Closing the Gaps on Inscrutability: Tackling Challenges with Knowledge Integration during AI development

Tapani Rinta-Kahila
Business School
The University of Queensland
Brisbane, Australia
Hanken School of Economics
Helsinki, Finland
Email: t.rintakahila@uq.edu.au

Ida Asadi Someh
Business School
ARC Training Centre for Information Resilience
The University of Queensland
Brisbane, Australia

Ali Darvishi
Office of the Deputy Vice-Chancellor (Academic)
The University of Queensland
Brisbane, Australia

Reihaneh Bidar
Business School
The University of Queensland
Brisbane, Australia

Marta Indulska
Business School
ARC Training Centre for Information Resilience
The University of Queensland
Brisbane, Australia

Abstract

The development of complex artificial intelligence (AI) systems presents a compelling knowledge integration challenge to organisations. As the organisations strive to integrate complex domain knowledge into algorithmic models, they also have to arm domain experts with the technical understanding of how such models work so they can be used responsibly. The inscrutability of AI technology – stemming from challenges related to both the technical explainability of the models as well as their social interpretability – makes knowledge integration particularly challenging by creating and deepening knowledge gaps between the AI model, its human users and domain reality. To increase understanding of how such knowledge gaps can be addressed in AI development, this study reports on three qualitative case studies on AI projects faced where inscrutability needed to be managed. Building on the gap model (Kayande et al., 2009), we identify three sociotechnical mechanisms for addressing knowledge gaps related to AI inscrutability and thus facilitating organisational learning. Our work provides contributions to both theory and practice.

Keywords: Artificial intelligence, Machine learning, Explainability, Inscrutability, Case study.

Received 27 Nov 2024, accepted 12 December 2025 after 2 rounds of review with A. Techatassanasoontorn as Senior Editor and M. Davern and S. Black as accepting Editors-in-Chief

1 Introduction

Public and private organisations are increasingly motivated to adopt artificial intelligence (AI), systems with a machine-learning model at their core, to improve their internal processes and customer-facing services (Business Wire, 2023). AI models have the potential to create new knowledge, optimise processes and mitigate human biases, possibly resulting in increased fairness and accuracy of organisational decision-making (Shollo et al., 2022). However, developing accurate AI models that are representative of the real world is far from straightforward. In addition to the technical work of data scientists to train machine learning (ML) models on available data, it requires the integration of deep and diverse domain knowledge (e.g., codified data, experiential and contextual insights) into those models (van den Broek et al., 2021). In this light, AI models are constructed through the synthesis of inputs from multiple disciplines and stakeholders (Asatiani et al. 2021; 2020). Yet, in practice, much of this knowledge might not be readily available to data scientists or may be hard to codify and incorporate into models. At the same time, domain experts tend to have limited technical skills and AI literacy, which can result in ineffective use or outright rejection of AI models. These tensions may lead to persistent knowledge gaps that can undermine the representativeness, accuracy and reliability of AI systems.

Research on decision-support systems (Kayande et al., 2009; Martens & Provost, 2014) suggests that ensuring high decision-support system accuracy and widespread user acceptance requires aligning (1) the model at the core of the intelligent system, (2) the mental models of its human users (e.g., domain experts, managers, system developers), (3) and the domain reality of their organisation. Gaps between these three "models" (i.e., knowledge gaps) can erode performance, e.g., when AI models fail to capture domain reality, thus performing

inadequately, or when the user's mental model does not align with that of the AI model, thus hampering user acceptance. When it comes to contemporary AI applications, inscrutability, defined as "deficiencies in the intelligibility of AI procedures and outputs in relation to a specific party" (Berente et al., 2021, p. 1441), exacerbates this challenge. AI inscrutability has both technical and social dimensions. Technically, the complexity of AI models can render their inner workings opaque, making it difficult or impossible to explain how the model produces outputs from inputs (i.e., the explainability challenge). Socially, even if such an explanation can be provided, it may not be accessible or interpretable to stakeholders with limited technical expertise, such as domain experts (i.e., the interpretability challenge). So, how can alignment between AI models, different users' mental models and domain reality be achieved if AI models operate in an opaque manner that may be fundamentally different from how their users think?

Ensuring sufficient explainability and considering different stakeholders' informational needs and perspectives is crucial for ensuring that AI systems are reliable, safe and trustworthy (Shneiderman, 2020). Failure to do this can result in insufficiently accurate AI models that are risky to use (Lebovitz et al., 2021). Furthermore, domain experts have been found to dismiss potentially helpful insights from AI systems when the systems are opaque and do not offer explanations for their outputs (Allen & Choudhury, 2022; Lebovitz et al., 2022), resulting in

missed opportunities to integrate knowledge. From the legal perspective, regulatory bodies may prohibit and penalise AI use due to inscrutability: an AI system deployed by the Dutch government was deemed unlawful by courts for being opaque and thus failing to satisfy the right for individuals to access meaningful explanations on decisions that affect their lives guaranteed by the European Union's GDPR legislation (Akbarighatar et al., 2025).

Knowledge integration is crucial for closing such knowledge gaps when developing AI systems, and this calls for organisations to commit to a sociotechnical learning process where AI models and the organisations within which they are developed and used, learn, adapt and improve iteratively in tandem (van den Broek et al., 2021; Rinta-Kahila et al. 2023a). As we set out to understand how to successfully navigate this learning process, we ask: *How does knowledge integration occur in the development and implementation of inscrutable AI systems, and what is the role of explanations in this process?* We approach this challenge by drawing on the gap model (Kayande et al., 2009; Martens & Provost, 2014), which posits that explanations can be used to bridge gaps in understanding or representation between an intelligent decision-support system, its human user, and reality. Our adaptation of this model as a lens through which we analyse three AI case studies shows how, through the process of 'explaining' both AI and domain knowledge, organisations can overcome these knowledge gaps that would otherwise hamper the success of their AI models.

We provide implications for theory and practical guidelines for organisations that develop and implement complex AI systems. In doing so, we respond to calls for a better understanding of AI inscrutability (Ågerfalk et al., 2022; Bauer et al., 2021; Berente et al., 2021) by presenting explanations as an approach for engaging with diverse social stakeholders and integrating their knowledge while developing and implementing AI. This process, aimed at aligning AI models with stakeholder-specific perspectives and knowledge, encourages multiple iterations and feedback loops in which AI models' learning is contrasted with the knowledge possessed by domain experts and domain reality.

2 Theoretical underpinnings

Here, we introduce the theoretical concepts that guide our study. We begin by positioning AI development as an Information Systems Development (ISD) process, which emphasizes the integration of diverse knowledge sources across technical and organisational domains. We then discuss how two key dimensions of AI inscrutability complicate knowledge integration during AI development. Finally, we introduce the three-gap model (Kayande et al., 2009) as a framework for understanding how AI inscrutability stimulates misalignments between AI models, users' mental models, and domain reality.

2.1 AI development as a knowledge integration process

AI development can be viewed as a form of ISD, which has long emphasized the integration of diverse forms of knowledge throughout the system lifecycle. Matook et al.'s (2021) curation of important research on this topic positions ISD as an inherently sociotechnical process involving iterative collaboration among stakeholders. In this process, knowledge integration plays a central role in shaping system design, functionality and performance. This research stream has established that effective ISD requires technical expertise along with the ability to incorporate domain-specific knowledge, user requirements, and relevant organisational

context into the system. Hahn and Lee (2021) reinforce this view by showing that cross-domain knowledge enhances ISD performance, particularly in complex projects. Their study demonstrates that teams with richer cross-domain knowledge are better equipped to navigate interdependencies in design decisions, leading to more robust and reliable outcomes. These findings align with broader ISD literature, such as Faraj and Sproull (2000), who emphasize the importance of coordinating expertise across domains, and Tiwana (2009), who shows that effective knowledge integration improves decision-making and system relevance.

At the core of AI development lies the construction and refinement of algorithmic models that encode knowledge about the world. These models are not static representations, but dynamic artifacts shaped through iterative integration of knowledge from diverse sources, including data, domain expertise and stakeholder input (Someh et al., 2023). Drawing on Grant's (1996) knowledge-based theory of the firm, we view AI models as organisational knowledge assets whose quality depends on the effective coordination and integration of specialized knowledge. Building high-performing models requires substantial effort in collaboration and feedback loops, where stakeholders contribute both requirements and contextual insights that guide model improvement. As such, AI development inherits and extends the collaborative, knowledge-intensive nature of ISD (Laato et al., 2024; Someh et al., 2023).

Where AI development diverges from traditional ISD is in its focus on building models that learn from data autonomously through the machine learning process, rather than being explicitly programmed. While traditional ISD projects produce rule-based systems with transparent logic, AI systems are often probabilistic and opaque, making their logic difficult to interpret even for their creators (Berente et al., 2021). This inscrutability introduces distinct challenges for knowledge integration and, thus, for building AI models that are accurate and accepted by their stakeholders.

2.2 AI inscrutability

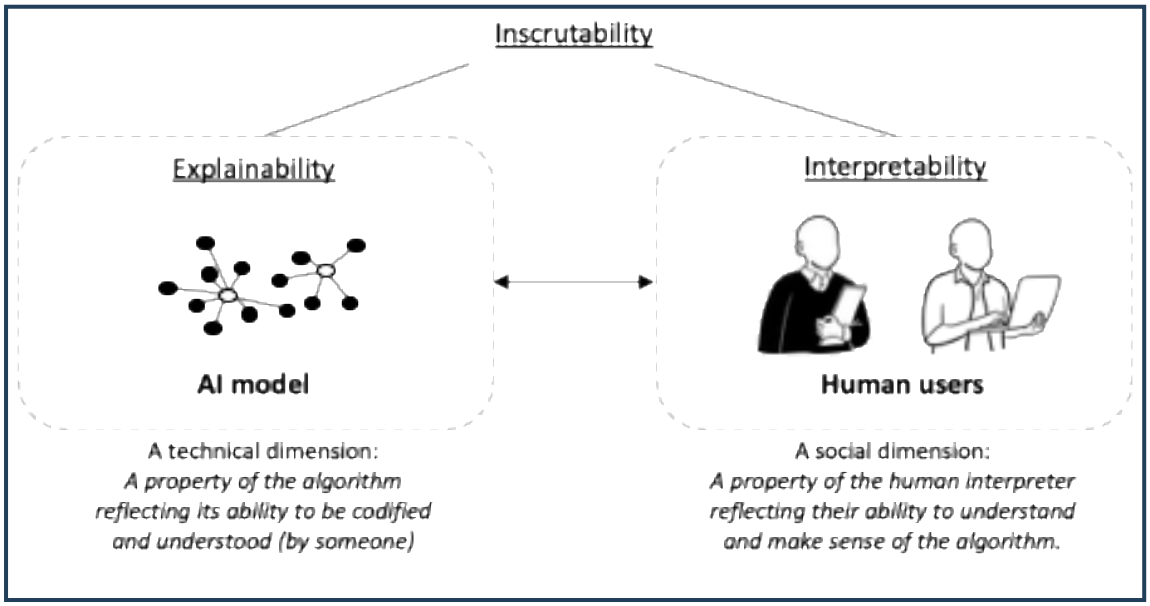

Figure 1: Different aspects of AI inscrutability

According to Berente et al. (2021, p. 1441), the key aspects of inscrutability include explainability (conversely, opacity) and interpretability (see Figure 1)[1]. While explainability is a property of the AI model that reflects its technical traceability, interpretability captures a social dimension of inscrutability by reflecting human stakeholders' ability to understand and make sense of the model's outputs.

