.Clarity in Complexity: Advancing AI Explainability through Sensemaking

Elisa Gagnon
Bishop's University
egagnon@ubishops.ca

Anouk DeRegt
Utrecht University
a.deregt@uu.nl

Liette Lapointe
McGill University
Liette.Lapointe@mcgill.ca

Abstract

This paper explores Explainable Artificial Intelligence (XAI) through a sensemaking lens, addressing the complexity in the extant literature and providing a comprehensive understanding of the process of explainability. Through an exhaustive review of relevant research, we develop a novel framework highlighting the dynamic interactions between AI systems and users in the co-construction of explanations. We conducted a thorough analysis and theoretical synthesis of the extant literature. Based on the results, we developed a framework that shows how explainability emerges as a shared process between humans and machines, rather than a one-sided output. The proposed framework offers valuable insights for enhancing human-AI interactions and contributes to the theoretical foundation of XAI. The findings pave the way for future research avenues, with implications for both academic investigation and practical applications in designing more transparent and effective AI systems.

Keywords: Explainable AI (XAI), explainability, sensemaking, human-in-the-loop, conceptualization.

1. Introduction

Explainable Artificial Intelligence (XAI) has emerged as a growing field of study within Artificial Intelligence (AI) and information systems (IS) fields, each contributing to a multifaceted understanding of its conceptual development and addressing the need for trust and understandability in various AI systems (Bitzer et al., 2023; Buomsoo et al., 2023; Stoffels et al., 2022). While XAI has been studied from various perspectives, to allow IS researchers to contribute to this emergent but critical domain of research, there is a pressing need to develop a better understanding of the XAI phenomenon (Gerlings et al., 2021).

Several XAI definitions and conceptualizations have been put forward in recent publications (Förster et al., 2020; Saeed & Omlin, 2023). For example, from a technical perspective, XAI is said to involve the deployment of methods through algorithms and models that can provide human-interpretable explanations of its predictions (Buomsoo et al., 2023; Lee & Ram, 2023; Stoffels et al., 2022), while from a cognitive perspective, it implicates how XAI methods influence human understanding of the explanations provided (Bauer et al., 2023). Most authors agree that providing a clear conceptualization for XAI is fundamentally complex, involving not just the technical mechanisms that render AI decisions explainable but also the cognitive processes that humans employ to understand these explanations and make decisions under uncertainty (Holzinger et al., 2017). The utilization of these explanations is contingent upon their reception by the intended users, which varies widely in terms of expertise, expectations, culture, and needs (Bauer et al., 2023). Indeed, the users' feedback and interactions with AI systems can also inform the iterative refinement of XAI approaches, ensuring they are aligned with user needs and context. The interaction between users and AI is thus a critical component of XAI, as it is through these interactions that the value and efficacy of AI explanations are ultimately realized (Bauer et al., 2023).

Despite the growing abundance of research on XAI, there remains a lack of consensus on how XAI should be conceptualized and operationalized. This is especially true when considering the contexts of how explainability is achieved from an AI use perspective and its interactions with different users (Lee & Ram, 2023; Stoffels et al., 2022). Additionally, from an IS perspective, there is a need to focus on the human and socio-technical aspects of XAI (Gerlings et al., 2021). This paper proposes to address this gap by adopting a sensemaking lens to examine the process underlying the actual "explainability" of AI for its users. By understanding XAI through the lens of sensemaking, we propose a nuanced framework that takes into consideration the diverse and dynamic ways in which humans interact with and make sense of AI systems, but also how AI technologies are trained to provide adequate explanations in their interactions with users. This approach not only shows how AI can achieve more transparency (opens the black-box of AI) but also fosters a more user-centric evolution of XAI for the IS field. By using a sensemaking lens, this paper aims to bridge the gap between black-box AI and human cognitive processes, explaining how AI explanations can be meaningful and actionable for various types of

URI: https://hdl.handle.net/10125/109008
978-0-9981331-8-8

users. Hence, this paper aims to answer the following research question: *How does a sensemaking perspective inform XAI, providing a more nuanced understanding of human-AI interaction, taking into account the contextual application of AI explainability?*

Studying XAI from a sensemaking perspective is expected to make significant contributions to the IS field and the broader body of knowledge on XAI. By elucidating the process through which users interact with and make sense of AI explanations, this research can offer novel insights into designing and implementing user-centric AI systems.

The remainder of the paper is organized as follows. In the next section, we present a comprehensive overview of the existing conceptualization of XAI and its dimensions. Then, an explanation of sensemaking theory is provided to motivate our approach as a foundation for our proposed framework. Lastly, in our conclusion, we propose avenues for future research and contributions to the field.

