment, maintenance and decommissioning, using the principle of CIA triad (Confidentiality, Integrity and Availability)*" (Expert F_BU).

In putting transparency into action, several recurring themes were highlighted that are aligned with earlier discussed theories, such as the need for audits and registries in the firms' governance and control apparatus. More specifically, in the financial sector, the standardisation of financial statements (e.g. International Financial Reporting Standards) and requirements for record-keeping are ubiquitous as processes for achieving transparency. It appears that as AI becomes more prevalent in the financial sector, these similar high standards of traceability and documentation will also be necessary: "*The next thing on our path, certainly for the more complex models, is what is termed AI audits […] to do assurance over the model itself"* (Expert F_BU). Some examples of AI audits adopted by the sector include algorithm performance assessments, model evaluations, and bias detection techniques. These audits aim "*to understand how algorithms perform and inspect for inherent bias such as data imbalance or discriminatory patterns*" (Expert A_BU). In a similar vein, registries are recommended to be utilised to build public trust in a sector often criticised for its scandals. As one interviewee said, "*The Flash Crash of 2010 showcased the risks of unregulated algorithmic trading. Algorithmic registries could offer transparency and oversight, guarding against future market turmoil*". Therefore, increasingly, financial institutions make their source codes available on public

registries, such as the Dutch government's algorithm registry, to gain customers' trust. Another interviewee pointed out that firms interested in garnering a higher level of transparency should be open to sharing their internal, non-client-facing AI tools, "*Firms often disregard this concern, yet recognising the limited trust society has in AI, one could argue that true transparency extends beyond what merely catches the client's eye*" (Expert G_BU).

4.2 Interpretability

From a technical perspective, companies, particularly those early in their AI capability maturity, can increasingly use tools and solutions. For instance, "*A tool like SHAP and LIME can be leveraged by any organisation utilising AI, providing a straightforward method for understanding and identifying the key factors influencing AI-generated decisions*" (Expert B_EN). While tools can help mitigate the earlier discussed black box problem, transparent business standards are also needed. In financial sectors, transparent business standards such as the GDPR help standardise work processes across the industry and contribute to favourable outcomes for consumers. These business standards should be equally adopted towards AI. The interviews also stressed a nuanced perspective that for companies with lower levels of AI ethics maturity, "*these regulations are often restrictive to business rather than supportive. A mature organisation should facilitate AI-enabled business innovation triggered by privacy considerations rather than despite it*" (Expert G_BU). Mature organisations are encouraged to embrace regulation and think along with it at the higher levels of the proposed framework.

The interviews repeated the criticality of governance structures to reach higher levels of maturity in AI ethics capability, "The [AI] systems are much more than a model, involving data access and authorisation, business and decision-making processes, which should be considered as part of an integrative system: defining roles and responsibilities, establishing communication lines between departments, and structuring teams to incorporate stakeholders perspectives effectively" (Expert A_AC). The governance structure facilitates information sharing and access among pertinent stakeholders and indicates the frequency and structure of stakeholders' consultation and review sessions, ensuring transparent accountability, such as employing multiple lines of defence in data governance, spanning from technical to business domains. It provides a distinct allocation of responsibilities, distinguishing between change-owners and run-owners, thus fostering efficient and effective decision-making processes; for more mature organisations, tailored explanations of how AI processes technically function must be delivered to various stakeholders. For one commercial bank, this means, "For every loan application we receive, we provide insights to the customer and the people who handle the application insight on how model output decisions were reached. For example, an associated explanation is given if the loan is denied" (Expert C_EN).

After establishing fundamental governance, companies are encouraged to integrate diverse viewpoints from different fields to combat the often-found incongruity between stakeholders: "*Often, Data and AI scientists solely focus on statistical perfection; however, these risks create incomprehensible models to crucial stakeholders. Integrating a range of viewpoints is key to ensuring meaningful understanding and collaboration*" (Expert G_BU). In addition, the concept of explainability-by-design is emphasised, for instance, awareness through training: "*We train our workforce to assess the ethics and explainability of the models which we build locally within the team instead of fully*

relying on a (central) unit that is looking for weak spots; after all, the creators are most knowledgeable about their creation" (Expert B_EN). The training, mainly when provided in a multidisciplinary fashion with mixed content on technical, legal, ethical, and business topics, can alleviate the potential disconnect between engineers and business agents. Similarly, a more diverse set of key performance indicators (KPIs) was suggested to be monitored in evaluating the extent of interpretability:

> "Firms rely on conventional KPIs regarding performance, e.g., how fast a process is performed. However, the challenge is how you measure interpretability. You need a mixture of qualitative and quantitative measures, which can only be formulated within a multidisciplinary team, for instance, by examining feature importance and conducting domain-specific evaluations" (Expert G_BU).

Complementary to the earlier themes, the importance of culture and behaviour was also discussed. To sustain the organisations' attention on the AI ethical issues, the more mature organisations are firmly committed to having a culture of interpretability; business and technical staff appeared to be aligned as one interviewee underscored his team's awareness of the ethical implications of the models they build and their receptive attitude towards business perspectives, "Openness is vital for all to voice concerns or ask questions about the model, as diverse stakeholders bring varied expertise; fostering curiosity requires a safe environment." (Expert B_EN). Similarly, the interviews revealed that in mature organisations, the culture is geared towards continuous improvement and adaptability:

> "Firms with more advanced ethical codes of conduct do not rely on a one-shot risk assessment project, as it will lead to a catch-up game with the continuously changing rules and directives. Instead, these firms invest in risk management capability that underscores a continuous review of all AI ethical and compliance aspects, from integrity to fairness, transparency, responsibility, and the like." (Expert G_BU).

4.3 Accountability

According to several interviewees, accountability can better be seen as a spectrum rather than a binary factor, and its extent is, among others, dependent on the firm's exposure to risk: "*Clients with significant exposure, who heavily utilise AI technologies, methods, or datasets, particularly in critical areas like credit risk, inherently face heightened levels of risk*" (Expert A_BU). In high-risk situations, the repercussions of adverse events are magnified. When companies anticipate being held accountable, they tend to exercise greater caution and diligence in risk assessment and management. Conversely, companies facing lower risks may not prioritise accountability practices as heavily, lacking the same sense of urgency. Hence, the highest levels of maturity for all aspects of explicability can be desirable but not necessary for every company.

Furthermore, when evaluating risk, as pointed out by one interviewee, it is necessary to take a high-level view, as "being mindful of the vulnerabilities of your value network partners is crucial when managing and mitigating AI risks, as every entity in the chain is exposed. Failing to do so is akin to applying a Band-Aid to a bullet wound" (Expert G_BU). The importance of value chain risk management is made evident by the fact that it features as a critical element of the EU AI Act, emphasising the necessity for

accountability and adherence to regulations irrespective of a company's role in the development or distribution of AI systems within the EU:

> "The EU AI Act, currently under consideration by the EU Council and Parliament, imposes strict accountability requirements on companies utilising GPAI models. It is crucial to note that these obligations apply regardless of whether a company is the developer of the AI system, importing it from outside the EU, or part of the supply chain facilitating its availability within the EU." (Expert G_BU)

As an early requirement for operationalising explicability, the interviews highlighted the case for implementing an AI officer, "*While very sporadically, I have observed some organisations establishing an AI officer whose responsibility is, among others, to look after algorithmic fairness… it would be beneficial to have more internal awareness of such initiatives, which ultimately aids in maintaining oversight*" (Expert B_AC). Another noted that in some organisations, "*the responsibility of AI compliance rests fully with the data office, with CIO [Chief Data Officer] or CDO [Chief Data Officer], while the problem clearly has a multidisciplinary nature and therefore warrants a position of its own*" (Expert G_BU). Strikingly, interviewees commonly appreciate having a designated AI officer to manage and streamline the audit process for transparency and accountability.