AI explainability refers to "an algorithm's ability to be codified and understood at least by some party" (Berente et al., 2021, p. 1441). At the core of any AI system is a model, an algorithm trained on data to mimic human decision-making (Russell & Norvig, 2010). The model is an abstract representation of reality that is typically tasked to predict domain-specific outcomes (e.g., patients at risk of sepsis). Unlike rule-based systems, which rely on human-defined rules, AI models derive decision rules from data with minimal human input. Deep-learning models, for instance, learn autonomously from data and propagate their learning across layers, resulting in highly complex structures. Such models suffer from lack of explainability, or opacity, defined as "the lack of visibility into an algorithm", relating to the fact that "the logic of some advanced algorithms is simply not accessible, even to the developers of those algorithms" (Berente et al. 2021, p. 1441). The inability to explain how an AI model produces an output can make knowledge integration difficult and hamper the model's acceptance among domain experts.

In response, a body of research has emerged to address the explainability challenge from a technical standpoint (Guidotti et al., 2018; Lipton, 2018). While complete explainability of advanced AI models cannot always be achieved, the emerging literature has introduced strategies for making decisions more traceable. One approach involves using inherently explainable models, often called white box models, like decision trees and linear regression. Although these models tend to offer lower performance than opaque black-box models (Asatiani et al., 2021), they offer clear, traceable logic. A second strategy involves pairing white-box models with black-box models to illuminate the inner workings of complex systems. Last, *post hoc* techniques like LIME (Local Interpretable Model-agnostic Explanations) have been developed to explain the predictions of complex black-box models. LIME works by creating a simpler model that approximates the behaviour of the black-box model in the local vicinity of a single prediction (Ribeiro et al., 2016).

Regulatory bodies are increasingly demanding a degree of explainability from AI systems that impact humans' lives. For example, the Australian government's AI ethics principles call for "reasonable justifications for AI systems outcomes" (Australian Department of Industry Science and Resources, 2024). Similarly, the European Union's AI Act and General Data Protection Regulation (GDPR), give individuals the right to receive *meaningful* explanations on decisions that affect their lives (Goodman & Flaxman, 2017). This notion of meaningfulness invites us to consider how stakeholders understand and make sense of AI's outputs.

AI interpretability captures the social aspect of inscrutability, as it refers to a particular stakeholder's ability to understand what the AI system is doing (Berente et al., 2021). Different stakeholders are equipped with varying levels of domain and technical skills, and thus an

[1]*Berente et al. (2021) also identify transparency as a dimension of inscrutability and define it as a strategic consideration regarding the extent to which the organisation using an AI system chooses to disclose information about that system. This dimension remains outside the scope of our research.*

algorithm that is interpretable to one stakeholder may not be so to another. Indeed, Bauer et al. (2021) note that "the majority of current designs meet their developers' demands but not their ultimate users' demands, who are typically domain, yet no technical experts" (p. 81). In the same vein, Berente et al. (2021) call for research on the social aspects of explanations, noting that "interpretation is always situated within a social context" (p. 1444). Hence, ensuring a model's technical explainability may not be enough for knowledge integration, as people may still not be able to interpret it.

This calls for consideration of the social context of AI use, which includes stakeholders who use the system for various purposes, as well as those who are subject to the model's decisions and actions (Ribera & Lapedriza, 2019). In this vein, Arrieta et al. (2020) provide a mapping of five relevant stakeholder groups for whom explanations have relevance, including: technical experts who create the AI system (e.g., software developers and data scientists), domain experts who use it (e.g., doctors and insurance agents), people affected by its outputs (e.g., citizens or customers), managers who make decisions about organisational AI implementations, and regulators who enforce compliance with law. Given that our focus is primarily on knowledge integration during the AI model development process, we focus on the perspectives of data scientists (who develop the model) and domain experts (who contribute to this process and ultimately use the model in their work).

Interpretability requires blending knowledge from both the data science and application domains. This can be challenging because these groups often operate independently, with distinct skill sets and mental models (Someh et al., 2023). Integrating domain knowledge with data-driven knowledge has been found to require multiple iterations wherein data scientists and domain experts engage in a sensemaking process when building training data, developing an AI model and using the model in practice (van den Broek et al. 2021). Such processes are encumbered by situations where the AI model's insights do not make sense to domain experts in light of their professional knowledge and experience. Failing to establish a reliable ground truth for assessing the model's performance can lead to the organisation abandoning the AI system (Lebovitz et al., 2021). These issues call for efforts on the part of data scientists to ensure domain experts can make sense of AI's insights, and on the part of domain experts to help data scientists integrate relevant factors into the models. In some cases, domain experts have been found to devise creative ways to interpret an opaque AI model. For example, radiologists have engaged in informal AI interrogation practices by scrutinizing the input data (i.e., CT scans) in cases where the AI model's recommendations differed from the human experts' own assessments (Lebovitz et al., 2022). These interpretations enabled them to integrate AI's insights into their own understanding and final decisions, resulting in learning and more accurate treatment of patients.

2.3 Knowledge gaps in AI development

Kayande et al.'s (2009) three-gap framework (Figure 2) provides an insightful perspective on the challenges of knowledge integration when implementing AI systems and how explanations can address these challenges. It posits that evaluating intelligent decision-support systems (such as AI) in data-rich domains, where decisions are repetitive and outcomes are uncertain, one should consider three models. In the context of AI, these three models include the algorithmic model behind the AI system (i.e., how it represents reality), its human user's mental model (i.e., how the user understands reality), and the "true model" (i.e., how things actually are in reality). These models may not always align, suggesting that

organisations should invest in bridging three essential gaps in understanding and/or representation.

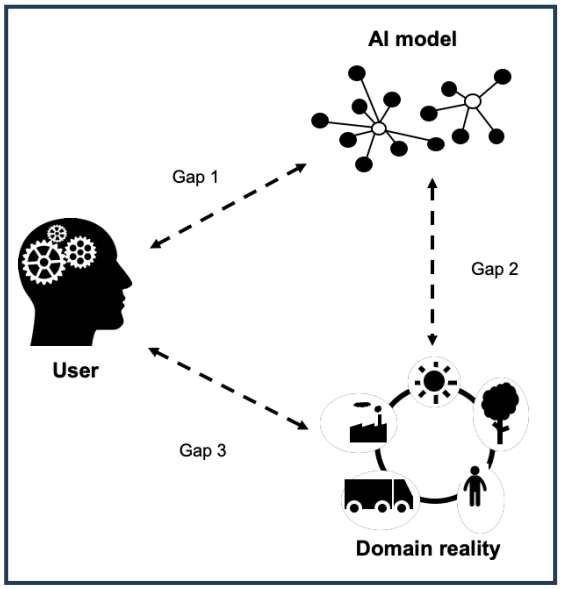

Figure 2: The Gaps Between the AI Model, User and Domain Reality

First, a situation of users not fully comprehending the AI model's logic indicates a gap between the user's mental model and the AI model (Gap 1). Such a gap is often rooted in AI inscrutability, whether stemming from the opacity of the AI model or the limited interpretational capability of the user. The gap makes it difficult for domain experts to detect mistakes made by AI systems and poses challenges for determining the accountability of decisions (Martin, 2019). It may further inhibit user trust and stifle acceptance of AI, as evidenced by cases of domain experts' aversion to opaque AI systems (Allen & Choudhury, 2022; Lebovitz et al., 2022). As such, the gap may prevent data-driven knowledge from being integrated into daily practice.

The second gap emerges in situations where the AI model does not represent domain reality in an accurate or meaningful manner[2] (Gap 2). AI models are only as good as the data used to train them, and they generally lack humans' ability to understand things within their context (Lebovitz et al., 2021). The resulting inconsistency between the AI model and domain reality results in poor performance and potentially in undesired effects such as bias against/for specific cohorts, which can then contribute to organisations' reluctance to leverage AI models (Lebovitz et al., 2021; Teodorescu et al., 2021). Inscrutability compounds this issue as it can hide models' biases and other unintended consequences. Depending on whether such models are implemented into use, this gap could result in either a failure to integrate useful data-driven knowledge or the integration of inaccurate and potentially harmful knowledge.

[2] *However, it is important to note that what is understood as "domain reality" can be a subject of debate and evolution over time. Indeed, previous AI studies have revealed that what is understood as "the ground truth" when training AI models may not in fact be true (Lebovitz et al. 2021). In this vein, Martens and Provost (2014) acknowledge that "'true" classifications of documents are subjective in certain domains, and it may be that a broadly used classification system changes the accepted subjective class definitions." (p. 79)*

Finally, a third gap concerns an inconsistency between the user's understanding and reality (Gap 3). A substantial body of research in psychology and behavioural economics suggests that human experts are subject to fallacies, often exhibiting subjective biases and limitations in their understanding of how things work in reality (Kahneman, 2011). Such biases hamper the integration of new knowledge and compromise the quality of domain experts' decisions, potentially curtailing their ability to achieve beneficial real-world impact. Moreover, as technology is increasingly automating knowledge work, domain experts may also experience loss of knowledge and expertise over time (Rinta-Kahila et al., 2023b), further amplifying the gap and the resulting difficulties in knowledge integration. Though AI models can help mitigate human biases, domain experts will struggle to learn from them if they operate inscrutably (Lebovitz et al., 2022).