2. Literature Review

As IS research explores and influences the dynamics of how individuals, groups, organizations, and markets engage with systems, the interaction between humans and AI emerges as a pivotal area of inquiry within the discipline (Brasse et al., 2023). The nature of human-AI interaction is influenced by various factors including the characteristics of the system and the surrounding context (Rzepka & Berger, 2018). When it comes to modern AI systems, they are characterized by continuously advancing frontiers of computational capabilities, heightened autonomy levels, and enhanced learning capacities, compared to conventional systems (Berente et al., 2021). In terms of context, there appears to be a growing trend toward outsourcing decisions with economic, social, and political consequences to artificial intelligence (Dabas et al., 2023), which is driving the need for more responsible AI and insights into the underlying decision-making models. These combined factors spurred broad agreement amongst academics and practitioners that Artificial Intelligence systems should be able to "explain" their behavior. This subsequently resulted in the rise of *eXplainable* AI as a field. Various researchers, all adopting a different approach and focus, have contributed over the years to the development and establishment of the field. As a result, a large number of different definitions, conceptualizations, models, and frameworks have been proposed which, ironically, contributes to the perceived complexity rather than enhanced clarity on what exactly constitutes XAI (Dhanorkar et al., 2021). This also becomes evident from the literature where an increasing number of publications refrain from defining XAI (Lee & Ram, 2023; Züllig et al., 2023), instead referring to the XAI field as a whole (Förster et al., 2020) or referencing the overarching goal of combating "the black-box problem" (Stoffels et al., 2022). The black box issue refers to the situation in which a system is unable to offer any reason or suitable explanation behind its decisions (Hassija et al., 2024). Accordingly, this black box problem is underpinned by three layers of obfuscation: Firstly, its opacity arises from the adoption of inherently opaque computational techniques (deep neural network machine learning algorithms), that are chosen for their unique capabilities and greater effectiveness over more explainable alternatives. Secondly, it remains unclear how corporations and government agencies overseeing AI use legal provisions to conceal the inner workings of their algorithms, protect their competitive advantage, and maintain their reputation. Thirdly, the complexity of the technology itself adds to its concealment, making it unintelligible to most people.

Specifically, this last point highlights that the notion of explainability does not exist without the interactions between the end-user and the AI technology. Some of the research that outlines XAI as a stand-alone construct, often solely, regarded as a technological problem that resides in the algorithm and its specifications (Waardenburg & Huysman, 2022) falls short. Additionally, looking at the extant body of IS literature, the lack of consensus might also stem from what exactly constitutes an "explanation". The reason that the concept of explanation within the context of AI remains inherently ambiguous has in part to do with the fact that "explanation" is not a technical term (Gilpin et al., 2018; Lipton, 2018). This has caused a disconnect between the explanations that systems produce in service of XAI and those explanations that different stakeholders and end-users may need (Kim et al., 2024). The disarray of the notion of "explanations" is further evidenced by the current scope of the literature where a myriad of words like "transparency, interpretability, understandability, comprehensibility, opacity" are (interchangeably) used to refer to explanations while creating circular arguments. As such, in the pursuit of creating explainable artificial intelligence (AI) models, a foundational step lies in defining what constitutes an explanation. This process not only establishes the groundwork for developing explainable models but also delineates the fundamental pillars upon which explanations are constructed: the definition, the purposes and objectives of an explanation, the content it entails, and the various types of explanations that can be generated by the system. Additionally, evaluating explanation quality emerges as a critical endeavor, essential for the advancement of explainable AI.

Following previous guidelines (Paré et al., 2015; Rivard, 2024; Webster & Watson, 2002), our work involved an exhaustive examination of the literature about the definition, conceptualization, and framing of XAI in the current literature. Initially, our focus was directed towards articles published in leading IS peer-reviewed journals and conference proceedings. More specifically, we first focused on the AIS basket of 11 journals in the field as well as the proceedings from the top conferences in the discipline. To gain a more comprehensive understanding of how XAI has been conceptualized, we also conducted a broader search in Google Scholar, using keywords such as "XAI", "explainable AI", "XAI definition" "XAI conceptualization", "XAI framework", "XAI model", "XAI research", "XAI review" "XAI taxonomy". Our overall literature search yielded a substantial set of papers, identifying close to 6000 articles.

To ensure the relevance and rigor of our paper, we established a set of exclusion criteria through consensus among the authors. Initially, we removed any articles not written in English. Following this, we excluded non-academic content such as practitioner articles, editorials, opinion pieces, and brief notes, as well as papers that only marginally addressed the topic of explainable AI (XAI). Further, we eliminated papers that were not published in peer-reviewed journals, ongoing empirical research lacking data, and publications focusing exclusively on the technical development of XAI without broader analysis. Subsequent exclusions were conducted by reviewing the titles and abstracts, followed by a detailed examination of the full texts. Our final selection comprised 114 papers, including 22 that were review articles. All papers were read and coded. In our analysis, we wanted to avoid adding yet a simple review paper on XAI. Accordingly, we built on the core ideas of Rivard (2024), keeping in mind that good literature reviews transcend mere synthesis of existing knowledge. While our initial objective was to propose a clear conceptualization of XAI for the IS field, our analysis revealed that to embrace the complexity of the XAI concept, the development of a clear, concise, and parsimonious yet comprehensive framework would better serve our purpose. Our goal was to progress beyond a descriptive summary of the literature to contribute meaningfully to theoretical advancements in IS research on XAI.

The analysis of the extant literature underscores the current oversight regarding a crucial aspect of the current knowledge on XAI: the intended users of explanations, which should be defined by the full spectrum of functional roles, users, and capabilities for explanation. As such, some would even go as far as to say that explanations cannot be uniform, as each stakeholder seeks them with distinct aims, expectations, backgrounds, and requirements (Ribera & Lapedriza, 2019). Accordingly, this research identifies two streams of XAI literature and associated definitions: (1) technologically-oriented and (2) human-oriented. Since information technology, as a medium, has traditionally been approached from a technological viewpoint, definitions originating from this perspective all describe XAI in terms of the accumulation, and technological specifications, of the hardware and software.

