IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

Interpretable Artificial Intelligence in Information Systems: Status Review and Future Research Directions

Saurabh Sudhakar Umredkar, Swapnil Anil Bagde, Sonu Ramkumar Shahu, Prof Nikita Khanzode
Tulsiramji Gaikwad Patil College of Engineering and Technology, Nagpur, India

Abstract: *Efforts to develop black-box artificial intelligence (AI) systems have become a phenomenon of emerging global interest in academia, business, and society, and have led to the development of the XAI research field. With its pluralistic perspective, information systems (IS) research is destined to contribute to this emerging field; thus, it is not surprising that the number of research publications at XAI has increased significantly. This paper aims to provide a comprehensive overview of XAI research in public and electronic markets, specifically using a structured literature review. Based on a literature review of 180 research papers, this work examines the most receptive points, the development of academic debates, and the most important concepts and methodologies. In addition, eight research areas with different levels of maturity in e-markets are identified. Finally, guidelines for the XAI research agenda in IS are presented.*

Keywords: Artificial intelligence.

I. INTRODUCTION

Artificial intelligence (AI) is already ubiquitous in work and everyday life: in the form of various technologies such as natural language processing or image recognition (Abdul et al., 2018; Berente et al., 2021) and in various application areas . , including electronic markets, finance, health, human resources, public administration and transportation (Collins et al., 2021; Meske et al., 2020). The availability of AI is expected to increase with around 70% of companies in the world intending to adopt AI by 2030 (Bughin et al., 2018). Therefore, AI is expected to change all aspects of society (Collins et al., 2021; Makridakis, 2017).

CEO of Alphabet Inc. now expect AI to be "more effective for humanity than fire, electricity, and the Internet" (Knowles, 2021). AI has great potential to achieve extraordinary efficiency and new data processing capabilities (Asatiani et al., 2021) and even exceed human performance in certain tasks (Meske et al., 2022). For example, AI has outperformed doctors in diagnosing breast cancer (eg, McKinney et al., 2020). At the same time, the use of AI is associated with serious risks, especially ethical issues such as ambiguity, fairness, justice and discrimination, and legal issues such as accountability, regulation and responsibility (Akter et al., 2021a; Asatiani et al. , 2021; Berente et al., 2021). The potential negative consequences of the use of AI affect not only individuals and organizations, but also society as a whole (Mirbabaie et al., 2022; Robert et al., 2020). For example, Robodebt, an AI-based debt recovery program, has claimed almost $2 billion from more than 400,000 Australian citizens (Australian Broadcasting Corporation, 2022). There is concern that the use of AI may exacerbate social or economic inequality (Gianfrancesco et al., 2018). For example, Amazon.com Inc. The AI-based recruitment engine used by Twitter Inc. downgrading women's resumes in favor of male candidates (Gonzalez, 2018), Twitter Inc. There is AI and AI controlled by. It is used by Google LLC, which provides racist results in image searches (Yampolsky, 2019).

The growing capabilities of AI models contribute to transparency in operations and results that cannot be interpreted by humans (Berente et al., 2021). Openness, on the one hand, can lead people to rely on AI results and replace their decisions with false decisions (Robert et al., 2020). On the other hand, lack of explanation can lead to reluctance to use AI. In the case of maternal diseases, AI-based decision support systems may fail to detect certain diseases, for example, due to biased training data. A doctor who shows correlation will not be able to detect this error; Doctors who do not trust AI systems and refuse to use them will not benefit from decision support.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

Interpretable AI (XAI) aims to reduce the risk of AI by exploiting its potential and improving interpretability. XAI aims to empower human stakeholders to understand, trust, and effectively manage AI (Arrieta et al., 2020; Langer et al., 2021). An explanation on the example of breast cancer diagnosis can help doctors understand the operation and results of AI-based decision support systems. So, it can help you have more confidence in the system's decisions and detect its errors. Ultimately, a partnership between doctors and AI can make better decisions than doctors or AI. It appears at different societal levels to improve understanding of AI systems. Companies are striving to make AI systems more intuitive (eg, Google, 2022; IBM, 2022). Regulators are taking steps to demand accountability and transparency of AI-based decision-making processes. For example, the European General Data Protection Regulation (GDPR) guarantees a "right to interpretation" for those affected by algorithmic decisions (Selbst & Powles, 2017). The EU's upcoming AI regulation requires human supervision to interpret and compete the results of AI systems in "high-risk" applications such as recruitment or credit assessment (European Commission, 2021). The economic and social importance of XAI has attracted the attention of researchers, appearing in a number of publications in recent years (Arrieta et al., 2020). For example, XAI researchers are working to unlock the functionality of AI-based applications such as user-friendly cancer detection systems (Kumar et al., 2021) and malware prediction systems (Iadarola et al., 2021). In addition, they investigate approaches to automatically generate annotations of AI decisions that can be used independently of existing AI models. Examples of use cases include credit risk assessment (Bastos & Matos, 2021) or fraud detection (Hardt et al., 2021). Information Systems (IS) Research.