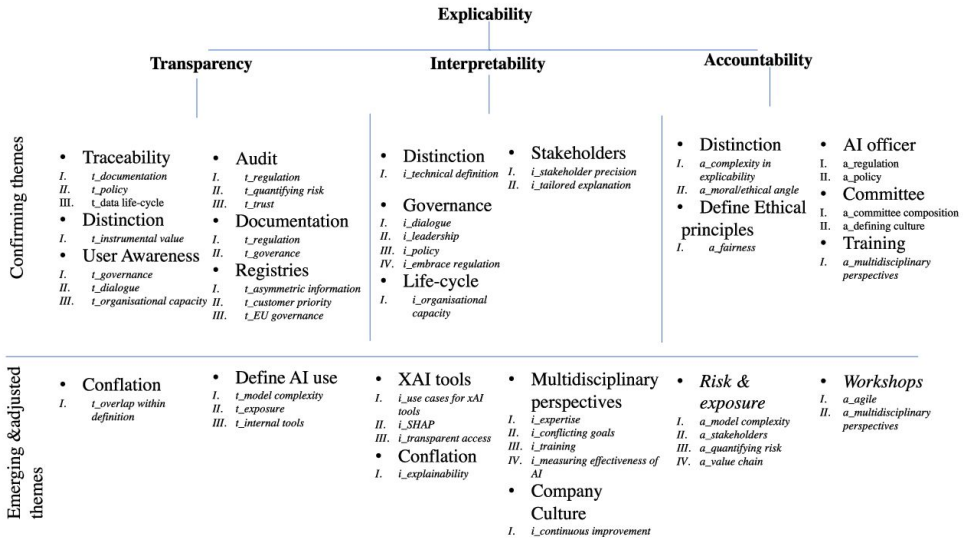
An additional way to operationalise accountability is to establish an advisory board to provide oversight, guidance, and strategic direction on ethical and responsible AI practices. The intervention is expected to be more for firms with mature AI ethics capability as appointing and organising such an entity, while aligned with other existing roles and functions, requires a delicate process. Also, an advisory board is expected to be proactive and visionary; e.g., "*an advisory board is not just a checkbox to mark off or a simple 'yes' or 'no' authority*; *it is a proactive group that offers insightful guidance and functions as an ethical compass*" (Expert D_BU).

Another recurring intervention was organising and participating in workshops on accountability, which covered topics such as bias detection and mitigation and ethical AI design. One interviewee had experience developing workshops within his research group and has created a 'game' that is "*aimed at helping people understand the ethical concerns of an AI system from the perspective of the system's stakeholders*" (Expert A_AC). Using tools like workshops to incorporate diverse perspectives is a novel approach for financial companies to foster accountability. This helps stakeholders grasp their new roles and act accordingly. Accordingly, Expert E (BU) emphasised:

> "Being accountable and behaving like one are two different things. The stakeholders need to understand their new or enhanced governance structures but then need to be facilitated in adapting to their new roles. Governance is the ideal structure on paper; accountability is about how stakeholders 'live' those structures by 'owning' the role and the inherent sense of responsibility."

Recognising a wide range of perspectives echoes earlier findings on the importance of interdisciplinary training for interpretability, once again showing the interdependence between interpretability and accountability. In sum, the interviews substantiated the earlier discussed interventions and added several new ideas; Figure 1 presents a summary of confirmed and emerging themes.

Figure 1 An overview of emerging, confirming, and adjusted themes



5 Discussion

While there are several ways to describe explicability, the existing literature and the empirical insights collected in this study converge around three dimensions, i.e., transparency, interpretability and accountability (e.g., Bankins and Formosa, 2023). Transparency in AI, as highlighted by Hermann (2022) and Omrani et al. (2022), is confirmed through empirical insights and functions as a crucial supporting element that facilitates the application of other ethical principles and enhances the trust and understanding of firms that utilise AI. Interpretability, identified by Hunkenschroer and Luetge (2022) as the antidote to black-box models, ensures that AI algorithms are understandable and can be scrutinised, thus mitigating potential risks and fostering accountability. As Tóth et al. (2022) argued, accountability encompasses the dispersion of responsibility within AI systems, prompting organisations to consider the ethical implications of their decisions and actions. Together, these three concepts form the basis of explicability in AI (Floridi et al., 2018).

5.1 Relational dynamics between the ethical constructs

The findings highlight that while transparency, interpretability, and accountability are distinct dimensions, they are strongly interdependent. For instance, appointing an AI officer not only strengthens governance structures but also enhances transparency and accountability, as seen in financial firms striving to comply with the EU AI Act. This supports Hermann's (2022) argument that transparency serves as a critical enabler of accountability. Also, interdisciplinary cooperation appeared to contribute to both interpretability and accountability (Larsson and Heintz, 2020). In sum, the findings imply that the three central constituent principles and their subgroups interrelate in various

ways, including facilitation, measurement, exemplification, enablement and dependency connections (see Appendix for a more detailed view of interdependencies). Therefore, it can be argued that explicability is better understood holistically where the constituent elements are interconnected (Larsson and Heintz, 2020), with their meanings derived from how they are used (Jackman, 2020).

5.2 Translating ethical principles into practice

The interview data underscores the gap between the theoretical understanding of explicability and its practical implementation. For financial institutions, operationalising explicability necessitates tailored interventions such as algorithmic audits, robust governance structures, and interdisciplinary collaboration, aligning with Hermann's (2022) emphasis on continuous reflection and improvement. The experts in this study, and as suggested by John-Mathews et al. (2022), confirm that high-level principles must be backed up with bottom-up pragmatic actions that are sensitive to the specific characteristics of the firm, the sector they operate in, and the complexity of the model chosen. According to some critics, and as explained by John-Mathews et al. (2022), AI ethics is excessively dominated by principlism, where theoretical moral frameworks are deductively derived from abstract principles and subsequently imposed on practical applications. Therefore, the methodology used in this study is highly inductive, beginning with the tools, techniques, and interventions actively employed in the industry and then abstracting from these practical applications to derive broader principles.