A clear example of this approach is the popular DARPA definition which simply refers to XAI as *machines* that understand the context and environment in which they operate, and over time build underlying explanatory models that allow them to characterize real-world phenomena. Furthermore, these types of definitions often highlight the assessment of XAI in terms of performance criteria, as exemplified by (Zhang et al., 2022) who state that "XAI is based on the idea that if the *system* can explain its reasoning, its operational validity can be assessed for a wide range of criteria". Technical XAI concepts have proven especially valuable for developers and coders of AI software, resulting also in a more prominent appearance of these definitions in practical business literature or fields such as computer science. However, focusing the definition of XAI around the technical system inherently means that the application of such a definition is limited.

This lack of insight into the processes and the effects of using these systems served as the main catalyst for researchers to adopt a human-in-the-loop approach. The rationale for a human-centered experiential definition of XAI is influenced by different aspects (Miller, 2019). One element to consider is the fact that XAI technology has not stabilized yet: the advancement of the technology is still ongoing and new hardware and software technologies are continuously being developed by a multitude of actors. Another advantage of shifting the locus of XAI away from the technical system and towards the human-centered perception is that the unit of analysis can now be specified – the user. This is evidenced by definitions that put the human, rather than the machine or system at the heart of the definition. For example, Adadi & Berrada (2018) emphasize that for AI to be explainable, it has to be *human-understandable*; and that it is "given to the *audience*" (Arrieta et al., 2020) that XAI produces details or reasons to make its functioning clear or easy to understand. Considering the user experience of XAI means that dependent measures and application of knowledge about perceptual processes and user differences can be used to determine the nature of XAI (Capel & Brereton, 2023). Thus, from an IS perspective, the interaction between the human and the machine appears critical for the field. While a handful of scholars have started to advocate for the adoption of a human-

centered or social-technical view of theoretical development, this area remains nascent (Atf & Lewis, 2023; Baber et al., 2024; Naiseh et al., 2024).

3. XAI Theory Development in IS Research

In our theory development, we argue that XAI exists solely through an understanding of the machine-human interaction, more precisely between the AI and the users. Explainability is not inherent to AI models and technologies; it emerges through iterative and interactive loops between AI and users (Miller, 2019). These loops facilitate a sensemaking process, involving mutual contribution of information and interpretation from both AI and users (Hoffman et al., 2018). This collaborative effort leads to the co-construction of meaning, enabling users to understand AI explanations and AI technologies to refine these explanations for clarity and effectiveness. From this perspective, we define explainability as *the process in which the interactions between the AI and the users enable the interpretation of the AI explanations by revealing the underlying sensemaking mechanisms.*

As illustrated in Figure 1, we primarily focus our theory development on this sensemaking process, including the AI and its explanations, the human-machine interactive loops, and the variety of users involved. While we do not delve into the elements that externally contribute to this process, we acknowledge three important elements that influence this sensemaking process. First, the interaction between the AI and the users does not occur in a void and the given context in which it occurs will color the process (Abedin, 2021). Here, context refers for example to the domain and business needs and requirements relevant to the practical utility of the system (Buomsoo et al., 2023). Understanding the context is important in terms of when and for what purpose to integrate AI and what data to include (Subramanian et al., 2024). Second, the rationale behind the need for explainability forms a fundamental aspect of AI development, serving as a cornerstone for effective control, evaluation, management, and continual improvement of AI systems (Adadi & Berrada, 2018; Chromik & Schuesschrler, 2020; Haque et al., 2023; Meske et al., 2022). Through explainability, users, including end-users and managers, can assess AI performance, manage its deployment effectively, and facilitate informed decision-making. Additionally, a deep understanding of the AI model helps developers identify and mitigate potential biases, enhancing the performance and fairness of AI systems. Hence, the reasons for which explainability is needed influence the overall process (Miller, 2019).

Lastly, the sensemaking process inherent in user-AI interactions plays a pivotal role in determining trust, understandability, decision quality, fairness, privacy, satisfaction, and overall usefulness of AI systems (Arrieta et al., 2020; Gerlings et al., 2021). This process establishes a crucial feedback loop, refining the mechanisms of explainability and generating insights into user experiences, thereby improving the quality and relevance of AI explanations.