II. THEORETICAL BACKGROUND AND RELATED WORK

Theoretical Foundations

Given that IT research investigates and explains "the interaction of individuals, groups, organizations and markets with IT" (Sidorova et al., 2008, p. 475), human-AI interaction is an important research topic for this course. In general, human-system communication occurs between an IT system and a user who wants to perform a specific task in a specific context (Rzepka & Berger, 2018). Business characteristics are defined by context, users and IT systems (Rzepka & Berger, 2018). When the human partner is an AI system, special features of the AI system must be considered. With the limits of computing capabilities constantly expanding, modern AI systems provide greater autonomy, deeper learning capabilities, and greater obscurity than previously studied IT systems (Baird & Maruping, 2021; Jiang et al., 2022). Rapid progress in AI mainly contributes to the development of machine learning (ML), defined as the ability to learn specific problems by building models based on data processing (Russel & Norwig, 2021). The autonomy and learning capabilities of ML-based AI systems further increase ambiguity (Berente et al., 2021). So, with ever-increasing levels of AI autonomy, learning ability, and obscurity, challenges arise to manage human-AI interactions.

From a managerial perspective, ambiguity has four interdependent values: openness, transparency, interpretation, and explanation (Berente et al., 2021). First, transparency is a property of AI systems and refers to the complex nature of AI that prevents humans from understanding the thought processes involved in AI (Meske et al., 2020). Most AI systems are "black boxes", meaning that the reasons for their results are often not clear to humans, not only for users but also for developers (Guidotti et al., 2019; Merry et al., 2021). A prominent example of this is the nervous system. Second, transparency refers to the willingness of AI system owners (parts) to open up and thus is considered a strategic management issue (Granados et al., 2010). Third, interpretability is a property of the AI system, which refers to the ability of the system to be understood at least to some extent by at least some parties (Gregor &Benbasat, 1999). Finally, interpretability means that an AI system can be understood from a human perspective. An AI system with some degree of explanation can be sufficiently explained to one person, but not necessarily to another (Berente et al., 2021). For example, the decision tree becomes inexplicable to some users as the complexity increases (Mittelstadt et al., 2019). Transparency significantly affects human-AI interaction: It prevents humans from observing or learning about the AI system's decision-making process (Arrieta et al., 2020). Faced with a transparent system, people cannot develop appropriate beliefs; they often ignore the decisions and recommendations of the system or do not use the system (Herse et al., 2018; Rader & Gray, 2015). Thus, transparency hinders human agency and AI adoption. XAI's research area deals with the transparency of AI systems. XAI aims for an approach that makes AI systems more understandable, sometimes called intelligibility (Doran et al., 2018) - generating automatic explanations for behavior and results while

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

maintaining high AI performance (Adadi& Berrada, 2018; Gregor &Benbasat). , 1999). In daily human interaction, "interpretation is a social and iterative process between the interpretant and the interpretant" (Hromik& Butz, 2021, p. 1). This translates into the context of human-AI interaction, where annotations can explain why an AI system associates a given input with a specific product (Abdul et al., 2018). Thus, annotations can resolve the ambiguity of AI systems and improve interpretability from the user's perspective. The researcher emphasizes that clarifying the role of XAI can make an important contribution to the ongoing debate on human-AI interaction (Sundar, 2020).

Terminological foundations

The XAI research framework is driven by four main objectives (Adadi& Berrada, 2018; Arrieta et al., 2020; Gerlings et al., 2021; Gilpin et al., 2018; Langer et al., 2021; Meske et al., 2020; Wang et al., 2020) 2021; Meske et al., 2020; Wang et al., 2019). For example, assessment in this context is used to identify and prevent disparities in marginalized communities (Arrieta et al., 2020). The second goal is to create explanations that help improve AI systems. In this case, annotations can be used by developers to improve model accuracy by deepening the understanding of AI system performance (Adadi& Berrada, 2018; Arrieta et al., 2020; Gilpin et al., 2018; Langer et al., 2021; Meske et al. ., 2020). Third, provide explanations that justify AI system decisions by increasing transparency and accountability (Adadi& Berrada, 2018; Gerlings et al.; 2021; Meske et al., 2020; Wang et al., 2019). A notable example is the need for justification based on the "right to be explained" for those affected by algorithmic decisions (eg with GDPR); Another example refers to a decision made by an expert who follows the advice of an AI system but is responsible for the decision (Arrieta et al., 2020). Finally, to provide explanations that allow learning from the system, revealing unknown connections that indicate causal relationships in the underlying data (Adadi& Berrada, 2018; Langer et al., 2021; Meske et al., 2020). improve, justify and learn AI systems by making explanations (Abdul et al., 2018; DARPA, 2018).