5.3 Proportionality and context in explicability

The empirical data also revealed that the elements of explicability operate on a spectrum, with the necessary degree of each principle contingent upon the unique circumstances of each firm, echoing Robbins' (2019) emphasis on a risk-based approach to principle adoption. Thus, the higher levels of explicability, while desirable, are not necessary for all firms. This interpretation aligns with Krishnan's (2019) perspective and the EU AI Act (European Commission, 2021), which advocates for a risk-based approach that offers guidance proportionate to the level of risk involved. The context for explicability is shaped by various factors, including the stakeholders involved, the sector's nature, whether it is public-facing, and the complexity of the model utilised, all collectively influencing the degree to which each principle is required (John-Mathews, 2022). A context-based ethical approach, such as the spectrum system described, aligns with the findings of Hermann (2022), who promotes a utilitarian approach to principle adoption that takes into account the benefits and costs across all stakeholders as opposed to the rigid deontological approach, where each principle should always be maximised. This spectrum of implementation extends to organisational integration as well, where some firms may approach AI ethics primarily as a compliance exercise, while for others, particularly those whose business models directly leverage AI for critical decision-making, ethical considerations must be embedded as a core organisational capability that informs strategic choices and shapes competitive advantage (Solaimani, 2024).

Table 3 Empirically evaluated framework for the operationalisation of AI explicability

	Explicability levels of maturity		
	Repeatable	*Defined*	*Managed*	*Optimised*
Transparency	• Transparency is informal, used reactively, and has instrumental value. (†) • Firms define AI use for stakeholders. (+) • Ad hoc understanding of transparency; minimal user notification practices. (†)	• Accessible and traceable AI processes established; impact on organisational processes is clear. (†) • Internal traceability methods implemented, with lifecycle monitoring of data and AI usage. (~) • Internal traceability ensures employees can access AI usage and impact assessments. (†) • Dataset sampling and algorithm audits initiated. (†) • Stakeholders are identified, and assessments are initiated to evaluate AI's impact systematically. (+)	• Users actively made aware of AI interaction; firm aligns its AI Code of Conduct with transparency goals. (†) • External algorithmic registries and audits are implemented to ensure trust. (†) • Transparency frameworks include algorithmic registries and regular external audits. (+) • Stakeholder feedback incorporated into periodic reviews of transparency practices. (+) • Advanced transparency mechanisms ensure real-time AI outputs are traceable and interpretable dynamically. (+)	• Transparency integrated into strategic goals and aligned with organisational values. (~) • External accreditation is sought, beyond compliance, to ensure business value. (†) • Real-time explainable AI decision-making access provided to stakeholders. (+) • Transparency protocols proactively monitored and accredited. (~) • Public trust is fostered by embedding transparency into innovation strategies. (+) • Transparency practices exceed compliance by providing publicly accessible registries, fostering trust and collaboration. (+)
Interpretability	• Consensus on interpretability definitions is lacking. (†) • AI model complexity minimally understood, limited to basic analyses. (~) • Developers rely on post-hoc tools and basic analysis techniques. (~)	• Governance structures guide dissemination of AI decisions. (†) • Developer training programs implemented to improve understanding of AI outputs. (+) • Governance ensures consistent dissemination of AI decisions throughout the organisation, tailored to different stakeholders (technical, managerial, end-users). (†) • Explainability processes and guidelines are formalised and institutionalised, supported by advanced tools. (†) • Interdisciplinary training aligns perspectives and mitigates risks of bias. (+)	• Multidisciplinary perspectives (e.g., legal, ethical, and technical) integrated to evaluate bias and ensure interpretability. (†) • Advanced explainability tools (e.g., SHAP, LIME, ELI5) systematically applied across models. (~) • Lifecycle monitoring ensures interpretability consistency across models. (+) • Comparisons with industry benchmarks ensure interpretability and validate models against standards. (+)	• Interpretability is embedded into workflows and aligned with strategic goals. (~) • Organisational culture fosters continuous improvement and collaboration across departments to enhance interpretability. (+) • Predictive monitoring systems fully embedded across the AI lifecycle to ensure interpretability. (+) • Best practices actively shared and adopted through collaboration with industry and regulators. (+) • Sandboxes enable proactive, innovative testing of explainability under evolving regulations and scenarios. (+)

Notes: ~ Revised /repositioned item
+ New item added
† Item remained unchanged

Table 3 Empirically evaluated framework for the operationalisation of AI explicability (continued)

	Explicability levels of maturity		
	Repeatable	*Defined*	*Managed*	*Optimised*
Accountability	• Roles and responsibilities for AI accountability are vague and reactive. (~) • Initial risk quantification performed to inform accountability. (+) • Internal processes lack formalisation; compliance awareness remains limited. (~)	• Internal accountability standards established; compliance with external legislation formalised. (†) • AI Officer role defined to oversee accountability processes. (~) • Formal accountability frameworks assign clear roles and responsibilities. (†) • Protocols for redress in unjust AI outcomes are defined and operationalised. (†) • Processes for reporting adverse incidents are formalised, ensuring stakeholders have mechanisms for redress. (~)	• Clear role identification supported by lifecycle audits to enable effective accountability. (+) • Justification of AI decisions and outcomes made available to stakeholders upon request. (†) • Accountability frameworks integrate diverse perspectives, ensuring inclusive AI governance. (+) • Ethical oversight is embedded in governance structures, with regular audits and reviews ensuring compliance and innovation. (+) • Comprehensive justifications for AI outcomes systematically shared with stakeholders. (+)	• C-level management establishes hierarchical accountability structures to ensure top-down responsibility. (~) • Ethics committee proactively governs AI accountability and aligns it with strategic goals. (†) • Ethics committee ensures accountability mechanisms exceed regulatory standards and promote ethical innovation. (+) • Accountability becomes a proactive, organisation-wide commitment. (~) • Organisational accountability frameworks are aligned globally, addressing cross-jurisdictional regulatory challenges. (+)

Notes: ~ Revised /repositioned item
+ New item added
† Item remained unchanged

For example, within transparency, for more mature firms, the onus is on implementing core tools such as algorithmic audits, documentation, and registries, which primarily aim to uphold traceability (Ada Lovelace Institute et al., 2021). For firms with lower AI capability, informing end users of their interaction with AI might be sufficient for upholding transparency (Van den Berg and Kuiper, 2020). Concerning interpretability, the interviews revealed that technically interpreting how models work is plausible for firms by using XAI tools, even at lower maturity levels, yet this should be complemented by developing governance structures that facilitate information sharing among stakeholders, aligning with the delineation made by Van den Berg and Kuiper (2020). For more mature firms, operationalising interpretability entails adopting a multidisciplinary approach, as supported by Tóth et al. (2022), for instance, through cross-expertise training. Such training fosters a deeper understanding of ethical implications across disciplines, thereby facilitating more inclusive and comprehensive decision-making processes. Thus, it is imperative that those tasked with designing regulations and overseeing political matters represent a diverse range of perspectives.

With regard to accountability, firms at all maturity levels found value in appointing a crossed skilled AI officer, particularly to help align the EU AI Act with the heavily regulated finance sector, while for more mature firms, establishing an AI ethics committee is expected, which confirms the findings of Theodorou and Dignum (2020) and Sandler and Basl (2019). At higher maturity levels, multidisciplinary workshops are recommended as a proactive approach to address ethical concerns from various stakeholder perspectives and go beyond governance on paper. Wang (2022) makes the observation that information on AI algorithms can be used by disciplinary groups to further their interests, either by how they frame the explanation or by omitting information:

> "When analysing algorithmic transparency…we should not only disclose the information about how algorithms work but also be alert to the hidden power structures and the way in which the disclosure happens can have profound and far-reaching effects that are often overlooked." (p.69)

The empirical findings of this research suggest that addressing the negative impacts of hidden power structures requires a multidisciplinary approach. This approach can be implemented through various tools, techniques, and interventions, including establishing clear governance structures, forming AI ethics committees, organising workshops, and providing cross-disciplinary training. Table 3 provides a summary of revisions to the earlier proposed maturity framework.