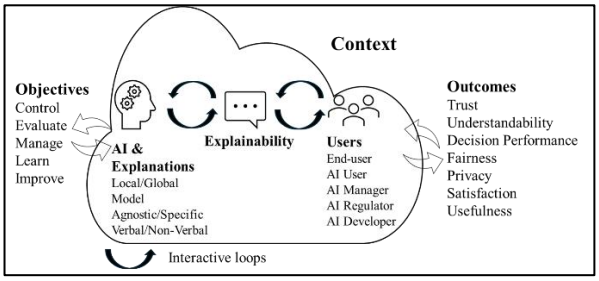

Figure 1. XAI Framework for IS Research

3.1. AI & Explanations

Various AI models, explanation methods, and types of explanations have been proposed based on the scope of explanation, the type of model it can explain, or combinations of these methods (Alarcon & Willis, 2023; Arrieta et al., 2020; Buomsoo et al., 2023; Zhou et al., 2021). Explanation methods can be categorized into three groups: ante-hoc (pre-model), in-model, and post-model (model-agnostic) (Zhou et al., 2021). These groups reflect the timing of explanation generation relative to the AI model development process. Explanations can also be classified based on the scope of explanations. For example, global explanations provide clarifications at the dataset level (Buomsoo et al., 2023). Examining the interplay of these explanation methods with different AI models - from transparent models to opaque "black-box" models - underlines the adaptability of explanation methods to different AI systems. This adaptability is central for tailoring explanations to the specific needs of various user groups, thus enhancing the effectiveness of the sensemaking process.

3.2. Users

In our theory development, it is imperative to address the diverse users that interact with AI. (Dhanorkar et al., 2021). As shown in Table 1, various user groups, each with a distinct focus and AI literacy levels, have been identified in the literature (Meske et al., 2022). These papers emphasize the importance of adapting explanations to suit the information needs and varying expertise levels of end-users, ensuring accessibility without oversimplification (Gerlings et al.,

2021; Simkute et al., 2021). For instance, end-users with little expertise in AI, such as consumers utilizing AI-driven recommendation systems, benefit from straightforward explanations (Mohseni et al., 2021). This enhances their trust in AI decisions (Ehsan et al., 2021). Similarly, domain experts, like clinicians using AI for medical diagnostics, require detailed explanations that bridge AI outputs with clinical reasoning to facilitate informed decision-making and reinforce trust in the technology.

Type	Focus	AI literacy	Selected references
End-user	Explanations that are accessible	Low Novice	(Dhanorkar et al., 2021; Haque et al., 2023; Hoffman et al., 2023; Laato et al., 2022; Meske et al., 2022; Mohseni et al., 2021)
AI End-User/ Domain Expert	Explanations to gain scientific knowledge	Medium / Data Experts	(Alarcon & Willis, 2023; Benjamin et al., 2022; Ehsan et al., 2021; Holzinger et al., 2019; Meske et al., 2022; Mohseni et al., 2021)
AI Manager	Explanations to protect users from false decisions.	Medium	(Hoffman et al., 2023; Meske et al., 2022)
AI Regulator	Explanations to assess compliance with regulatory standards.	Medium	(Hoffman et al., 2023; Meske et al., 2022; Tomsett et al., 2018)
AI Developer	Explanations to evaluate and improve the AI. Quality assurance.	High/AI Expert	(Adadi & Berrada, 2018; Meske et al., 2022; Mohseni et al., 2021; Tomsett et al., 2018)

Table 1. Principal types of users

In extant research, the significance of making explanations accessible is emphasized, ensuring that they are not only technically accurate but also comprehensible by non-experts (Chromik & Schuessler, 2020). It aims to democratize AI, making it usable and trustworthy across various users. This reveals significant differences in explanation needs and preferences among different groups, which informs the development of more user-centric XAI systems (Haque et al., 2023). We argue that a user-focused approach is crucial for the successful integration of AI technologies into diverse real-world applications, enhancing user understanding, trust, and satisfaction with AI systems across a broad spectrum of users.

3.3. Explainability as a Sensemaking Process

3.3.1 XAI-Human interactions framed as a sensemaking process. Guiding principles for effective human-AI interaction have been a focus of research. A critical aspect of these interactions is ensuring that users understand not only the AI system's capabilities but also the reasoning behind its outputs (Amershi et al., 2019). This focus on explainability can be addressed through different human-AI interaction models (see Table 2). These models can be broadly categorized into two main groups based on the information flow: one-way and multi-directional interactions.

In these interactions, the focus on explainability aligns with the concept of sensemaking, which involves actively processing information to achieve an understanding of complex situations (Pirolli & Russell, 2011). To date, sensemaking models do not include non-human actors Nonetheless, Weick's sensemaking theory (Weick, 1995) offers a valuable lens for understanding AI-human interactions. Indeed, this theory emphasizes that people construct meaning from ambiguous information, prioritizing plausible narratives over absolute accuracy. In the context of XAI, this suggests that interpretability goes beyond simply providing perfect explanations. On the one hand, when faced with AI outputs, humans can enact sensemaking by actively asking questions, seeking diverse explanations, and considering the AI's limitations. This critical analysis allows users to weave AI's insights with their knowledge and experience, allowing for a richer understanding of the information presented. On the other hand, XAI can improve how humans and machines collaborate by understanding how people make sense of complex situations. For example, AI systems could enact sensemaking by tailoring explanations to the user. Providing multiple perspectives, mimicking how humans build understanding from incomplete information, allows users to actively participate in constructing meaning from the AI's outputs.

While sensemaking theory offers a powerful framework for human-machine interaction, it remains largely untapped within XAI research (Kaur et al., 2022). This gap presents a significant opportunity to bridge the explainability gap by understanding how

people, not just machines, construct meaning from AI outputs (Kaur et al., 2022; Shen et al., 2021). A key weakness of the few papers that mobilize sensemaking as part of the extant research is that these papers do not delve into how the AI system itself enacts sensemaking. Ideally, a comprehensive approach would explore how the AI reasons about the situation, gathers information, and tailors explanations to the user's background and understanding. If the focus remains solely on how humans make sense of pre-determined explanations, it does not fully take into account the interactions between the AI and its users.