In sum, while knowledge integration is crucial to addressing these gaps, AI inscrutability makes this challenging because it obscures how AI models process inputs, making it difficult to embed domain knowledge effectively into these models and hindering alignment between AI's outputs and stakeholder expectations. These challenges complicate the development of mutual understanding between data scientists and domain experts. While Kayande et al. (2009) showed how system explanations can reduce these gaps by enabling domain experts' learning and facilitating their system acceptance, Martens and Provost (2014) proposed that gaps can also be addressed by improving the system with the help of explanations that afford a better understanding of its workings. They further demonstrated the importance of considering the mental models of different organisational stakeholders. As we adopt this model as a framework for our study, we will next briefly reflect on different types of explanations that can help address knowledge gaps between the said entities.

2.4 AI explanations

In general terms, explanations refer to the reasoning of "why particular facts (events, properties, decisions, etc.) occurred" (Miller, 2019, p. 3), typically aimed at "communicating an understanding" (Keil, 2006) or clarifying a matter (Gregor & Benbasat, 1999). Hence, the act of explaining refers to the provision of explanations to an interested party. Explanations can involve one or more stakeholders. A person can explain an event to oneself, using explanation as a "verbal strategy" aimed at making sense of an event or solving a problem. However, more typically, explanations occur "between individuals and reflect an attempt to communicate an understanding" (Keil 2006, p. 2). Doing this can clarify an unclear matter or rectify a misunderstanding (Gregor & Benbasat, 1999). We refer to such explanations as *human-to-human explanations* for clarity. Such explanations play a key role during AI development projects, e.g., when data scientists explain an AI system's decision-making logic to domain experts (Asatiani et al., 2021; 2020), when domain experts explain the importance of various business factors to data scientists (van den Broek et al., 2021), and when managers explain the AI model's value to other stakeholders (Someh et al., 2022).

The proliferation of "intelligent systems", first in the form of rule-based expert systems of the 1980s, later in today's predictive machine-learning systems and large language models, has given rise to explanations that are provided by information systems (IS) to their human users

(Gregor & Benbasat 1999). We refer to such explanations as *machine-to-human explanations*.[3] According to Gregor and Benbasat (1999), machine-to-human explanations may manifest as: 1) a line of reasoning that explains "why certain decisions were or were not made by reference to the data and rules used in a particular case", 2) a justification that explains "part of a reasoning process by linking it to the deep knowledge from which it was derived", 3) a strategic explanation that sheds light into "the system's control behavior and problem solving strategies", or terminological explanation that provide definitional information (e.g., by defining a given term) (p. 503). If designed well, such explanations can significantly bridge the knowledge gaps: empirical research on rule-based systems has confirmed that domain experts are more likely to adhere to a decision-support system's recommendations if the system provides explanations and if the explanations have a good fit with the user (Arnold et al., 2006). However, if explanations are not available, organisations may end up over-relying on the system even if it is producing erroneous outputs (e.g., Rinta-Kahila et al., 2022). Further, research on both AI and rule-based systems suggests that mere availability of explanations does not necessarily prevent over-reliance: domain experts may overlook them if they are not perceived to offer enough immediate benefits (Rinta-Kahila et al., 2023b; Vasconcelos et al., 2023). While building explanation facilities into rule-based systems is somewhat straightforward, the provision of explanations has become a pertinent challenge with opaque AI systems. Hence, some organisations have been found to appoint "brokers" to translate algorithmic knowledge across data scientists and domain experts (Waardenburg et al., 2022).

There is a need for a deeper understanding of how explanations, whether human-to-human or machine-to-human, shape (and are shaped by) knowledge integration when developing inscrutable AI systems and implementing them for use. Building on this perspective, we take a sociotechnical approach to understanding knowledge integration during AI development. As such, we qualitatively explore the role of explanations in addressing knowledge gaps in AI development (Kayande et al., 2009; Martens & Provost, 2014).

3 Methods

Prior applications of the gap model have been conducted via hypothetical experiments (Kayande et al., 2009; Martens & Provost, 2014), leaving the model's potential to generate rich insights untapped. To gain a nuanced understanding of how knowledge integration occurs when developing AI models, we adopted the qualitative case study approach (Yin, 2018). Considering the scarcity of empirical knowledge on how to manage AI inscrutability in sociotechnical contexts (Berente et al., 2021), the case study method was deemed a suitable approach to teasing out rich insights into this poorly understood phenomenon. The method proved invaluable to understanding how explanations shape and are shaped by interactions among the AI model, users and domain reality within an organisation's sociotechnical context. Next, we present the three case studies chosen for this inquiry, after which we discuss our approach to the collection and analysis of data.

[3]*Recent advances in large-language models (LLMs) have caused a proliferation of could be termed as "human-to-machine explanations" (e.g., a human user explaining their informational needs to an LLM chatbot). Such explanations are beyond the scope of this study.*

3.1 Three case studies

We studied AI projects undertaken by three organisations: two public agencies in Australia and one private US-based conglomerate. The case organisations were identified as a part of a larger project on AI. The systems, which we will refer to as Tax AI, Health AI, and Contractor AI, leverage various ML-based technologies aimed at improving specific facets of the focal organisation's operations. A common characteristic of all three contexts is that decisions made by the organisation impact human stakeholders, whether private citizens or employees. Hence, all three AI systems have been designed to inform and augment human experts' decision-making, not replace it. These cases were selected as each one involved notable knowledge integration challenges wherein the organisations were faced with inscrutability as they strived to reconcile algorithmic knowledge with domain knowledge. While the degree of implementation success varied, the organisations demonstrated concrete efforts to deal with AI inscrutability with the use of explanations. Thus, they exhibited high theoretical relevance to our inquiry. Table 1 summarizes key information about each case.

Project	Organisation	Geographical area	Purpose	Stage of the project
Tax AI	State revenue agency	Australasia	Improve citizens' land tax compliance through proactive identification of taxpayers at risk of becoming tax debtors	Implemented for use
Health AI	State hospital's emergency ward	Australasia	Reduce unnecessary sepsis deaths through timely identification and treatment of patients at risk	Pilot testing completed
Contractor AI	Multinational conglomerate's Environment, Health, and Safety team	North America	Improve the management of contractor risk	Implemented for use

Table 1: AI project case studies

3.1.1 Tax AI

The Tax AI project aimed to identify taxpayers who exhibited a risk of becoming chronic debtors by failing to make their payments on time. The state tax revenue-management office turned to AI technology in its search for ways to enhance taxpayer services and improve tax collection rates. They had a vision of AI giving the office's call-centre workers access to richer and more accurate insights so that they could implement appropriate intervention strategies tailored to individual taxpayer circumstances. In 2018, the agency commissioned a software vendor to develop an AI model for land-tax debt due to that tax area's high rates of payment default. The model was trained with roughly 200 million data records of nearly 100,000 taxpayers, spanning seven years.

3.1.2 Health AI

The Health AI project aimed to identify patients at risk of developing sepsis while waiting for treatment in a hospital's emergency department (ED). Many patients arriving at the ED are susceptible to this life-threatening condition, which is caused by the human body responding to infection in a way that damages its own tissues and organs. While hospital personnel can

treat sepsis effectively at low cost with antivirals and antibiotics, detecting it in a patient early enough is far from straightforward. Detection involves a combination of various confounding factors and humans' cognitive limits in considering these factors lead to unnecessary deaths from the condition. Giving antibiotics to patients 'just in case' is also problematic due to growing issues caused by antibiotic resistance. Seeking to better manage this wicked problem, a state health department developed an AI system to detect signs of sepsis in ED waiting-room patients and alert triage nurses to the possible need for rapid treatment.

3.1.3 Contractor AI

The Contractor AI project was developed at a large conglomerate offering a variety of high-tech products and services in different areas, including health, aviation, and energy, to name a few. The scale of the conglomerate's operations necessitates the use of contractors for various purposes. The organisation's Environment, Health, and Safety (EHS) team delivers company-wide governance and oversight, including the management of contractor risk. Since 2016, the team has aimed to proactively identify and prevent the possibility of high-risk contractor operations by implementing "Life Saving Principles" (LSP) standards, which are designed to guide work practices in high-risk operations. Contractor onboarding had already been a labour-intensive process, and incorporating LSP evaluations made it even more so. To evaluate each contractor's alignment with LSP standards, hundreds of EHS professionals had to manually vet the written policies of potential contractors (approximately 80,000 new ones annually).

To improve efficiency, a team of data scientists was tasked to develop an ML-based contractor document assessment tool. Contractor AI analyses the document and indicates whether the LSP criteria are likely to be satisfied or not. A human expert then decides whether to accept Contractor AI's assessment, investigate the documents further, or formulate a different assessment. If the contractor disagrees with the resulting assessment, the document is reviewed by a second EHS human expert. The project team concluded that the AI-driven LSP review process was simpler and more consistent than any manual process. Contractor AI frees EHS professionals to focus their expertise on higher-value work and assesses contractors in a more standardized and streamlined manner.