6 Conclusions

The rise of AI is significantly disrupting legal and societal norms, forcing us to rethink how responsibility is assigned when human control is removed from decision-making. This is particularly relevant in the finance sector, where AI technology introduces both opportunities and risks. Current trends of hyper-automation in areas such as insurance, trading, and credit scoring come with the significant risk of (unintended) discrimination and algorithmic bias (Willems and Hafermalz, 2021). The risk is amplified by concerns around the lock-in effect of new technology, where firms become accustomed to new technology despite its adverse effects and the risk of diminished human critical

reasoning, whereby model hallucinations go undetected because they appear credible at a surface level (Stahl et al., 2023). Given the exponential rise in the utilisation of Large Language Models (LLMs) in 2023, it becomes essential for both regulatory entities and businesses to prioritise the ethical considerations surrounding AI. The introduction of legislation such as the EU AI Act, AI Liability Act, and NIS2 underscores the urgency of addressing these issues effectively (European Commission, 2021).

6.1 Contributions for theory and practice

To combat these emerging challenges in AI adoption, explicability is proposed in the literature as a high-level principle that can help guide ethical AI implementation. It encompasses the need for AI technology to be interpretable, transparent, and supported with appropriate infrastructure to uphold accountability. While the notion of explicability remains definitionally fluid, influenced by evolving regulations, technological trends, societal dynamics, and client preferences (Krishnan, 2019), this study makes several distinct contributions to advance both theory and practice.

First, this research addresses the increasing calls in the literature to move beyond principlism to actionable frameworks (e.g., Floridi et al., 2018; John-Mathews et al., 2022). The study provides the first empirically validated maturity framework that integrates transparency, interpretability, and accountability into a coherent assessment tool tailored to financial institutions. Second, this work makes a theoretical contribution by synthesising previously fragmented approaches to explicability, demonstrating through empirical evidence that these elements are interdependent rather than isolated. This holistic conceptualisation advances the understanding of unbiased AI in general and explicability in particular (Hermann, 2022; Larsson and Heintz, 2020). Third, the research reveals the contextual nature of explicability, showing that appropriate levels of each principle depend on organisational characteristics, model complexity, and risk exposure; a nuanced perspective that refines current theoretical approaches.

Practically speaking, the study lays the path for practitioners, policymakers, auditors, regulatory authorities, and compliance officers to take pragmatic steps toward explicability appropriate for the level of AI maturity needed in their specific context (Hermann, 2022). The framework serves as both an assessment tool and implementation guide, offering concrete interventions that can be tailored to specific needs and circumstances, and used cyclically for ongoing improvement and evolution of capabilities over time (Kazanjian and Drazin, 1990). After all, actions speak louder than words.

6.2 Research limitations

Needless to say, this study bears some limitations that should be acknowledged. First, the narrow focus on the financial sector, while allowing for in-depth industry-specific insights, limits the generalisability of findings to other domains where AI is increasingly deployed. The financial sector's highly regulated nature and specific governance structures may create a context that differs significantly from other industries. Second, while appropriate for the research objectives, the exploratory qualitative approach brings inherent methodological limitations. The relatively small sample size of 16 interviews, though sufficient for reaching theoretical saturation in qualitative research, may not capture the full spectrum of perspectives on explicability. Third, despite efforts to ensure a reasonable heterogeneity of data sources, the geographical distribution of participants

was primarily concentrated in Western contexts (UK, USA, Netherlands, and Sweden). This Western-centric perspective may not account for important global cultural, regulatory, and ethical variations in explicability requirements. Finally, the rapid evolution of AI technologies, particularly the emergence of generative AI since data collection, creates a temporal limitation that may affect the long-term applicability of some specific operationalisation recommendations, even as the core framework remains relevant.

6.3 Implications for future research

Several promising avenues for future research emerge from this study. First, extending the explicability maturity framework to other sectors, such as healthcare, could validate its broader applicability and identify sector-specific adaptations needed (Arora et al., 2023; Shah et al., 2024). Healthcare presents particularly interesting opportunities for comparative analysis, given its similarly stringent regulatory environment but different stakeholder dynamics and risk profiles. Second, future studies could employ quantitative or mixed-method approaches to statistically validate the suggested interventions across larger samples, potentially developing standardised assessment instruments based on the maturity framework proposed here (Prabhu, 2020). This could enable benchmarking and comparative analysis across organisations and sectors (e.g., Zand et al., 2015).

Third, identifying the critical success factors (CSFs) for implementing the explicability maturity framework would provide valuable guidance for practitioners. Research could examine which organisational, technological, and environmental factors most significantly influence successful transitions between maturity levels. Focus on CSFs is widely adopted across many topics (e.g., Bullen and Rockart, 1981; Rasuli et al., 2016; Solaimani et al., 2013), and in this specific context, it would help organisations prioritise their efforts and resources when operationalising explicability principles, ultimately increasing the practical impact of the framework.

Fourth, conducting cross-cultural research in different global locations would enrich the understanding of explicability and its operationalisation in diverse regulatory environments. Particularly valuable would be research in rapidly developing AI ecosystems such as China and India, where different approaches to AI governance are emerging (Li et al., 2021). Such research could identify cultural variations in transparency expectations, accountability mechanisms, and interpretability requirements. Fourth, integrating explicability more explicitly into broader AI adoption frameworks represents another promising direction. Building on existing adoption models (Solaimani and Swaak, 2023; Solaimani et al., 2024), researchers could examine how explicability considerations influence adoption decisions, implementation strategies, and success measures. This connection would be particularly valuable for understanding how ethical considerations shape organisational strategies for AI deployment.

Finally, exploring the implications of explicability for emerging technologies such as generative AI could yield valuable insights. Future research might investigate how the rapid evolution of GenAI capabilities affects business models (Dabestani et al., 2025; Kanbach et al., 2024) and how explicability principles can be effectively integrated within firms' business processes to ensure responsible innovation (Rosemann et al., 2024). As AI capabilities continue to advance, explicability frameworks need to evolve accordingly, suggesting the need for longitudinal studies examining the dynamic

relationship between AI capabilities, explicability requirements, and effective governance mechanisms.

Declarations

All authors declare that they have no conflicts of interest.

References

Ada Lovelace Institute, AI Now Institute and Open Government Partnership (2021) *Algorithmic Accountability for the Public Sector* [online] https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/ (accessed 21 April 2025).

Adadi, A. and Berrada, M. (2018) 'Peeking inside the black box: a survey on explainable artificial intelligence (XAI)', *IEEE Access*, Vol. 6, pp.52138–52160, DOI: 10.1109/ACCESS.2018.2870052.

Adams, W. (2015) 'Conducting semi-structured interviews', in Newcomer, K., Hatry, H. and Wholey, J. (Eds.): *Handbook of Practical Program Evaluation*, 4th ed., pp.492–505, John Wiley & Sons, Hoboken, NJ.