Type	Characteristic	Selected Sources
One-way interaction (Interpretability, explanation, trace, justification, and strategy, transparency)	A one-way direction where the AI provides explanations to the users.	(Biran & Cotton, 2017; Ribeiro et al., 2016; Ye & Johnson, 1995)
Multi-directional interaction (Explanatory debugging, explanation & feedback, interactivity)	A multi-directional and dynamic communication exchange involving multiple entities, where information is co-created through feedback loops in which the AI explains predictions to users, who respond with necessary corrections.	(Amershi et al., 2019; Chromik & Schuessler, 2020; Kulesza et al., 2015; Martens & Provost, 2014; Shen et al., 2021; Smith-Renner et al., 2020; Stumpf et al., 2009)

Table 2. Types of Human-AI Interactions

3.3.2 Making Sense of AI: The Human Challenge. XAI presents a unique challenge – interpreting its outputs that are too often deemed enigmatic for various users (Benjamin et al., 2022). As explained above, humans can rely on sensemaking to tackle this complexity. Sensemaking goes far beyond simply finding information (Pirolli & Russell, 2011). It is an active process of grappling with new and ambiguous situations. This involves gathering information, actively learning new domains, solving complex problems, and most importantly, building a shared understanding.

At its core, sensemaking is about creating coherence – a web of meaning – by connecting observations, beliefs, and actions (Schildt et al., 2020). Humans rely on existing mental frameworks, like categories or assumptions, to interpret the world (Weick, 1993). These frameworks can be incredibly helpful, allowing for quick comprehension and decision-making. However, they can also be a double-edged sword. Frameworks can limit understanding (Weick, 1993), leading to missed possibilities or clinging to outdated perspectives, especially when faced with the complexities of XAI outputs.

There are various ways to understand how humans engage in sensemaking. One theory suggests a cost-benefit analysis (Russell et al., 1993). Here, people construct mental models – simplified representations of reality – to answer specific questions. These models help us understand situations and make decisions efficiently. However, these models might not be sufficient for the unfamiliar territory of XAI outputs.

Another perspective focuses on how mental frameworks, like data frames, organize information (Klein et al., 2006). These frameworks act as filters, influencing what information is noticed and how it's interpreted. They can also help understand past events and predict future ones. However, in the context of XAI, where the inner workings of the AI system might be opaque, these frameworks can lead to misinterpretations or missed crucial details.

The social aspect of sensemaking is crucial, especially when dealing with the complexities of XAI. Sharing information, discussing interpretations, and revising understanding based on others' insights helps overcome the limitations of individual frameworks (Weick, 1995). This collaborative approach is even more important in the context of XAI, where interpreting AI outputs can be challenging. By working together, people can leverage the strengths of different perspectives and arrive at a more nuanced and accurate understanding of what the AI is communicating.

3.3.3. AI-powered Sensemaking: Unveiling the Black Box. Within the field of AI, the concept of sensemaking – the active processing of information to achieve comprehension of intricate situations – has typically been viewed as an exclusively human capability. A groundbreaking possibility emerges in the potential to endow AI with its sensemaking capacity. An AI system capable of not only excelling at information processing but also critically evaluating its reasoning processes can add value to the challenge of explainability. Such innovative development would empower AI to generate more comprehensive explanations of its outputs. Consequently, a more collaborative human-AI problem-solving environment could emerge, capitalizing on the strengths of both human intuition and AI's analytical capabilities.

One way AI can leverage sensemaking is through identifying inconsistencies. Drawing inspiration from cost-based sensemaking models (Russell et al., 1993), AI can construct multiple models to explain a situation. By analyzing these models for inconsistencies, AI can

achieve a form of self-reflection, similar to how humans might recognize inconsistencies in their thought processes (Weick, 1993). This self-reflection can flag potential biases or limitations in the AI's understanding, leading to more robust and trustworthy explanations. Another powerful technique is analyzing data from multiple perspectives. Just like humans use different mental frameworks to understand information (Klein et al., 2006), AI systems can be programmed with various data structures. This allows the AI to analyze information from multiple angles, highlighting potential biases inherent in any single approach (Schildt et al., 2020). The presentation of AI conclusions alongside alternative explanations based on diverse data structures could present a valuable opportunity. This approach would indeed foster a more nuanced understanding for the human user, enabling them to critically evaluate the rationale underlying the AI's outputs. Social interaction plays a crucial role in human sensemaking (Weick, 1995). AI systems can be designed to leverage this principle by learning from past interactions. By analyzing past interactions with users, the AI can gain insights into their goals and expectations (Miller, 2019). This would allow the AI to tailor its explanations, using language and examples that resonate with the user's specific needs. Fundamentally, such a process echoes the importance of the social exchange of knowledge in human sensemaking, leading to a more effective and collaborative interaction.