3.2 The collection and analysis of data

For all three projects, we collected data by interviewing key stakeholders and obtaining additional documentation material (see Table 2). We followed each case up longitudinally to get a deep understanding of the systems as they progressed from initial model development toward implementation and use. Because the development of the Tax AI system had been commissioned to an external vendor, we started by interviewing the vendor side (i.e., the data scientist and business architect) and then proceeded to consult staff at the tax agency, where the system was implemented. Both Health AI and Contractor AI were developed internally, though the former leveraged talent from outside the organisation (i.e., the UX design lead was a university researcher). In the initial interviews, the scope of the inquiry was relatively wide, and we asked about various aspects of the AI projects, including explainability considerations. We started each interview by asking for basic information about the informant's background and role in the organisation, the intent behind the AI project, the relevant stakeholders involved, the surrounding organisational environment, etc (see the initial interview protocol

Project	**Informants**	**Time and length**	**Additional data**
Tax AI	Data scientist	Nov 2019 (30 min)	-Slide deck on the use case by the system vendor -The state tax agency's user guide for citizens on paying land tax
	Business architect	Nov 2019 (30 min)	
	Principal solution architect and technology architect (group interview)	April 2021 (40 min)	
	Tax collections team leader	April 2022 (47 min)	
Health AI	Chief data scientist	Aug 2020 (59 min)	-The state agency's official "sepsis pathway" process map -A video showcasing the functionality of the Health AI system
	Clinical director	Aug 2020 (61 min) March 2021 (51 min)	
	Data analytics director	Aug 2020 (58 min) Mar 2021 (40 min) July 2022 (56 min)	
	UX design lead	July 2022 (54 min)	
Contractor AI	Compliance manager	Aug 2019 (60 min)	-Screenshots of the system interface -Email follow-ups with informants to ensure accuracy of case description
	Data scientist	Aug 2019 (60 min)	
	VP product manager	Aug 2019 (60 min) Aug 2020 (30 min)	
Total	11 informants	15 interviews (736 min)	

Table 2: Data collection

in Appendix A). We then delved deeper into the AI project by inquiring about the type of model ("What kind of ML algorithm did you choose and why?"), challenges encountered during the project, and ways in which these were overcome or managed. We specifically probed about explanations with questions such as "Can you explain how the AI is making decisions?", which was typically followed by asking about explanatory interfaces, if the informants did not mention them otherwise. As we followed up on each project, we delved more deeply into the explainability dimension, asking further questions about the role of explanations in the projects. The protocols for the follow-up interviews were composed separately for the specific informational needs we had for each project and stakeholder to ensure data relevance. In general, though we used a semi-structured interview protocol, its role was to serve as a loose guide as we allowed the conversations to flow naturally and made follow-up questions whenever interesting insights emerged. Interviews were conducted remotely via video call software. Each interview was recorded and transcribed.

We started our analysis by listening to the interviews and checking the transcriptions for accuracy. We then applied open coding (Strauss & Corbin, 1998) on the corpus of data by assigning codes to excerpts that related to our initial research interest in managing inscrutability during AI development, as well as to other relevant or interesting observations. These codes were kept close to the data by maintaining the language of the informants. The collection and analysis of data was an iterative process in which the former stage informed the subsequent stage. In parallel to this empirical work, we had begun to do conceptual work around the gap model, inspired by the multi-stakeholder approach of Martens and Provost (2014). By the time the first rounds of interviews had been conducted, we had identified the gap model as a suitable conceptual framework for understanding how explanations can help

organisations manage AI inscrutability. We then engaged in a top-down style of analysis by applying the gap model as an analytical lens. Here, we considered all 1st-level indicators that related to AI explainability or explanations in general and organised them under the model's concepts, i.e., gaps and alignments between the human stakeholder, the AI model, and domain reality. In line with Martens and Provost (2014), we also considered the direction to which the gaps were being closed, e.g., whether the gap between tax compliance officers and the Tax AI model diminished due to increased technical understanding of the officers or due to a well-designed interface and explanation provision of the AI model. In most cases, we found evidence of effects in both directions. For instance, while the gap between the Health AI model and domain reality of sepsis detection diminished through efforts to increase the AI model's accuracy (changing the AI model), we found that it was also reduced by the data-driven discovery that rule-based sepsis thresholds were very limited predictors of sepsis, resulting in reconsideration of the criteria for detection as well as academic publications about effective detection methods (changing the domain reality). We leveraged both the Nvivo software and Excel sheets as tools to facilitate the coding process.

The analytical process was highly iterative as the authors would regularly gather to discuss the emerging empirical insights and debate over their meaning in relation to the gap model. Sometimes disagreements among the authors needed to be resolved through lengthy discussions that drilled deeply into the philosophical meanings of explanation and domain reality. These were resolved by elaborating different perspectives and finding a common ground. The additional documentation data was used to enrich our understanding of the cases and to triangulate insights from the interviews. For instance, videos and screenshots of Health AI and Contractor AI helped us understand how machine-to-human explanations could help address gaps between domain users and the AI model. Similarly, a live demonstration of Tax AI's user interface at the organisation's premises provided us with an important glimpse into how the system's users experience the AI.

Our analysis led us to identify data scientists and domain experts as the key stakeholders of interest. This decision was partially shaped by empirical convenience, as we were able to access these particular stakeholders across all three projects. Hence, we extended previous gap frameworks by developing the six-gap model to capture the gaps between these two groups, as well as the AI model and domain reality. The analysis culminated in the identification of six mechanisms that addressed the knowledge gaps during AI development and implementation. As we reflected on these findings in the broader context of IS literature, we ended up casting AI development into the context of knowledge integration during ISD, which helped us further refine the mechanisms in light of this rich body of knowledge.

4 Findings

Given our focus on challenges of knowledge integration during AI development, we began our empirical data analysis by identifying the key stakeholders and the associated knowledge gaps they encountered when participating in AI projects. Since data scientists and domain experts are the primary actors directly interacting with AI models, we focused on these user profiles. Drawing on the frameworks proposed by Kayande et al. (2009) and Martens and Provost (2014), we developed *a six-gap framework*, as an extension of their work to capture misalignments in understanding between data scientists, domain experts, (inscrutable) AI models, and domain reality. Figure 3 and Table 3 illustrate the components of this adapted framework.

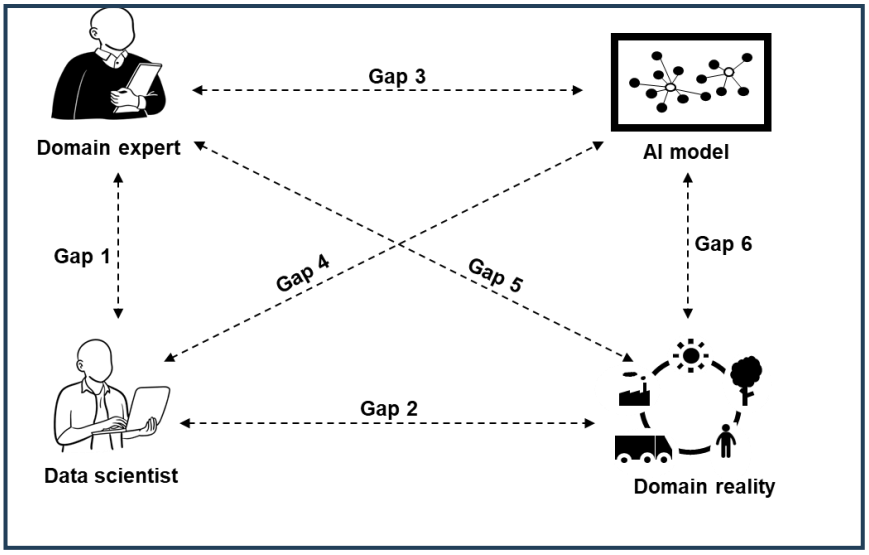

Figure 3: The six-gap model

Gap	Between	Description
Gap 1	Data Scientist ↔ Domain Expert	Misalignment between data scientists' and domain experts' mental models
Gap 2	Data Scientist ↔ Domain Reality	Data scientists lack the deep domain knowledge required for building relevant AI models
Gap 3	Domain Expert ↔ AI Model	Domain experts lack the technical skills required for understanding and interpreting AI models
Gap 4	Data Scientist ↔ AI Model	Data scientists may not fully understand the workings of opaque AI models
Gap 5	Domain Expert ↔ Domain Reality	Domain experts' understanding of reality is biased or incomplete
Gap 6	AI Model ↔ Domain Reality	The AI model does not accurately represent domain reality

Table 3: Knowledge Gaps in AI Development

Below, we provide a findings narrative structured around three mechanisms we identified for addressing the six key knowledge gaps. These mechanisms vary depending on the stage of AI development and involve two primary forms of explanation: (1) human-to-human explanations, which play a central role in the early stages of development, and (2) machine-to-human explanations (i.e., system-generated explanations), which provide some support for closing gaps during AI development but become more important during implementation and use.

4.1 Developing a shared understanding

The first step in integrating knowledge during AI development is establishing a shared understanding of domain reality. Our case studies revealed that effective AI development requires intensive cross-disciplinary collaboration between data scientists and domain experts whose mental models had to be aligned by developing a shared language and contextual understanding of the domain. This process was essential for reducing knowledge gaps between data scientists and domain experts (Gap 1), between data scientists and domain reality (Gap 2), and between domain experts and AI systems (Gap 3). In all three cases, human-to-human explanations played a central role in facilitating this alignment, enabling data

scientists to learn from domain experts about relevant decision factors, contextual nuances, and organisational practices. These interactions were often mediated by boundary-spanning roles, such as business architects or clinical directors, who facilitated alignment between stakeholders.

4.1.1 Developing a shared understanding of domain reality

To initiate model development, data scientists needed to gain contextual knowledge from domain experts on how the domain reality works (Gap 2). In our cases, building such a knowledge foundation relied on human-to-human explanations. Because the domain experts were the ones with insight into which factors are signals of a citizen failing to pay land tax, a patient developing sepsis, or a contractor failing to follow responsible practices, the data scientists worked alongside them to identify relevant features and gauge the models' outputs against human expertise. Domain experts equipped data scientists with the knowledge of different data points, decision variables, and how and why the variables are related to one another.

The tax collections team leader said:

> *"…when I initially sat down with them, I outlined what our business practices were with each collection step within land tax and kind of explained to them what happens with the first notice down to our legal stage and how to identify that in the pulse. So, it was just my role to provide the business knowledge so they could feed it into the actual program."*

The two parties' mental models came closer to each other as they found a common language. However, this process required facilitation. Oftentimes, the interaction between domain experts and data scientists was mediated by an intermediary who drew together the two fields' knowledge. In Tax AI, business architect acted in such role:

> *"What does it mean if a taxpayer has three different late payments? Is that a high risk? Is that not? For them to be able to train the model, we need to get all that business context and basically take that from those business users (and) put them into (the heads of) our data scientists. (business architect for Tax AI)"*

Similarly, in Health AI, the clinical director acted as a bridge between data science and domain knowledge, by translating clinical processes and contextual nuances into a language that data scientists could understand:

> *"a lot of times I'll let them connect together and I sit back […] I'm always there on call so if they need some quick query they might ask me […] so I think it's not just about me talking the language and being in the middle, but actually just building that bridge, so if I'm not there, they can still cross across (clinical director for Health AI)."*

In sum, domain experts' explanations to data scientists reduced Gap 2 by increasing the latter's understanding of the business domain. Moreover, these efforts decreased the gap between data scientists and domain experts (Gap 1), as the former ones became more familiar with how the latter ones think and work.