Adekunle, S.A., Aigbavboa, C., Ejohwomu, O., Ikuabe, M. and Ogunbayo, B. (2022) 'A critical review of maturity model development in the digitisation era', *Buildings*, Vol. 12, No. 6, p.858, DOI: 10.3390/buildings12060858.

Ahmadi, T. and Solaimani, S. (2021) 'Past and future of demand forecasting models', in Hemmati, M. and Sajadieh, M.S. (Eds.): *Influencing Customer Demand*, CRC Press, Boca Raton, FL, ISBN 9781003107446.

American Institute of Certified Public Accountants and Canadian Institute of Chartered Accountants (2011) *AICPA/CICA Privacy Maturity Model* [online] https://vvena.nl/wp-content/uploads/2018/04/aicpa_cica_privacy_maturity_model.pdf (accessed 21 April 2025).

Ananny, M. and Crawford, K. (2018) 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability', *New Media & Society*, Vol. 20, No. 3, pp.973–989, DOI: 10.1177/1461444816676645.

Arora, A., Barrett, M., Lee, E., Oborn, E. and Prince, K. (2023) 'Risk and the future of AI: algorithmic bias, data colonialism, and marginalization', *Information & Organization*, Vol. 33, No. 3, p.100478, DOI: 10.1016/j.infoandorg.2023.100478.

Arthur, W.B. (1989) 'Competing technologies, increasing returns, and lock-in by historical events', *The Economic Journal*, Vol. 99, No. 394, pp.116–131, DOI: 10.2307/2234208.

Baker-Brunnbauer, J. (2020) 'Management perspective of ethics in artificial intelligence', *AI & Ethics*, Vol. 1, No. 2, pp.173–181, DOI: 10.1007/s43681-020-00024-5.

Bankins, S. and Formosa, P. (2023) 'The ethical implications of artificial intelligence (AI) for meaningful work', *Journal of Business Ethics*, Vol. 185, pp.725–740, DOI: 10.1007/s10551-022-05041-2.

Bartlett, R., Morse, A., Stanton, R. and Wallace, N. (2022) 'Consumer lending discrimination in the FinTech era', *Journal of Financial Economics*, Vol. 143, No. 1, pp.30–56, DOI: 10.1016/j.jfineco.2021.04.030.

Belenguer, L. (2022) 'AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry', *AI & Ethics*, Vol. 2, No. 4, pp.771–787, DOI: 10.1007/s43681-022-00178-1.

Biran, O. and Cotton, C. (2017) *Explanation and Justification in Machine Learning: A Survey* [online] http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf (accessed 21 April 2025).

Bouasabah, M. (2024) 'Analysis of machine learning's performance in stock market prediction, compared to traditional technical analysis indicators', *International Journal of Data Analysis Techniques and Strategies*, Vol. 16, No. 1, pp.32–46, DOI: 10.1504/IJDATS.2024.10061891.

Brinkmann, S. and Kvale, S. (2009) *Interviews: Learning the Craft of Qualitative Research Interviewing*, Sage Publications, Thousand Oaks, CA.

Brooks, J. and King, N. (2014) 'Doing template analysis: evaluating an end-of-life care service', in *SAGE Research Methods Cases*, Part 1, SAGE Publications, Thousand Oaks, CA.

Bullen, C.V. and Rockart, J.F. (1981) *A Primer on Critical Success Factors*, No. 69, pp.1220–1281, Center for Information Systems Research Sloan School of Management, MIT, Boston, USA.

Capability Maturity Model Integration (CMMI) (2002) *CMMI for Systems Engineering, Software Engineering, Integrated Product and Process Development, and Supplier Sourcing (Version 1.1)*, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

Castelvecchi, D. (2016) 'Can we open the black box of AI?', *Nature*, Vol. 538, No. 7623, pp.20–23, DOI: 10.1038/538020a.

Chandani, A. and Bhatia, A. (2025) 'Are robo advisors robbing financial advisors? A systematic literature review paper', *International Journal of Business and Globalisation*, Vol. 39, No. 1, pp.22–43, DOI: 10.1504/IJBG.2025.143577.

Chen, T-H. (2020) 'Do you know your customer? Bank risk assessment based on machine learning', *Applied Soft Computing*, Vol. 86, p.105779, DOI: 10.1016/j.asoc.2019.105779.

Colmenarejo, A., Nannini, L., Rieger, A., Scott, K.M., Zhao, X., Patro, G.K., Kasneci, G. and Kinder-Kurlanda, K. (2022) 'Fairness in agreement with European values', *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, DOI: 10.1145/3527613.3532478.

Conaty, F. (2021) 'Abduction as a methodological approach to case study research in management accounting – an illustrative case', *Accounting, Finance & Governance Review*, Vol. 27, DOI: 10.52399/001c.22171.

Constantinides, P., Monteiro, E. and Mathiassen, L. (2024) 'Human-AI joint task performance: Learning from uncertainty in autonomous driving systems', *Information & Organization*, Vol. 34, No. 2, p.100502, DOI: 10.1016/j.infoandorg.2024.100502.

Council of the European Union (2000) *Council Directive 2000/43/EC of 29 June 2000: Implementing the Principle of Equal Treatment between Persons Irrespective of Racial or Ethnic Origin*, *Official Journal of the European Communities*.

Cunha, S.L.d.R., Costa, A., Gonçalves, R., Pereira, L., Dias, Á.R.V. and Silva, A. (2023) 'Smart systems adoption in management', *International Journal of Business and Systems Research*, Vol. 17, No. 6, pp.703–727, DOI: 10.1504/IJBSR.2023.134465.

Dabestani, R., Solaimani, S., Ajroemjan, G. and Koelemeijer, K. (2025) 'Exploring the enablers of data-driven business models: a mixed-methods approach', *Technological Forecasting and Social Change*, Vol. 213, p.124036, DOI:10.1016/j.techfore.2025.124036.

Daly, J., Willis, K., Small, R., Green, J., Welch, N., Kealy, M. and Hughes, E. (2007) 'A hierarchy of evidence for assessing qualitative health research', *Journal of Clinical Epidemiology*, Vol. 60, No. 1, pp.43–49, DOI: 10.1016/j.jclinepi.2006.03.014.

Dastin, J. (2018) 'Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters* [online] https://www.reuters.com/technology/amazon-scraps-secret-ai-recruiting-tool-2018-10-10/ (accessed 21 April 2025).

Deloitte Netherlands (2023) *The NIS2 Directive* [online] https://www2.deloitte.com/nl/risk/articles/the-nis2-directive.html (accessed 21 April 2025).

Dewey, J. (2008) *Ethics* (orig. 1908), SIU Press, Carbondale, USA.

Dignum, V. et al. (2018) 'Ethics by design: necessity or curse?', *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, DOI: 10.1145/3278721.3278745.

European Commission (2021) *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, Brussels.

European Commission (2023) *Artificial Intelligence Liability Directive* [online] https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI_739342_EN pdf (accessed 21 April 2025).

European Commission (2024) *Data Governance Act Explained* [online] https://digital-strategy.ec.europa.eu/en/policies/data-governance-act-explained (accessed 21 April 2025).