Finally, humans are susceptible to confirmation bias, focusing on information that confirms existing beliefs (Weick, 1995). AI systems can be programmed to address this by actively surfacing hidden data. Similar to how some sensemaking models address the issue of "residue" (Calvard, 2016), AI can be designed to identify data that contradicts its initial conclusions. Presenting this "contradictory evidence" alongside the main explanation prompts humans to question assumptions and arrive at a more comprehensive understanding. By incorporating these sensemaking techniques, the AI could explain its reasoning, which has the potential to foster trust and collaboration between humans and AI. The development of an AI system capable of not only generating solutions but also explicating its reasoning processes in a transparent, unbiased, and user-centric manner would thus present a significant advancement towards explainability. Such an AI system would possess the ability to provide clear explanations of its decision-making processes, mitigating potential biases and tailoring the explanations to the specific understanding of the human user. This advancement has the potential to revolutionize human-AI interaction, fostering trust and collaboration in problem-solving endeavors. Such a collaborative sensemaking process, where AI and humans leverage their respective strengths – human intuition and AI's analytical power (Naikar & Elix, 2019) – represents a promising approach.

4. Discussion and Future Research

The growing and diverse literature on XAI underscores the evolving nature of this field and the ongoing efforts to establish standardized practices and frameworks. Our analysis of the literature revealed that there is confusion and a lack of convergence in extant research (Amann et al., 2022; Smith et al., 2023). To ensure XAI knowledge building in the IS field (Barki, 2008), this paper aims to contribute to the conceptualization of AI explainability, by providing a comprehensive yet parsimonious framework that shows that the very existence of XAI cannot exist outside the interaction between the AI and the users. Specifically, it can inform the development of XAI research that is not only technically robust but also adjusted to the cognitive and contextual needs of users across various domains. Such contributions are expected to advance the discourse on human-AI interaction, emphasizing the importance of user engagement, understanding, and trust in the effective deployment of AI technologies. By integrating sensemaking into the XAI narrative, this research underscores the role of human cognition and context in interpreting AI outputs, thereby promoting the design of explanations that are meaningful and actionable for users.

All in all, our analysis of extant literature and the theory development efforts that led to our proposed framework pave the way to a rich and meaningful research agenda (Rivard, 2024) for XAI's future research in the IS domain. One of our objectives is to identify fruitful areas for development, but avoiding to add to the current confusion and complexity around XAI in the extant literature. Hence, we chose to frame avenues for research development around the key elements of our framework.

Core in our framework is the sensemaking process that embeds the AI and its explanations, the users, and their interaction. *First*, concerning AI and its explanations, a critical future direction lies in achieving a balance between the comprehensiveness and fidelity of explanations and the preservation of optimal AI performance. In terms of AI’s explanations, future research can explore the optimal timing of explanations (upfront, after decisions, or both) as well as the ideal amount of explanation that should be provided, knowing the risk of overwhelming the users. Avoiding potential biases embedded within the AI's reasoning is also vital. Developing robust methods and metrics to measure the quality of explanations – such as clarity, trustworthiness, and addressing user concerns – is

another area that can lead to critical implications for XAI research (Miller, 2019). *Second*, our framework identifies the users as essential partners in the explainability challenge. Future research could focus on tailoring explanations to user characteristics like background, skills, and experience. This might include user profiling for adaptive explanations, developing literacy-aware explanation methods, calibrating explanations based on prior experience, and creating evaluation frameworks that assess explanation effectiveness for diverse user groups. By personalizing explanations, XAI can provide better user interactions with AI technologies and, as a result, better explainability. *Third*, we showed that the interaction between the human and the machine is paramount in the quest for explainability. A critical future direction in XAI research lies at the intersection of explainability and actionable decision-making, particularly in refining the human-AI partnership. While explaining the "what" of an AI model's actions is well-established, less focus has been placed on how users can leverage this information for informed collaboration. Future research can delve into the actionability of the sensemaking process. This may include exploring the optimal direction and level of user interaction with explanations. Other research could see if effective XAI design might consider whether explanations should provide a prescriptive course of action or a broader framework that facilitates collaborative sensemaking. Another avenue could be to see how AI-human interactions can allow users to harness the power of AI while maintaining control and accountability.

In our framework, we also identified three other key elements that influence the sensemaking process that leads to explainability, or lack thereof. *First*, a crucial avenue for XAI research lies in the influence of context on explainability. Different domains present unique challenges. *Second*, another promising avenue for XAI research is the necessity to delve into the fundamental drivers behind the need for explainability. Understanding the specific reasons for explanations can lead to more targeted XAI solutions. Our *third* influential element lies in the outcomes of the sensemaking process that leads to explainability. From this perspective, a potential area for XAI research involves comprehensively understanding the downstream impacts of explainability. While the absence of transparency is likely to hinder accountability and erode user trust, the actual influence of XAI itself warrants further investigation.

Finally, above and beyond these six elements of our framework that offer opportunities for research development, we want to urge researchers to consider two additional areas for future research. *First*, the extant discourse surrounding XAI is often hampered by confusion and a lack of standardized terminology. To bridge this gap, future research should prioritize a more rigorous conceptualization of core XAI terms. This includes, for example, establishing clear distinctions between explainability and interpretability. Additionally, research should clarify the relationships between other interconnected concepts like transparency, understandability, comprehensibility, accountability, and responsibility. *Second*, XAI research lies in its scope and methodologies. In this paper, we focused on the individual user's interaction with AI's explanations. Future research could explore explainability across different levels of analysis, including organizational, industry-wide, governmental, and societal contexts. Furthermore, future XAI research might rely on true interdisciplinary collaboration. While the technical methods of explanation development fall within the realm of computer science, a truly comprehensive understanding of XAI necessitates a broader perspective. By fostering collaboration across disciplines and levels of analysis, XAI research can ensure the development of explanations that are not just technically sound, but also user-centric, trustworthy, and impactful.