4.1.2 Developing a shared understanding of AI models

To ensure the models would be accepted and used effectively by the domain experts, data scientists needed to ensure that domain experts stay in the loop and understand how AI models learn and produce results (i.e., to reduce Gap 3). Here too, human-to-human

explanations played an important part, but this time through explanations by data scientists to domain experts.

In the Tax AI project, data scientists began educating tax experts on the principles that formed the foundation of the AI system's design, including its input data, functionality, limitations, and output decisions. The team had to explain the overall logic of the AI's operation so that customer-service staff could understand how the model takes various factors into consideration and recognize the probabilistic nature of its outputs. Illustrations of specific scenarios, with visualizations, enriched the team's description of how the AI reached decisions about taxpayers.

The business architect for Tax AI elaborated on how this was explained to the users:

> *"We told them specifically: You cannot automate everything end to end. For example, when the machine tells 'this is a 90% probability [someone] will become a debtor...', you can't have that without human input and automatically let the machine send letters to the taxpayers. Because the machine itself is not 100% bulletproof."*

Furthermore, data scientists demonstrating the AI's logic highlighted to tax officers the need to collect more detailed textual data from customer interactions so that the model could be even more sensitive to the reasons for taxpayers' delayed payment. Tax AI's business architect stated that the customer-service staff "had to understand what they were doing. How do we now capture data, how do we now use this tool?"

The data science lead of Health AI, in turn, talked about their gradual approach to educating clinicians about AI to make sure the complexity does not overwhelm them:

> *"obviously when you start bringing in those interaction effects and variables that aren't necessarily a part of the current clinical sepsis pathways -- the variables that they'd be familiar with -- we didn't want to overwhelm them with that complexity from the start; we want to slowly build it up to the point where there's enough trust in the model without introducing such a cognitive load that the clinician disengages."*

In sum, data scientists' explanations to domain experts reduced Gap 3 through educating the latter ones on how AI models function and utilize data, contributing to an improvement in their technical skills. Similarly to the previous section, these efforts decreased the gap between data scientists and domain experts (Gap 1), as the latter ones become more informed about algorithms, enabling the two parties to start 'speaking the same language'.

4.2 Discovering new domain knowledge

The development of AI systems not only requires integrating existing domain knowledge but also offers opportunities to discover new knowledge. This occurs when the development process surfaces patterns that challenge established assumptions, reveal blind spots, or prompt domain experts to revise their understanding of reality (Gap 5). In this way, AI systems do not merely replicate human expertise; they extend it. The discovery of new domain knowledge is particularly important in contexts where human decision-making is shaped by heuristics, institutionalized practices or incomplete information. In our case studies, human-to-human explanations, often supported by machine-to-human ones, played a central role in enabling domain experts to engage with AI outputs, reflect on their own assumptions, and recalibrate their understanding of the domain. However, there are limits to the extent that AI models can produce new knowledge that is relevant to domain reality (van den Broek et al.,

2021). This limit became visible as the case organisations grappled with the challenge of maximising AI's accuracy while maintaining adequate explainability (Gap 6).

4.2.1 Revealing blind spots

While domain experts possess valuable knowledge, as humans, they are also fallible and susceptible to various personal biases (Gap 5). If such biases and blind spots have become institutionalized, what is generally understood as domain reality may not in fact represent the true state of the matter. In our case studies, explanations were instrumental in revealing biases, errors and blind spots in the domain experts' decision-making. This effect unfolded through showing domain experts system explanations of why the AI made specific decisions. As domain experts engaged with AI explanations and became aware of their own biases, they often began to reframe their understanding of how the domain operates. This reframing involves moving beyond entrenched assumptions and adopting more nuanced, data-informed perspectives.

In Tax AI, by 'unwrapping' the AI decision-making process, the data scientists were able to deepen the domain experts' understanding of why and when taxpayers defaulted on their debts. For example, people who have been on interest-free payment plans in previous years have a tendency to become debtors later by failing to pay by the due date without any apparent reason. The AI system's customer-journey visualizations helped data scientists demonstrate to the tax officers that many of these people believed that they were still on extended payment plans and thus had thought they had been doing the correct thing. In another example, some individual customers aged under 35 were failing to pay even after the agency had sent them numerous whitemail letters. This data, coupled with the recorded absence of digital contact, led the agency to realise that these customers may in fact be willing to pay; they just had not kept their address information up to date.

With the Health AI project, hospital nurses had been relying on simple cut-off values to identify whether a patient was likely to have sepsis. The clinical director said: "people would look at patients and (declare that) a pulse of more than 120 is sepsis-positive (so) pulse 119 is not sepsis-positive." Data scientists and the clinical director explained how AI development revealed these thresholds as problematic because the reality is more complex than binomial cut-offs, and reliance on such a simplification renders a nurse likely to miss many sepsis cases. This helped the nurses - and their organisation more broadly - move beyond the old paradigms and gain a more comprehensive understanding of the decision features involved in the detection of sepsis.

At EHS, data scientists showed decision outputs made by Contractor AI, along with system explanations for the AI model's decision criteria. The domain experts found these explanations as revelatory as they revealed mistakes in the experts' decision-making, alarming them about the possibility of higher-than-expected error rates in the manual process.

> *"So we started off very simple with a few lockout tag out documents. So [data scientist] built a really early prototype and he tried it out on a couple and what we did is we did a blind test. And his simple out of the blocks model showed that I was making errors in my judgment. And so that was the human error piece – it was like, Oh. And then I had a hard time getting my head around that because one of his data scientists came back and I'd say, "Well no, that's not it." Like I don't see that. And then he came back with, "Well, it's described here this way." And I'm like, "Oh, yep, I made a mistake. Oh, absolutely. I made a mistake." And it was not only faster than I could*

> *do, but that's what really got me convinced of the human error. And I developed the criteria. I developed the bag of words. I really think that I've got a pretty good handle on evaluating these documents. And I'm probably one of the best to do that. Right? So, if I'm making errors..."*

In sum, as the data-science team took advantage of data and AI's power to expose and explain previously overlooked features to domain experts, the latter ones' understanding of their domain was refreshed and deepened (decreasing Gap 5).

4.2.2 Optimising for both model performance and explainability

As the case organisations explored how AI models could surface new domain insights, they also encountered the model performance-explainability trade-off. Too much opacity makes AI models' workings difficult to understand for data scientists (Gap 4), which can hamper their ability to manage and improve the models. Yet, if explainability is pursued at the expense of performance, the model may not provide an accurate view of domain reality (Gap 6). Pursuing both performance and explainability required selecting and combining modelling approaches that best captured domain reality while remaining intelligible to both data scientists and domain experts.

When evaluating the models' performance against real-world outcomes, and as project teams moved from simple machine-learning algorithms to more complex ones, such as deep neural networks, they noticed that increasing accuracy came with the cost of decreasing explainability. The Tax AI data scientist said:

> *"Even as a data scientist, when we run a neural network, we still don't really fully understand (what is) happening inside the neural network; (…) random forest is much easier, because we can visualize the decision tree and show how the decision is made."*

However, our findings show that one does not necessarily have to sacrifice performance for explainability. Because different datasets require different kinds of modelling approaches, the Tax AI team adopted an approach of combining a random-forest model and a deep-learning model. This combination of models not only helped boost performance but also enabled them to increase explainability, as the resulting ensemble model was far more interpretable than the largely opaque approach of a deep neural network. The Tax AI data scientist remarked:

> *"We tested (…) ensemble methods, and we've just (…combined) some of the algorithms together (…) we ended up using random forest and also the neural network(…) (We) mash the two results together: (for the) neural network, it is hard to explain, but for random forest, we can actually show them (that) these are the feature importance – why the model says it this way."*

In the case of Health AI, the team was highly conscious of the explainability-accuracy trade-off when they explored competing models. The team started with logistic regression because it was "(the) most transparent and (integrable) with clinical workflows", according to the data analytics director. As they moved to more complex models, such as boosting techniques and neural networks, they scrutinized both the accuracy and explainability of each model. If an opaque model seemed to deliver notable performance benefits over an explainable one, the team would try "to build some transparency back in" by seeing whether linear approximations could be run on top of the opaque model.

Hence, by prioritising more explainable models, the data scientists were able to maintain a sufficient understanding of their workings (reducing Gap 4) while using techniques to

improve model accuracy in representing the real world to diminish the gap between AI models and domain reality (Gap 6).

4.3 Shaping domain workflows

Once AI models have been developed and refined, the challenge shifts to embedding them into real-world organisational contexts. This stage is not merely about deploying technology: it also involves aligning AI-generated insights with domain workflows, stakeholder expectations, and decision-making processes. Applying AI models to domain practice requires careful attention to how domain experts interact with the system, interpret its outputs, and incorporate its recommendations into their work. Our case studies show that successful application hinges on the provision of meaningful explanations, particularly machine-to-human explanations, which help domain experts understand, accept, trust and act upon AI outputs. These explanations not only facilitate adoption (reducing Gap 3) but also enable domain experts to augment their decision-making (reducing Gap 5). Further, AI development also prompted changes in domain reality as the data-enabled insights led organisations to revise their practices and what they considered as "ground truth" to bridge the gaps between domain reality and domain experts (Gap 5) as well as domain reality and AI model (Gap 6).

4.3.1 Augmenting domain experts

While the previously described explanations by data scientists contributed to domain experts' understanding of the AI model's general operating principles (global explanations), integrating the model effectively into their work required providing justifications for specific AI decisions (local explanations). Otherwise, the domain experts might doubt the AI system's recommendations and fail to adopt them (giving rise to Gap 3). Ensuring acceptance and continued use required the development of automated system explanations. The project teams experimented with simple, approachable AI interfaces to present the model's insights and the recommended actions, alongside explanations of their basis.