Faraj, S., Pachidi, S. and Sayegh, K. (2018) 'Working and organizing in the age of the learning algorithm', *Information & Organization*, Vol. 28, No. 1, pp.62–70, DOI: 10.1016 /j.infoandorg.2018.02.005.

Firth, N. (2021) 'Apple Card is being investigated over claims it gives women lower credit limits', *MIT Technology Review* [online] https://www.technologyreview.com/2021/03/12/1020316 /apple-card-gender-bias-investigation/ (accessed 21 April 2025).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. and Vayena, E. (2018) 'AI4People – an ethical framework for a good AI society: opportunities, risks, principles, and recommendations', *Minds & Machines*, Vol. 28, No. 4, pp.689–707, DOI: 10.1007/s11023-018-9482-5.

Gegenhuber, T., Mair, J., Lührsen, R. and Thäter, L. (2023) 'Orchestrating distributed data governance in open social innovation', *Information & Organization*, Vol. 33, No. 1, p.100453, DOI: 10.1016/j.infoandorg.2023.100453.

Gita, P.C. and Krishnakumar, S. (2024) 'The ripple effect of organisational inclusiveness on perception of ethical climate – an empirical investigation', *International Journal of Business Governance and Ethics*, Vol. 18, No. 1, pp.84–103, DOI: 10.1504/IJBGE.2022.10051609.

Gkeredakis, M., Lifshitz-Assaf, H. and Barrett, M. (2021) 'Crisis as opportunity, disruption and exposure: Exploring emergent responses to crisis through digital technology', *Information & Organization*, Vol. 31, No. 1, p.100344, DOI: 10.1016/j.infoandorg.2021.100344.

Glaser, B. and Strauss, A. (1999) *Discovery of Grounded Theory: Strategies for Qualitative Research*, Routledge, London.

Glavina, S. (2024) 'AI in financial industry: ethic issues', *Journal of Trends and Challenges in Artificial Intelligence*, Vol. 1, No. 1, pp.15–20, DOI: 10.61552/JAI.2024.01.002.

Guest, G., Bunce, A. and Johnson, L. (2006) 'How many interviews are enough?', *Field Methods*, Vol. 18, No. 1, pp.59–82, DOI: 10.1177/1525822X05279903.

Gujar, S.N. et al. (2025) 'Brain tumour detection and multi-classification using GNB-based machine learning architecture', *International Journal of Data Analysis Techniques and Strategies*, Vol. 17, No. 1, pp.20–35, DOI: 10.1504/IJDATS.2025.144963.

Hennink, M. and Kaiser, B.N. (2022) 'Sample sizes for saturation in qualitative research: a systematic review of empirical tests', *Social Science & Medicine*, Vol. 292, p.114523, DOI: 10.1016/j.socscimed.2021.114523.

Hermann, E. (2022) 'Leveraging artificial intelligence in marketing for social good – an ethical perspective', *Journal of Business Ethics*, Vol. 179, pp.43–61, DOI: 10.1007/s10551-021-04843-y.

High-Level Expert Group on Artificial Intelligence (2018) *A Definition of AI: Main Capabilities and Scientific Disciplines*, European Commission [online] https://ec.europa.eu/futurium /en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf (accessed 21 April 2025).

High-Level Expert Group on Artificial Intelligence (2019) *Ethics Guidelines for Trustworthy AI* [online] European Commission [online] https://ec.europa.eu/digital-single-market/en/news /ethics-guidelines-trustworthy-ai (accessed 21 April 2025).

Hunkenschroer, A.L. and Luetge, C. (2022) 'Ethics of AI-enabled recruiting and selection: a review and research agenda', *Journal of Business Ethics*, Vol. 178, pp.977–1007, DOI: 10.1007 /s10551-022-05049-6.

Jackman, H. (2020) 'Meaning holism', *Stanford Encyclopedia of Philosophy* [online] https://plato.stanford.edu/entries/meaning-holism (accessed 21 April 2025).

John-Mathews, J.M. (2022) 'Some critical and ethical perspectives on the empirical turn of AI interpretability', *Technology Forecasting & Social Change*, Vol. 174, 121209, DOI: 10.1016/j.techfore.2021.121209.

John-Mathews, J.M., Cardon, D. and Balagué, C. (2022) 'From reality to world: a critical perspective on AI fairness', *Journal of Business Ethics*, Vol. 178, pp.945–959, DOI: 10.1007/s10551-022-05055-8.

Kanbach, D.K., Heiduk, L., Blueher, G., Schreiter, M. and Lahmann, A. (2024) 'The GenAI is out of the bottle: Generative artificial intelligence from a business model innovation perspective', *Review of Managerial Science*, Vol. 18, No. 4, pp.1189–1220, DOI: 10.1007/s11846-023-00696-z.

Kazanjian, R.K. and Drazin, R. (1990) 'A stage-contingent model of design and growth for technology-based new ventures', *Journal of Business Venturing*, Vol. 5, No. 3, pp.137–150, DOI: 10.1016/0883-9026(90)90028-R.

Korneeva, E., Salge, T.O., Teubner, T. and Antons, D. (2023) 'Tracing the legitimacy of artificial intelligence: a longitudinal analysis of media discourse', *Technology Forecasting & Social Change*, Vol. 192, p.122467, DOI: 10.1016/j.techfore.2023.122467.

Krippendorff, K. and Craggs, R. (2016) 'The reliability of multi-valued coding of data', *Communication Methods & Measures*, Vol. 10, No. 4, pp.181–198, DOI: 10.1080/19312458.2016.1228863.

Krishnan, M. (2019) 'Against interpretability: a critical examination of the interpretability problem in machine learning', *Philosophy & Technology*, Vol. 33, No. 3, pp.487–502, DOI: 10.1007/s13347-019-00372-9.

Kubica, M.L. (2022) 'Autonomous vehicles and liability law', *American Journal of Comparative Law*, Vol. 70, No. 1, pp.39–69.

Kuiper, O., van den Berg, M., van den Burgt, J. and van Leijnen, S. (2021) 'Exploring explainable AI in the financial sector: Perspectives of banks and supervisory authorities', in *Proceedings of the Benelux Conference on Artificial Intelligence*, Springer, Cham., pp.105–119, DOI: 10.1007/978-3-030-93842-0_6.

Lainhart, J., Conboy, M. and Saull, R. (2019) 'COBIT 2019 framework: introduction and methodology', *ISACA Online Forum*.

Larsson, S. and Heintz, F. (2020) 'Transparency in artificial intelligence', *Internet Policy Review*, Vol. 9, No. 2, DOI: 10.14763/2020.2.1469.

Lehmann, R.J. (2021) 'Why 'big data' will force insurance companies to think hard about race', *Insurance Journal* [online] https://www.insurancejournal.com/blogs/right-street/2018/03/27/484530.htm (accessed 21 April 2025).

Li, D., Tong, T. and Xiao, Y. (2021) 'Is China emerging as the global leader in AI?', *Harvard Business Review* [online] https://hbr.org/2021/02/is-china-emerging-as-the-global-leader-in-ai.