This research also identifies some potential limitations in adopting the sensemaking approach. *First*, adopting the sensemaking paradigm might become more limited in scenarios where the direct human decision-maker is not in charge of the full process flow. For example, in situations where straightforward translations of predictions into decisions might be complicated by a plethora of actors involved. Since sensemaking relies heavily on individual and collective interpretation natural variation in understanding can pose challenges when striving for unified and consensus-based decisions. What one person sees as a key insight, another might interpret differently based on their background and viewpoint. This inherent diversity of thought, while valuable, requires careful navigation to ensure everyone is working towards a shared understanding. *Second*, while AI can augment sensemaking, over-reliance on AI systems can inadvertently amplify existing biases and diminish the role of human intuition and experience, potentially leading to issues within the decision-making framework. Thus highlighting the need to prioritize human-centered design principles when designing AI sensemaking systems. *Third*, the absence of widely accepted ethical frameworks specifically designed for human-AI collaboration in decision-making further complicates human-AI sensemaking. Clear guidelines are essential to ensure responsible development, deployment, and use of AI in sensemaking. By acknowledging these limitations and proactively addressing them, organizations can harness the power of

human-AI sensemaking while mitigating potential drawbacks.

5. Conclusion

This paper aligns with broader initiatives in responsible and accountable AI research. The quest for XAI resonates with the overarching goal of developing AI systems that are not only efficient and accurate but also accountable and responsible in their decision-making processes. As the ethical dimensions of AI gain prominence, addressing the explainability dimension that underlies XAI emerges as a critical component of ensuring that AI systems align with societal values and expectations. We hope that this paper's contributions help move in this direction.

6. References

Abedin, B. (2021). Managing the tension between opposing effects of explainability of artificial intelligence: A contingency theory perspective. *Internet Research*, *32*(2), 425–453.

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Alarcon, G., & Willis, S. (2023). Explaining Explainable Artificial Intelligence: An integrative model of objective and subjective influences on XAI. *Proceedings of the 56th Hawaii International Conference on System Sciences*, 1095–1104.

Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., Gilbert, T. K., Hagendorff, T., Holm, S., Livne, M., Spezzatti, A., Strümke, I., Zicari, R. V., Madai, V. I., & on behalf of the Z-Inspection initiative. (2022). To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, *1*(2), e0000016.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

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

Barki, H. (2008). Thar's Gold in Them Thar Constructs. *The DATA BASE for Advances in Information Systems*, *39*(3), 9–20.

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Information Systems Research*, *34*(4), 1582–1602.

Benjamin, J. J., Kinkeldey, C., Müller-Birn, C., Korjakow, T., & Herbst, E.-M. (2022). Explanation Strategies as an Empirical-Analytical Lens for Socio-Technical Contextualization of Machine Learning Interpretability. *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP), 1–25.

Berente, N., Bin Gu, Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, *45*(3), 1433–1450.

Biran, O., & Cotton, C. V. (2017). Explanation and Justification in Machine Learning: A Survey. *IJCAI Workshop on Explainable Artificial Intelligence (XAI)*, 8–13.

Bitzer, T., Wiener, M., & Cram, W. A. (2023). Algorithmic Transparency: Concepts, Antecedents, and Consequences – A Review and Research Framework. *Communications of the Association for Information Systems*, *52*, 293–331.

Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, *33*(1), 26.

Buomsoo, K., Srinivasan, K., Sung Hye Kong, Jung Hee Kim, Chan Soo Shin, & Ram, S. (2023). Rolex: A Novel Method for Interpretable Machine Learning Using Robust Local Explanations. *MIS Quarterly*, *47*(3), 1303–1332.

Calvard, T. S. (2016). Big data, organizational learning, and sensemaking: Theorizing interpretive challenges under conditions of dynamic complexity. *Management Learning*, *47*(1), 65–82.

Capel, T., & Brereton, M. (2023). What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–23.

Chromik, M., & Schuessler, M. (2020). A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. *Proceedings of the IUI Workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies*, 1–7.

Dabas, V., Thomas, A., Khatri, P., Iandolo, F., & Usai, A. (2023). Decrypting Disruptive Technologies: Review and Research Agenda of Explainable AI as a Game Changer. *Proceedings of the International Conference on Technology Management, Operations and Decisions*, 1–6.

Dhanorkar, S., Wolf, C. T., Qian, K., Xu, A., Popa, L., & Li, Y. (2021). Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. *Designing Interactive Systems Conference 2021*, 1591–1602.

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2021). *The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations* (arXiv:2107.13509). arXiv.

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020). Fostering Human Agency: A Process for the Design of User-Centric XAI Systems. *ICIS 2020 Proceedings*.