In the Tax AI project, the data-science team collaborated with an IT application team to develop a user interface that provides system explanations by translating the AI's insights into a graphical depiction of a customer journey. The interface applies simple textual and visual cues pointing to whether and why a given taxpayer seems likely to become a debtor and, backed up by these explanations, suggestions for actions that may constitute appropriate intervention. The business architect for Tax AI noted:

> *"Like a traffic light, green to red, to show increasing risk to low risks. So really easy to understand, and we just used the percentage: were they 85% or…?"*

Along similar lines, the Health AI interface displayed patients' estimated risk of sepsis by means of distinct colours, alongside explanations. The UX design lead emphasized the user-centred nature of the interface design as triage nurses were involved in the process. The data-analytics director further elaborated on how explanations were then integrated into the design:

> *"So we had red and orange as sort of different levels in the mock-up, but we're probably going to try just red highlighting to keep it simple for the initial launch of this tool, and we're going to get their feedback on whether there's any value in having a nuanced approach (...) you can click the patient and get the full range of factors that are leading to that prediction."*

Thus, the AI models, by informing decisions and processes that affect citizens, suggest ways to make a real-world impact. Moreover, arming upskilled domain experts with AI interfaces

enables them to exert effects in an informed manner within their domains. As the data scientist for Tax AI noted:

> *"It basically augments your job. So, it helps you in your daily job, to help you (gain a better) understanding about your clients – about your taxpayers, for example. So, the next time, for when the taxpayers call, you'll be able to understand what they have done in the past, what action (…) had been taken so that you will be able to advise better."*

EHS experts, specialists in safety, first struggled to trust the AI decisions without clear evidence and safeguards. In response, the Digital Team developed a contractor assessment interface to make the decision-making process transparent and actionable for EHS reviewers. The UX displayed criterion results for each safety principle using color-coded indicators (blue = satisfied, red = not satisfied) and allowed reviewers to drill down into the original document for verification. To enhance interpretability, the interface showed probability scores for each outcome and provided dashboards summarizing review history, including high-risk versus low-risk criteria. The compliance manager said:

> *"So [the data science] team created a user interface for us so that we could visualize. And really what it is - is that when you run through the documents, through the tool, against the criteria, it will display through all of the criteria we either passed or not. Is it blue? It met the requirement. If it's red, it did not meet the requirement. And then we're able to go in there and see if the tool got it right."*

An additional explainability feature was the use of a text summarization technique that highlighted the minimum number of words or excerpts from the document that informed the model's decision (whether a requirement was satisfied or not). This feature was embedded in the UX to help reviewers quickly understand the rationale behind the model's assessment. Lastly, the data science team implemented conservative classification controls and showed how they minimised false positives to ensure safety and compliance as the tool scaled.

In sum, machine-to-human explanations increased the systems' interpretability to domain experts, resulting in their increased AI acceptance (reduced Gap 3) and a better understanding of domain reality (reduced Gap 5).

4.3.2 Revising the ground truth

As the domain experts' understanding of the domain evolved, so did what the organisations understood as "ground truth." Since in our cases, expert knowledge was used to train and evaluate AI models, as experts learned from AI, they updated the criteria used to define correct outcomes, thereby improving both model performance and domain practices (addressing Gaps 3 and 6).

In Tax AI, the data-driven insights about why some people were becoming debtors (see p. 17) led the agency to revise its understanding of factors that contribute to becoming a tax debtor. As such, they updated their ground truth about the matter and revised some of their operating practices. For instance, they changed the contact method for people under 35 from whitemail to email, with the positive consequence of an increased rate of timely payment.

The Health AI case differs from the other two in that, as a public healthcare institution, it is subject to strict regulatory oversight and faces accountability over patients' health outcomes. The nurses had to operate in line with the official sepsis pathway, and there were strict limits to the extent to which the process could be altered. Hence, the AI system was designed to

augment the existing process of nurses checking people's health information from a computer screen (e.g., age, blood pressure) coupled with making rounds at the ED waiting area to observe any changes in patients' conditions. However, the AI development process revealed that the nurses did not, in fact, make these rounds because they were too busy with all the paperwork and clinical work at the department. In other words, what was previously assumed as ground truth was not, in fact, true. Because the plan for AI integration had previously operated on the assumption that the sepsis process operates as expected, the development team had to revise their assumptions and think of how to account for this unexpected issue.

Finally, to improve that their AI model's representation of domain reality over time, the EHS team built a feedback interface into Contractor AI that allowed domain experts to evaluate and comment on algorithmic outputs. The data scientist at EHS explained how Contractor AI's performance improved in response to domain expert feedback:

> *"I think it adds to users but the very first and foremost thing – the reason we did set up is the data scientists want to consume that information, take that feedback, and then rebuild the algorithm to make it more robust."*

Data from reviewer feedback and override controls (thumbs up/down, comments, and ability to override satisfied/not satisfied) were captured for model retraining, creating a feedback loop of explanation and learning. As domain experts engaged with AI outputs and provided corrections or refinements, their feedback - a form of human-to-human explanation - contributed to model training. It enabled data scientists to revise the ground truth used to evaluate model performance, leading to more accurate and context-sensitive AI systems. In doing so, the EHS not only improved its model but also aligned it more closely with its domain reality.

Beyond improving individual decision-making, the implementation of AI systems also led to broader shifts in organisational practices. As illustrated in the Contractor AI case, the system prompted the EHS team to adopt a more systematic and consistent approach to evaluating contractor documents, an outcome that had not been achieved across the organisation prior to AI adoption:

> *"And then the three people we have doing the inspections, agreeing on what the process is and doing that every time. We read the document first from start to finish. And then we take criteria by criteria and we look for things that trigger the criteria and then does it meet the criteria. So we have a very standard process. I would say that [prior to Contractor AI] we didn't have one, nor did any of the other 3000 EHS professionals in evaluating contractor documents."*

Importantly, the integration of AI required not only changes in processes but also in the design of artefacts. Documents and workflows were restructured to be machine-readable, enabling smoother algorithmic processing and reducing ambiguity in interpretation. This shift from human-operated to machine-ready artefacts fostered greater consistency in how work was conducted, aligning organisational routines with the logic and capabilities of AI systems.

The case examples above demonstrate how implementing AI did not merely support existing practices—it reshaped them. This reduced the gaps between the AI model, domain experts and domain reality (Gap 5 and 6).
Table 4 provides an overview of the three sociotechnical mechanisms we identified.

Mechanism	Sub-mechanism	Empirical observations	Gaps Addressed
Developing a shared understanding	Developing a shared understanding of domain reality	• Human-to-human explanations from domain experts to data scientists • Use of boundary-spanning roles (e.g., business architect, clinical director) • Iterative dialogue to surface relevant features, decision variables, and contextual nuances • Breaking down functional silos	Gap 1: Data Scientist ↔ Domain Expert Gap 2: Data Scientist ↔ Domain Reality
	Developing a shared understanding of AI models	• Human-to-human explanations from data scientists to domain experts • Visualizations and scenario-based walkthroughs Gradual onboarding to avoid cognitive overload	Gap 1: Data Scientist ↔ Domain Expert Gap 3: AI Model ↔ Domain Expert
Discovering new domain knowledge	Revealing blind spots	• Machine-to-human explanations that contrast AI decisions with expert judgments • Surfacing overlooked patterns and biases • Reframing domain understanding through interaction with AI outputs	Gap 5: Domain Expert ↔ Domain Reality Gap 6: AI Model ↔ Domain Reality
	Balancing performance and explainability	• Combining interpretable and high-performing models • Iterative model refinement based on domain feedback • Revising ground truth based on expert learning	Gap 3: AI Model ↔ Domain Expert Gap 4: AI Model ↔ Data Scientist Gap 6: AI Model ↔ Domain Reality
Shaping domain workflows	Augmenting domain experts	• Machine-to-human explanations embedded in user interfaces • Local explanations for specific decisions • Visual cues and contextualized recommendations	Gap 3: AI Model ↔ Domain Expert Gap 1: Data Scientist ↔ Domain Expert
	Revising ground truth	• Feedback loops from domain experts • New data points generated from expert corrections • Continuous refinement of model evaluation criteria	Gap 3: Data Scientist ↔ Domain Reality Gap 6: AI Model ↔ Domain Reality

Table 4: Mechanisms for knowledge integration in AI development

5 Discussion and conclusions

Our study set out to shed light on the following question: How does knowledge integration occur in the development and implementation of inscrutable AI systems, and what is the role of explanations in this process? Our empirical findings shed light on this question by identifying specific knowledge gaps among human stakeholders, AI models, and domain reality, showing how they shape knowledge integration during AI development, and uncovering organisational strategies for addressing the gaps. We identified three sociotechnical mechanisms that facilitate knowledge integration during AI development: 1. developing a shared understanding; 2. discovering new domain knowledge; and 3. shaping domain workflows. These mechanisms are actualised by the provision of human-to-human and machine-to-human explanations. Next, we discuss the implications of our findings.

5.1 Theoretical implications

5.1.1 Knowledge integration during AI development

Knowledge integration has long been recognized as a critical process in ISD, where diverse technical and domain knowledge must be combined to create effective systems (Matook et al., 2021; Hahn & Lee, 2021; Tiwana, 2009; Faraj & Sproull, 2000). However, our study shows that this process plays out differently in the development of AI systems. Unlike traditional ISD projects that produce rule-based systems with transparent logic, AI development involves building predictive models characterised by inscrutability (Berente et al., 2021). The challenges of inscrutability, whether related to explainability or interpretability, complicate efforts to fully align AI models with users' mental models and domain reality. Our study identifies these misalignments as knowledge gaps and advances a set of mechanisms to address them, thereby extending ISD research by explaining how knowledge integration unfolds in the context of probabilistic, evolving, and opaque technologies.

First, *developing a shared understanding* highlights the need to align the mental models of data scientists and domain experts, an issue recognised in the ISD literature but amplified in AI contexts due to the opaque and probabilistic nature of models. While previous research has explored successful configuration of teams for analytics initiatives (Someh et al., 2023), the dynamics of such interaction in the presence of an inscrutable model has remained unexplored. Our study shows how a shared understanding can be achieved through sustained, two-way explanations between data scientists and domain experts, supported by roles that bridge disciplinary boundaries. Domain experts provided contextual knowledge and clarified decision variables, while data scientists explained model logic and limitations using accessible tools and gradual onboarding. This mechanism enabled human experts to retain their competence and update their skills in response to technological progress is crucial in an age where cutting-edge AI models are taking over an increasing share of knowledge work (Eloundou et al., 2023) and threatening humans' expertise and relevance (Rinta-Kahila et al. 2023b; Strich et al., 2021). Developing technical literacy, in particular, is an important challenge for any organisation today.