Li, S., Garces, E. and Daim, T. (2019) 'Technology forecasting by analogy based on social network analysis: the case of autonomous vehicles', *Technology Forecasting & Social Change*, Vol. 148, p.119731, DOI: 10.1016/j.techfore.2019.119731.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. and Zou, J. (2023) 'GPT detectors are biased against non-native English writers', *Patterns*, Vol. 4, No. 7, p.100779, DOI: 10.1016/j.patter.2023.100779.

Martin, K. (2019) 'Ethical implications and accountability of algorithms', *Journal of Business Ethics*, Vol. 160, pp.835–850, DOI: 10.1007/s10551-018-3921-3.

Martineau, K. (2023) 'What is generative AI?', *IBM Research Blog* [online] https://research.ibm.com/blog/what-is-generative-ai (accessed 21 April 2025).

Meske, C., Bunde, E., Schneider, J. and Gersch, M. (2022) 'Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities', *Information Systems Management*, Vol. 39, No. 1, pp.53–63, DOI: 10.1080/10580530.2020.1849465.

Metz, C. and Schmidt, G. (2023) 'Elon Musk and others call for a pause on AI, citing 'profound risks to society'', *New York Times* [online] https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html (accessed 21 April 2025).

Miles, M.B. and Huberman, A.M. (1994) *Qualitative Data Analysis: An Expanded Sourcebook*, 2nd ed., SAGE Publications, Thousand Oaks, CA.

Monod, E., Mayer, A.S., Straub, D., Joyce, E. and Qi, J. (2024) 'From worker empowerment to managerial control: The devolution of AI tools' intended positive implementation to their negative consequences', *Information & Organization*, Vol. 34, No. 1, 100498, DOI: 10.1016/j.infoandorg.2023.100498.

Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E. and Sicilia, M.A. (2021) 'Traceability for trustworthy AI: a review of models and tools', *Big Data & Cognitive Computing*, Vol. 5, No. 20, DOI: 10.3390/bdcc5020020.

Murphy, K.P. (2012) *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MA.

Nanda, P. and Kumar, V. (2024) 'New generation technologies: development vs. ethical challenges', *International Journal of Business Information Systems*, Vol. 1, No. 1, p.1, DOI: 10.1504/IJBIS.2022.10047594.

NIST (2018) *Risk Management Framework for Information Systems and Organizations*, DOI: 10.6028/NIST.SP.800-37r2.

Ntoutsi, E. et al. (2020) 'Bias in data-driven artificial intelligence systems – an introductory survey', *Wiley Interdisciplinary Reviews: Data Mining & Knowledge Discovery*, Vol. 10, No. 3, DOI: 10.1002/widm.1356.

Ollagnier, J-M. (2024) 'AI meets regulation: Driving innovation within the EU AI Act', *Forbes* [online] https://www.forbes.com/sites/jeanmarcollagnier/ (accessed 21 April 2025).

Omrani, N., Rivieccio, G., Fiore, U., Schiavone, F. and Agreda, S.G. (2022) 'To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts', *Technology Forecasting & Social Change*, Vol. 181, p.121763, DOI: 10.1016/j.techfore.2022.121763.

Oruthotaarachchi, C.R. and Wijayanayake, W.M.J.I. (2023) 'Developing a multi-perspective capability model for organisational business process management maturity assessment in digital era', *International Journal of Information Systems and Change Management*, Vol. 13, No. 3, pp.191–208, DOI: 10.1504/IJISCM.2023.133356.

Pereira, L., Nobre, R., Dias, Á., Costa, A.d.R. and Gonçalves, R. (2024b) 'Artificial intelligence and project management: exploring the contributions and implications', *International Journal of Logistics Systems and Management*, Vol. 47, No. 4, pp.432–467, DOI: 10.1504/IJLSM.2024.138916.

Pesch, U. (2014) 'Engineers and active responsibility', *Science & Engineering Ethics*, Vol. 21, No. 4, pp.925–939, DOI: 10.1007/s11948-014-9571-7.

Prabhu, G. (2020) 'Teaching the scope and limits of generalizability in qualitative research', *New Trends in Qualitative Research*, Vol. 1, pp.186–192, DOI: 10.36367/ntqr.1.2020.186-192.

Prince, A. and Schwarcz, D. (2020) 'Proxy discrimination in the age of artificial intelligence and big data', *Iowa Law Review*, Vol. 105, pp.1257–1318.

Rana, S. (2023) 'AI and GPT for management scholars and practitioners: guidelines and implications', *FIIB Business Review*, Vol. 12, No. 1, pp.7–9, DOI: 10.1177/23197145231161408.

Rasuli, B., Alipour-Hafezi, M. and Solaimani, S. (2016) 'The identification of criticalsuccess factors in the development of national ETDs programs', in *19th International Symposium on Electronic Theses and Dissertations (ETD 2016): 'Data and Dissertations'*, July, Villeneuve d'Ascq, France [online] http://hal.univ-lille3.fr/hal-01396672 (accessed 21 April 2025).

Ratten, V. (2024) 'Artificial intelligence, digital trends and globalization: future research trends', *FIIB Business Review*, DOI: 10.1177/23197145231222774.

Robbins, S. (2019) 'A misdirected principle with a catch: explicability for AI', *Minds & Machines*, Vol. 29, No. 4, pp.495–514, DOI: 10.1007/s11023-019-09509-3.

Rodríguez-Pérez, R. and Bajorath, J. (2020) 'Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions', *Journal of Computer-Aided Molecular Design*, Vol. 34, No. 10, pp.1013–1026, DOI: 10.1007 /s10822-020-00314-0.

Rosemann, M., Brocke, J.V., Van Looy, A. and Santoro, F. (2024) 'Business process management in the age of AI – three essential drifts', *Information Systems & e-Business Management*, DOI: 10.1007/s10257-024-00689-9.

Sandler, R. and Basl, J. (2019) *Building Data and AI Ethics Committees*, Accenture [online] https://www.accenture.com/_acnmedia/PDF-107/Accenture-AI-And-Data-Ethics-Committee-Report-11.pdf (accessed 21 April 2025).

Santoni de Sio, F. and Mecacci, G. (2021) 'Four responsibility gaps with artificial intelligence: Why they matter and how to address them', *Philosophy & Technology*, Vol. 34, No. 4, pp.1057–1084, DOI: 10.1007/s13347-021-00450-x.

Saura, J.R. (2024) 'Algorithms in digital marketing: Does smart personalization promote a privacy paradox?', *FIIB Business Review*, Vol. 13, No. 5, pp.499–502, DOI: 10.1177 /23197145241276898.

Schneider, J., Abraham, R., Meske, C. and Vom Brocke, J. (2023) 'Artificial intelligence governance for businesses', *Information Systems Management*, Vol. 40, No. 3, pp.229–249.

Shah, W.S., Elkhwesky, Z., Jasim, K.M., Elkhwesky, E.F.Y. and Elkhwesky, F.F.Y. (2024) 'Artificial intelligence in healthcare services: Past, present, and future research directions', *Review of Managerial Science*, Vol. 18, No. 3, pp.941–963, DOI: 10.1007/s11846-023-00699-w.