Gerlings, J., Shollo, A., & Constantiou, I. (2021). Reviewing the Need for Explainable Artificial Intelligence (xAI). *Proceedings of the 54th Hawaii International Conference on System Sciences*, 1284–1293.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89.

Haque, A. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, *186*, 122120.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, *16*, 45–74.

Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining Explanation For "Explainable AI." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 197–201.

Hoffman, R. R., Mueller, S. T., Klein, G., Jalaeian, M., & Tate, C. (2023). Explainable AI: Roles and stakeholders, desirements and challenges. *Frontiers in Computer Science*, *5*.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). *What do we need to build explainable AI systems for the medical domain?* (arXiv:1712.09923). arXiv.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, *9*(4), e1312.

Kaur, H., Adar, E., Gilbert, E., & Lampe, C. (2022). Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 702–714.

Kim, M., Kim, S., Kim, J., Song, T.-J., & Kim, Y. (2024). Do stakeholder needs differ? - Designing stakeholder-tailored Explainable Artificial Intelligence (XAI) interfaces. *International Journal of Human-Computer Studies*, *181*, 103160.

Klein, G., Moon, B., & Hoffman, R. R. (2006). Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems*, *21*(4), 70–73. https://doi.org/10.1109/MIS.2006.75

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of Explanatory Debugging to Personalize Interactive Machine Learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 126–137.

Laato, S., Tiainen, M., Najmul Islam, A. K. M., & Mäntymäki, M. (2022). How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research*, *32*(7), 1–31.

Lee, K., & Ram, S. (2023). Explainable Deep Learning for False Information Identification: An Argumentation Theory Approach. *Information Systems Research*, *0*.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43.

Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. *MIS Quarterly*, *38*(1), 73–100.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, *39*(1), 53–63.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, *11*(3–4), 1–45.

Naikar, N., & Elix, B. (2019). Designing for self-organisation in sociotechnical systems: Resilience engineering, cognitive work analysis, and the diagram of work organisation possibilities. *Cognition, Technology & Work*, *58*, 1–15.

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183–199.

Pirolli, P., & Russell, D. M. (2011). Introduction to this Special Issue on Sensemaking. *Human–Computer Interaction*, *26*(1–2), 1–8.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938*.

Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? A proposal of User-Centered Explainable AI. *Joint Proceedings of the ACM IUI 2019 Workshops.*, 1–7.

Rivard, S. (2024). Unpacking the process of conceptual leaping in the conduct of literature reviews. *The Journal of Strategic Information Systems*, *33*(1), 101822.

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, 269–276.

Rzepka, C., & Berger, B. (2018). User Interaction with AI-enabled Systems: A Systematic Review of IS Research. *ICIS 2018 Proceedings*, 1–17.

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*, 110273.

Schildt, H., Mantere, S., & Cornelissen, J. (2020). Power in Sensemaking Processes. *Organization Studies*, *41*(2), 241–265.

Shen, H., Liao, K., Liao, Z., Doornberg, J., Qiao, M., Van Den Hengel, A., & Verjans, J. W. (2021). Human-AI Interactive and Continuous Sensemaking: A Case Study of Image Classification using Scribble Attention Maps. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–8.

Simkute, A., Luger, E., Jones, B., Evans, M., & Jones, R. (2021). Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable. *Journal of Responsible Technology*, *7–8*, 100017.

Smith, M., Acquesta, E., Smutz, C., Rushdi, A., & Moss, B. (2023). Assessing the Fidelity of Explanations with Global Sensitivity Analysis. *Proceedings of the 56th Hawaii International Conference on System Sciences*, 1085–1094.

Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S., & Findlater, L. (2020). No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Stoffels, D., Faltermaier, S., Strunk, K. S., & Fiedler, M. (2022). Opening the Black-Box of AI: Challenging Pattern Robustness and Improving Theorizing through Explainable AI Methods. *ICIS2022 Proceedings*.

Stumpf, S., Rajaram, V., Li, L., Wong, W.-K., Burnett, M., Dietterich, T., Sullivan, E., & Herlocker, J. (2009). Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, *67*(8), 639–662.

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *2018 ICML Workshop on Human Interpretability in Machine Learning*.

Waardenburg, L., & Huysman, M. (2022). From coexistence to co-creation: Blurring boundaries in the age of AI. *Information and Organization*, *32*(4), 100432.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii–xxiii.

Weick, K. (1993). The Collapse of Sensemaking in Organizations: The Mann Gulch Disaster. *Administrative Science Quarterly*, *38*(4), 628–652.

Weick, K. (1995). *Sensemaking in Organizations* (1 edition). Sage Publications.

Ye, L. R., & Johnson, P. E. (1995). The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice. *MIS Quarterly*, *19*(2), 157–172.

Zhang, Q., Hall, M., Johansen, M., Galetic, V., Grange, J., Quintana-Amate, S., Nottle, A., Jones, D. M., & Morgan, P. L. (2022). Towards an Integrated Evaluation Framework for XAI: An Experimental Study. *Procedia Computer Science*, *207*, 3884–3893.

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, *10*(5), Article 5.

Züllig, K., Hühn, P., Bohnen, E., & Obermeier, A. (2023). Tell Me Why (I Want It That Way) – Effects of Explanations and Online Customer Reviews on Trust in Recommender Systems. *Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS2023*. ICIS.