Second, *discovering new domain knowledge* connects data-driven learning processes into the AI development process. This phenomenon has been covered in recent literature on AI implementation (e.g., Shollo et al., 2022). We enrich this understanding by giving specific attention to inscrutability, which continues to be a challenge for many organisations striving to use AI responsibly (Asatiani et al., 2021; 2020). Our findings lend support to recent arguments that explainability may not need to be sacrificed for accuracy – interpretable models can be surprisingly effective despite their relative simplicity (Mahya and Fürnkranz, 2023), and they can help provide explanations when run in conjunction with black-box models (Someh et al., 2022).

The third mechanism, *shaping domain workflows*, reveals how AI and data-driven knowledge can be leveraged to achieve a positive change in the real world. It highlights the importance of communicating AI insights to domain experts via dashboards and reconsidering the structure of work processes and the roles of people. Informing the organization with dashboards (Zuboff, 1991) and conducting large-scale process re-engineering based on analytical insights (Hammer, 1990) are not new ideas. However, our findings show that to deal with highly complex and virtually inscrutable, our case organizations developed interfaces to provide domain experts with a sufficient level of explainability, before task augmentation could occur.

A key aspect of the emerging workflow was to enable the generation of new ground truth to continue with model learning and adaptation over time.

All told, our case analysis suggests that AI development is not a one-off process, but instead, AI models are "incomplete by design" (Garud et al., 2008); they must continuously adapt to evolving user capabilities and shifting domain realities. This translates into constant iterations of social learning and technical refinement to models. In fact, we view AI development as a never-ending process, driven by ongoing shifts in the real world. This means that knowledge gaps cannot be simply closed and eliminated - without conscious efforts to monitor and address them, they are likely to grow over time and jeopardise organisational knowledge assets and learning capacity. What organizations can activate through this process is data-enabled learning, a self-reinforcing cycle where AI systems constantly improve based on user interaction data (Hagiu & Wright, 2023)

5.1.2 Managing AI inscrutability by closing knowledge gaps

Our study echoes previous work suggesting that managing AI inscrutability goes far beyond scrutinizing the technical traceability of AI models (Ågerfalk et al., 2022; Bauer et al., 2021; Berente et al., 2021) and provides much-needed empirical evidence for how the challenge can be managed. While other studies have treated explainability as a static property of the AI model that needs to be managed via sociotechnical arrangements (e.g., Asatiani et al., 2021), we provide more nuance to the matter by showing how inscrutability of AI necessitates a sociotechnical learning process involving both human-to-human and machine-to-human explanations at various stages. By doing this, we answer the calls to "regard machine learning as a special case of organisational learning" (Ågerfalk et al., 2022, p. 17) and understand inscrutability not only as a technical or cognitive problem but as a social problem (Miller, 2019; Berente et al., 2021). Our insights about how inscrutability affects knowledge integration in both social and technical ways can help future research further unpack different sociotechnical strategies for managing this challenge.

We contribute to theory by extending the gap model by Kayande et al. (2009) and examining its less-researched aspects (as highlighted by Martens & Provost, 2014). Specifically, our six-gap adaptation extends the model's previous versions (Kayande et al., 2009; Martens & Provost, 2014) by specifying two key human stakeholders and establishing a vertical knowledge gap between them. Our adaptation also specifically relates explainability and interpretability to the model's knowledge gaps, providing a theoretical connection between the gap model and AI inscrutability. In our empirical application of the model, we explicitly address the vertical gap between different human users by presenting evidence of how the knowledge gap between domain experts and data scientists manifests and how explanations can reduce it. Moreover, we show how the gap between AI models and domain reality can be addressed by reshaping what is considered true in the domain (as opposed to just revising the AI model to correspond to the assumed domain reality), something that has remained outside the scope of previous gap model studies (see Martens & Provost, 2014). More broadly, our work contributes empirically to the gap model by providing qualitative insights into how the model's gaps emerge in real-life organisational settings, how they shape AI development processes, and how they can be addressed by the interaction of various stakeholders. As such, we show how the gap model can provide rich insights into the sociotechnical processes around AI development. Our adaptation of the model can serve as a helpful framing device for future empirical studies, whether quantitative or qualitative.

While AI models become more accurate and inscrutable via advances in ML, our findings suggest that inscrutability should be understood in both technical and social terms by considering the model's codifiability in relation to different social actors who interpret the model's outputs with the help of explanations they provide. This finding provides further evidence for considering explainability as a capability organisations need to master (Someh et al., 2022; Rinta-Kahila et al. 2023a). All three case studies highlight the need for organizations to embark on a learning journey when implementing AI, one that has uncertainties and that stands in stark contrast with traditional IT implementation projects. This observation resonates with recent distinctions between digital transformation and IT-enabled change (Wessel et al., 2021). The learning journey in the case of AI suggests a process wherein both humans and machines accumulate organisational knowledge capital jointly and iteratively. As organisations and their stakeholders continue to learn about the workings of ever-evolving AI models, they also gain new insights about their own work processes, employees, and customers. The implication of this is that at the dawn of an AI-driven world, learning has finally become truly sociotechnical in nature.

5.2 Managerial implications

In addition to contributions to the current body of academic knowledge, our research identifies concrete practices that have the potential to help organisations carry out impactful AI projects with success. First, we show that it is possible to build explainability into complex AI models by examining and incorporating alternative traceable models. While many advanced AI models are inscrutable black boxes, they can be examined and tightly correlated relative to traceable ones. Combining different models may add some level of transparency to the decision-making process and even enhance performance. Following this approach requires heavy scrutiny from the explainability perspective: consistent, reliable decision-making necessitates assessing, comparing, and calibrating the models' performance against one another.

Second, we advise managers to move beyond technical traceability to consider explanations that engage and involve stakeholders in AI-model development. Our study highlights the importance of considering the users of AI models, and their knowledge, values, and perspectives when building AI. Training good AI models requires input from stakeholders. These stakeholders range from domain experts and executives to citizens, for many of whom the concept and potential value of AI remain unclear. Presentations of a model's technical operations and trace will not be meaningful for these stakeholders. Shifting from technical traceability to explanations of the decisions, actions, and mechanics relevant to the stakeholders can encourage their deep engagement with the model, whereby they can inform the model's training and guide its evolution. This enables user upskilling and helps to overcome user resistance that typically plagues IT deployments. Requesting stakeholder feedback and incorporating it into the AI model should be done on a continuous basis as part of the overall governance mechanism – not just in the development and implementation phases. Users who are kept educated and informed with explanations that clarify the models' boundaries and limitations add substantial value when empowered to exercise decision-making authority and override AI decisions. Furthermore, models can learn from cases of users questioning or overriding their decisions.

Third, we highlight the importance of developing user-friendly explanatory interfaces. When a model's complexity is mirrored by a highly technical application interface, the system is not

a good tool for many non-technical stakeholders. Regardless of how advanced the code behind the interface is, domain experts need simple tools built with specific user requirements in mind: interfaces that deliver an end-to-end process or service experience, preferably with clear explanatory visualisations. The AI models' integration into existing workflows, products, and services proves just as vital. Clear, uncluttered interfaces with visual aids get the most from humans in AI-augmented decision-making processes.

Finally, managers should plan for an iterative process. AI technologies are still nascent and emerging. Therefore, their implications for human stakeholders, work processes, and organizational arrangements remain poorly understood. This demands awareness and flexibility: organisations must exercise prudence via an iterative process wherein the business goals behind the AI system are refined and refocused, and in which the AI models get scrutinized, both periodically and in response to stakeholder feedback. Explanations are crucial for the detection of issues that necessitate revisions to the organisation's AI systems.

5.3 Limitations and avenues for future research

Our work is not without limitations. First, our empirical inquiry involved only three AI implementation cases and a limited number of informants. While two of the AI systems we studied had been implemented for use, one (Health AI) was still in the piloting stage. Hence, it is possible that further insights could be gathered once the system enters everyday use. Still, considering our focus on how inscrutability shapes knowledge integration during the AI development process, we do not find this problematic. Nevertheless, we caution that the long-term implications of the AI models in each case organisation are not yet known. This means that some knowledge integration mechanisms we describe here may not turn out as successful in the long run as they seemed at the time of conducting the study. It could also be that mechanisms that have not been identified here will emerge as the implementations mature. Considering more cases and informants may result in additional findings regarding AI explanations. We invite future research to extend our work to overcome these limitations.

Second, following from the previous point, our identification of mechanisms may have been limited by our chosen theoretical framework. We acknowledge that alternative frameworks might reveal additional mechanisms, for instance, by considering various other organisational stakeholders (see Arrieta et al., 2020). Moreover, considering concepts from cognitive load theory, organizational learning, and behavioural decision-making in relation to the gap model could add depth to the analysis. They could shed light on questions such as what happens when explanations contradict each other, and how such contradictions should be managed in organisations with AI-driven decision-making. Similarly, they could help explore how explanations diffuse within groups, and what factors shape their acceptance or rejection in such settings. Incorporating alternative or complementary theoretical lenses represents a future research opportunity to provide a richer and more nuanced understanding of how AI inscrutability affects knowledge integration. studying

Third, our findings mainly apply to organisations that have control and a say in the development process of the AI system they are using. Today, many organisations are buying off-the-shelf AI packages, such as LLM chatbots, from AI developer behemoths like OpenAI and Google. In such projects, the user organisation may have little influence on the final system's explainability. This brings up questions about power and politics in AI development, inviting future researchers to study questions such as what the distribution of power is in the generation of explanations, and how AI influences this.

Finally, our observation of AI development as a sociotechnical learning process calls for theoretical development on the idea of sociotechnical learning. Surprisingly, the Information Systems field lacks such theorising despite the highly sociotechnical and informational nature of our discipline. Our findings can help pave the way for such theory development by showing how bridging the gaps in representation and understanding between different models of reality can ultimately contribute to a more informed world. In general,

References

Ågerfalk, P. J., Conboy, K., Crowston, K., Eriksson Lundström, J., Jarvenpaa, S. L., Ram, S., & Mikalef, P. (2022). Artificial Intelligence in Information Systems: State of the Art and Research Roadmap. *Communications of the Association for Information Systems*, *50*(1), 420–438. doi.org/10.17705/1CAIS.05017

Akbarighatar, P., Rinta-Kahila, T., & Someh, I. (2025). When Welfare Goes Digital: Lessons from the Dutch Syri Risk Indicator in Public Sector. *Academy of Management Proceedings 2025 (1), 10358, Copenhagen, Denmark.*

Allen, R. T., & Choudhury, P. (2022). Algorithm-Augmented Work and Domain Experience: The Countervailing Forces of Ability and Aversion. *Organization Science*, *33*(1), 149–169. doi.org/10.1287/ORSC.2021.1554

Arnold, V., Clark, N., Collier, P. A., Leech, S. A., & Sutton, S. G. (2006). The Differential Use and Effect of Knowledge-Based System Explanations in Novice and Expert Judgment Decisions. *MIS Quarterly*, *30*(1), 79–97.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.