Simonsson, M., Johnson, P. and Ekstedt, M. (2010) 'The effect of IT governance maturity on IT governance performance', *Information Systems Management*, Vol. 27, No. 1, pp.10–24, DOI: 10.1080/10580530903455106.

Solaimani, S. (2024) 'From compliance to capability: on the role of data and technology in environment, social, and governance', *Sustainability*, Vol. 16, No. 14, p.6061. DOI: 10.3390 /su16146061

Solaimani, S. and Swaak, L. (2023) 'Critical success factors in a multi-stage adoption of artificial intelligence: a necessary condition analysis', *Journal of Engineering and Technology Management*, Vol. 69, p.101760, DOI: 10.1016/j.jengtecman.2023.101760.

Solaimani, S., Bouwman, H. and Secomandi, F. (2013) 'Critical design issues for the development of Smart Home technologies', *Journal of Design Research*, Vol. 11, No. 1, pp.72–90, DOI: 10.1504/JDR.2013.054067.

Solaimani, S., Dabestani, R., Harrison-Prentice, T., Ellis, E., Kerr, M., Choudhury, A. and Bakhshi, N. (2024) 'Exploration and prioritisation of critical success factors in adoption of artificial intelligence: a mixed-methods study', *International Journal of Business Information Systems*, Vol. 45, No. 4, pp.429–453, DOI: 10.1504/IJBIS.2024.10048145.

Solaimani, S., van Eck, T., Kievit, H. and Koelemeijer, K. (2022) 'An exploration of the applicability of lean startup in small non-digital firms: An effectuation perspective', *International Journal of Entrepreneurial Behavior & Research*, Vol. 28, No. 9, pp.198–218, DOI: 10.1108/IJEBR-04-2021-0270.

Sonntag, M., Mehmann, S., Mehmann, J. and Teuteberg, F. (2024) 'Development and evaluation of a maturity model for AI deployment capability of manufacturing companies', *Information Systems Management*, DOI: 10.1080/10580530.2024.2319041.

Srivastava, P., Mishra, N., Srivastava, S. and Shivani, S. (2024) 'Banking with chatbots: the role of demographic and personality traits', *FIIB Business Review*, DOI: 10.1177 /23197145241227757.

Stahl, B.C. (2012) 'Morality, ethics, and reflection: a categorization of normative IS research', *Journal of the Association for Information Systems*, Vol. 13, No. 8, Article 3, DOI: 10.17705/1jais.00304.

Stahl, B.C., Brooks, L., Hatzakis, T., Santiago, N. and Wright, D. (2023) 'Exploring ethics and human rights in artificial intelligence – a Delphi study', *Technological Forecasting & Social Change*, Vol. 191, p.122502, DOI: 10.1016/j.techfore.2023.122502.

Sullivan, Y.W. and Fosso Wamba, S. (2022) 'Moral judgments in the age of artificial intelligence', *Journal of Business Ethics*, Vol. 178, No. 4, pp.917–943, DOI: 10.1007/s10551-022-05053-w.

Tarigan, M.K., Simatupang, T.M. and Bangun, Y.R. (2025) 'Building resilience through digital transformation: a systematic literature review and comprehensive framework for large enterprises', *International Journal of Business Innovation and Research*, Vol. 36, No. 5, DOI: 10.1504/IJBIR.2025.10070413.

Tavory, I. and Timmermans, S. (2014) *Abductive Analysis: Theorizing Qualitative Research*, University of Chicago Press, Chicago.

Theodorou, A. and Dignum, V. (2020) 'Towards ethical and socio-legal governance in AI', *Nature Machine Intelligence*, Vol. 2, No. 1, pp.10–12, DOI: 10.1038/s42256-019-0136-y.

Tóth, Z., Caruana, R., Gruber, T. and Loebbecke, C. (2022) 'The dawn of the AI robots: towards a new framework of AI robot accountability', *Journal of Business Ethics*, Vol. 178, pp.895–916, DOI: 10.1007/s10551-022-05050-z.

Townson, S. (2020) 'AI can make bank loans more fair', *Harvard Business Review* [online] https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair.

Uddin, M.S. and Chowdhury, M. (2024) 'A user-friendly anger and anxiety disorder prediction scheme using machine learning and a mobile application for mental healthcare', *International Journal of Data Analysis Techniques and Strategies*, Vol. 16, No. 1, pp.47–81, DOI: 10.1504/IJDATS.2024.137466.

Van de Poel, I. (2013) 'Translating values into design requirements', in Michelfelder, D.P., McCarthy, N. and Goldberg, D.E. (Eds.): *Philosophy and Engineering: Reflections on Practice, Principles, and Process*, pp.253–266, Springer, Dordrecht.

Van den Berg, M. and Kuiper, O.X. (2020) *XAI in the Financial Sector* [online] https://www.internationalhu.com/research/projects/explainable-ai-in-the-financial-sector (accessed 21 April 2025).

Vishwarupe, V., Joshi, P.M., Mathias, N., Maheshwari, S., Mhaisalkar, S. and Pawar, V. (2022) 'Explainable AI and interpretable machine learning: a case study in perspective', *Procedia Computer Science*, Vol. 204, pp.869–876, DOI: 10.1016/j.procs.2022.08.105.

Wade, L. (2010) 'HP software doesn't see black people', *Sociological Images* [online] https://thesocietypages.org/socimages/2010/01/05/hp-software-doesnt-see-black-people/ (accessed 21 April 2025).

Wang, H. (2022) 'Transparency as manipulation? Uncovering the disciplinary power of algorithmic transparency', *Philosophy & Technology*, Vol. 35, Article 23, DOI: 10.1007/s13347-022-00564-w.

Wehrli, S., Hertweck, C., Amirian, M., Glüge, S. and Stadelmann, T. (2021) 'Bias, awareness, and ignorance in deep-learning-based face recognition', *AI & Ethics*, Vol. 2, No. 3, pp.509–522, DOI: 10.1007/s43681-021-00108-6.

Willems, T. and Hafermalz, E. (2021) 'Distributed seeing: algorithms and the reconfiguration of the workplace, a case of 'automated' trading', *Information & Organization*, Vol. 31, No. 4, p.100376, DOI: 10.1016/j.infoandorg.2021.100376.

Zand, F., Solaimani, S. and van Beers, C. (2015) 'A role-based typology of information technology: model development and assessment', *Information systems management*, Vol. 32, No. 2, pp.119–135, DOI: 10.1080/10580530.2015.1018770.

Notes

1 Explicability is closely linked with Explainable AI (xAI) and these two expressions are often used interchangeably. However, explicability is chosen because it captures the more nuanced concepts of accountability and responsibility.

2 Post-hoc analysis refers to the process of reverse engineering outcomes of sophisticated AI to identify explainable decision features (Adadi and Berrada, 2018).

3 In the proposed framework the level of ad hoc is omitted, this is because it represents an undesirable state.

4 Full interview protocol available upon request.

5 The tool SHAP (abbreviation for Shapley additive explanation method) is an "approach [that] enables the identification and prioritisation of features that determine compound classification and activity prediction" (Rodríguez-Pérez and Bajorath, 2020). This tool is rapidly growing in popularity due to its success in helping data scientists to interpret deep neural networks.

Appendix

Figure A1 The relationships and interdependencies between the constituents of explicability