Asatiani, A., Malo, P., Nagbol, P., Penttinen, E., Rinta-Kahila, T., Salovaara, A. (2020) "Challenges of Explaining the Behavior of Black-box AI Systems," *MIS Quarterly Executive* 19(4), pp. 259-274.

Asatiani, A., Malo, P., Nagbol, P., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). "Sociotechnical Envelopment of Artificial Intelligence: Resolving the Challenges of Explainability in an Organization. *Journal of the Association for Information Systems,* 22(2), 325-352, doi.org/10.17705/1jais.00664

Australian Department of Industry Science and Resources. (2024). Australia's AI Ethics Principles. In *Australia's Artificial Intelligence Ethics Framework*. https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles#principle-6

Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Expl(AI)n It to Me – Explainable AI and Information Systems Research. *Business and Information Systems Engineering*, *63*(2), 79–82. doi.org/10.1007/s12599-021-00683-2

Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. *ACM International Conference Proceeding Series*, 248–266. doi.org/10.1145/3531146.3533090

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, *45*(3), 1433–1450. doi.org/10.4324/9781315691398-22

Business Wire. (2023). *Survey: AI Adoption Among Federal Agencies Is Up But Trust Continues to Be An Obstacle to Future Adoption and Use*. https://www.businesswire.com/news/home/20231214493999/en/

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. *ArXiv, Working pa*. http://arxiv.org/abs/2303.10130

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*, *Fall*, 50–57. doi.org/10.1609/aimag.v38i3.2741

Gregor, S., & Benbasat, I. (1999). Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly*, *23*(4), 497–530. doi.org/10.2307/249487

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *Journal of Strategic Information Systems*, *29*(2), 101614. doi.org/10.1016/j.jsis.2020.101614

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A Survey Of Methods For Explaining Black Box Models. *ACM Computing Surveys*, *51*(5). arxiv.org/abs/1802.01933

Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA 's explainable AI (XAI) program: A retrospective . *Applied AI Letters*, *2*(4), 1–12. doi.org/10.1002/ail2.61

Hagiu, A., & Wright, J. (2023). Data-enabled learning, network effects, and competitive advantage. *The RAND Journal of Economics*, *54*(4), 638-667.

Hahn, J., & Lee, G. (2021). The complex effects of cross-domain knowledge on IS development: A simulation-based theory development. *MIS Quarterly*, 45(4), 2023–2054. doi.org/10.25300/MISQ/2022/16292

Hammer, M. (1990). Reengineering work: don't automate, obliterate. *Harvard Business Review*, *68*(4), 104–112.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kayande, U., De Bruyn, A., Lilien, G. L., Rangaswamy, A., & van Bruggen, G. H. (2009). How incorporating feedback mechanisms in a DSS affects DSS evaluations. *Information Systems Research*, *20*(4), 527–546. doi.org/10.1287/isre.1080.0198

Keil, F. C. (2006). Explanation and understanding. *Annu. Rev. Psychol.*, *57*(1), 227-254.

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly: Management Information Systems*, *45*(3), 1501–1525. doi.org/10.25300/MISQ/2021/16564

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. *Organization Science*, *33*(1), 126–148. doi.org/10.1287/ORSC.2021.1549

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *ACM Queue*, *16*(3), 30. doi.org/10.1145/3233231

Mahya, P., & Fürnkranz, J. (2023). An Empirical Comparison of Interpretable Models to Post-Hoc Explanations. *AI (Switzerland)*, *4*(2), 426–436. doi.org/10.3390/ai4020023

Marjanovic, O., Cecez-Kecmanovic, D., & Vidgen, R. (2022). Theorising Algorithmic Justice. *European Journal of Information Systems*, *31*(3), 269–287. doi.org/10.1080/0960085X.2021.1934130

Martens, D., & Provost, F. (2014). Explaining Data-Driven Document Classifications. *MIS Quarterly*, *38*(1), 73–99. doi.org/10.25300/misq/2014/38.1.04

Martin, K. (2019). Designing ethical algorithms. *MIS Quarterly Executive*, *18*(2), 129–142. doi.org/10.17705/2msqe.00012

Matook, S., Lee, G., & Fitzgerald, B. (2021). *Information Systems Development*. In A. Burton-Jones & P. Seetharaman (Eds.), MIS Quarterly Research Curations. Retrieved from http://misq.org/research-curations

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. doi.org/10.1016/j.artint.2018.07.007

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. In *Artificial Intelligence Review* (Vol. 55, Issue 5). Springer Netherlands. doi.org/10.1007/s10462-021-10088-y

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Prediction of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 1135–1144. doi.org/10.1145/2939672.2939778

Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? A proposal of user-centered explainable AI. *CEUR Workshop Proceedings*, *2327*.

Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., & Gregor, S. (2022). Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. *European Journal of Information Systems*, vol. 31, no. 3, pp. 313–338, doi.org/10.1080/0960085x.2021.1960905

Rinta-Kahila, T., Someh, I., Indulska, M., & Ryan, I. (2023a). Building Artificial Intelligence capability in the public sector. *Australasian Conference on Information Systems 2023, Wellington, New Zealand.*

Rinta-Kahila, T., Penttinen, E., Salovaara, A., Soliman, W., & Ruissalo, J. (2023b). The Vicious Circles of Skill Erosion: A Case Study of Cognitive Automation. *Journal of the Association for Information* Systems 24 (5), 1378-1412. Article 2. Doi.org/10.17705/1jais.00829

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*. doi.org/10.1007/s10458-019-09408-y

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. doi.org/10.1038/s42256-019-0048-x

Russell, S., & Norvig, P. (2010). Artificial Intelligence. A Modern Approach. In *Pearson*

Education. doi.org/10.1119/1.15422

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

Someh, I., Wixom, B. H., Beath, C. M., & Zutavern, A. (2022). Building an Artificial Intelligence Explanation Capability. *MIS Quarterly Executive*, *21*(2).

Someh, I., Wixom, B., Davern, M., & Shanks, G. (2023). Configuring relationships between analytics and business domain groups for knowledge integration. *Journal of the Association for Information Systems*, *24*(2), 592-618.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research techniques*. Sage Publications.

Strich, F., Mayer, A. S., & Fiedler, M. (2021). What Do I Do in a World of Artificial Intelligence? Investigating the Impact of Substitutive Decision-Making AI Systems on Employees' Professional Role Identity. *Journal of the Association for Information Systems*, *22*(2), 304–324. doi.org/10.17705/1jais.00663

Teodorescu, M. H. M., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of fairness in automation require a deeper understanding of human–ML augmentation. *MIS Quarterly: Management Information Systems*, *45*(3), 1483–1499. doi.org/10.25300/MISQ/2021/16535

Van den Broek, E., Sergeeva, A., & Huysman, M. (2021). When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly*, *45*(3), 1557-1580. doi.org/10.25300/MISQ/2021/16559

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1-38.

Waardenburg, L., Huysman, M., & Sergeeva, A. V. (2022). In the land of the blind, the one-eyed man is king: Knowledge brokerage in the age of learning algorithms. *Organization Science*, *33*(1), 59-82. doi.org/10.1287/orsc.2021.1544

Wessel, L., Baiyere, A., Ologeanu-Taddei, R., Cha, J., & Jensen, T. B. (2021). Unpacking the difference between digital transformation and IT-enabled organizational transformation. *Journal of the Association for Information Systems*, *22*(1), 102–129. doi.org/10.17705/1jais.00655

Yin, R. K. (2018). *Case Study Research and Applications: Design and Methods* (6th ed.). SAGE Publications Inc.

Zacharias, J., von Zahn, M., Chen, J., & Hinz, O. (2022). Designing a feature selection method based on explainable artificial intelligence. *Electronic Markets*, *32*(4), 2159–2184. doi.org/10.1007/s12525-022-00608-1

Zuboff, S. (1991). Informate the enterprise: An agenda for the twenty-first century. *National Forum, Honor Society of Phi Kappa Phi*, *71*(3).

Acknowledgements

We thank the editor and anonymous reviewers for their helpful comments. Moreover, we are indebted to the three anonymous case organisations and all the interviewees for sharing their experiences to us. This research was supported (partially or fully) by the Australian

Government through the Australian Research Council's Industrial Transformation Training Centre for Information Resilience (CIRES) project number IC200100022. Furthermore, Tapani Rinta-Kahila is grateful for the generous support his research receives from the Australian Research Council (DE240100269) and the Research Council of Finland (370017).

Copyright

Copyright © 2025 Rinta-Kahila, T., Someh, I., Bidar, R., Ali Darvishi, A., and Indulska, M. This is an open-access article licensed under a Creative Commons Attribution-Non-Commercial 4.0 Australia License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and AJIS are credited.

doi: https://doi.org/10.3127/ajis.v29.5567

Appendix A: The interview protocol

-Warm-up questions about the informant's role and background (e.g., position in the organisation, professional and educational background, etc).

-Describe a specific AI initiative at your organisation. In your example…

1. What does the AI model do? What is your role in relation to the AI model?
2. What kind of ML algorithm did you choose and why?
3. Can you explain how the AI is making decisions? Can you justify the decisions? Is the decision-making process fully transparent for you?
4. How do you interpret AI-based decisions, recommendations or actions?
5. What inputs do you use for developing AI applications over time?
6. What new skills did you need to work with the AI model? What has changed?
7. What part of the organisation is driving this effort? What parts are involved? How do you coordinate across different groups?
8. How do you establish accountability for AI-based decisions?
9. How do you trust the AI model? How do you make sure AI is doing what was intended?
10. How do you identify and manage business rules? What are the roles of AI vs domain experts in relation to business processes?
11. What value has AI created for your company? How do you assess AI project value and risks?
12. What are some lessons learned from this project?