To achieve this goal, XAI research provides a variety of approaches that can be classified along two dimensions: interpretation limits and dependency models (Adadi& Berrada, 2018; Arrieta et al., 2020; Vilone& Longo, 2020). The scope of interpretation can be global or local (Adadi& Berrada, 2018; Arrieta et al., 2020; Heuillet et al., 2021; Payrownaziri et al., 2020; Vilone& Longo, 2020). Global annotation targets the entire operation of the AI model. Using the example of a credit line decision, a global description can show the most important criteria used by the AI model to make a credit line decision. Local explanations, on the other hand, aim to rationalize specific outputs of AI models. Returning to the example of the credit line decision, local interpretation may provide the most important criteria for personal rejection or approval. The second dimension, AI model dependence, distinguishes between two approaches: model-specific and model-agnostic (Adadi& Berrada, 2018; Arrieta et al., 2020; Rawal et al., 2021). Model-specific approaches focus on providing explanations for AI models or classes of models (Arrieta et al., 2020; Rawal et al., 2021), such as neural networks (Montavon et al., 2018), by considering internal components. AI models (classes), such as structural data. The model-agnostic approach, on the other hand, ignores the internal components of the model and is therefore used in various AI models (Adadi& Berrada, 2018; Rawal et al., 2021; Ribeiro et al.; 2016; Vilone& Longo, 2020).

Designing or selecting the best XAI approach for a given problem is similar to solving a "human-agent interaction problem" (Miller, 2019, p. 5). Therefore, it is important to focus on the interpretation of the audience. Three main target groups are the focus of XAI research (Bertrand et al., 2022; Cooper, 2004; Ribera &Lapedriza, 2019; Wang et al., 2019). The first group includes developers who build AI systems, namely data scientists, computer engineers, and researchers (Bertrand et al., 2022; Ribera &Lapedriza, 2019; Wang et al., 2019). For example, using the credit policy decision example, this is the team that created the AI system or is responsible for supporting it. The second group includes domain experts who share skills based on formal education or professional experience in practical fields (Bertrand et al., 2022; Ribera &Lapedriza, 2019; Wang et al., 2019). In terms of loan decisions, it will be the bank's advisor for loan decisions. The last group includes individuals who are influenced by AI decisions among users (Bertrand et al., 2022; Cooper, 2004; Ribera &Lapedriza, 2019), such as the AI system recommendations of bank customers approved or rejected for loans (Mittelstadt et al. et al., 2019). In addition, this third group includes layer users who interact with AI, such as customers who explore credit lines with the help of AI-based agents.

To investigate how XAI approaches to overcome this "human interaction problem", the literature draws on three different approaches.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

Table 1 Key concepts in XAI research

Concept	Definition	Source
Dependency on the AI model		
Model-specific	Approaches that focus on providing explanations for specific AI models or model classes	Adadi & Berrada, 2018; Arrieta et al., 2020; Rawal et al., 2021
Model-agnostic	Approaches that disregard the underlying AI model's internal components and are thus applicable across a wide range of AI models	Adadi & Berrada, 2018; Rawal et al., 2021; Ribeiro et al., 2016; Vilone & Longo, 2020
Scope of explainability		
Global explainability	An explanation that targets explaining the functioning of the entire AI model	Adadi & Berrada, 2018; Arrieta et al., 2020; Heuillet et al., 2021; Payrovnaziri et al., 2020; Vilone & Longo, 2020
Local explainability	An explanation that focuses on rationalizing a specific outcome of an AI model	Adadi & Berrada, 2018; Arrieta et al., 2020; Heuillet et al., 2021; Payrovnaziri et al., 2020; Vilone & Longo, 2020
Explanation's target group		
Developers and AI researchers	Data scientists, computer engineers, and researchers who build or maintain AI systems	Bertrand et al., 2022; Ribera & Lapedriza, 2019; Wang et al., 2019
Domain experts	Experts who share expertise in the field of application based on formal education or professional experience	Bertrand et al., 2022; Ribera & Lapedriza, 2019; Wang et al., 2019
Lay users	Non-expert individuals who are affected by AI decisions or who interact with AI systems	Bertrand et al., 2022; Cooper, 2004; Ribera & Lapedriza, 2019
Explanation's goal		
Evaluate the system	Evaluate an AI system to detect its flaws and prevent unwanted behavior	Arrieta et al., 2020; Adadi & Berrada, 2018; Gerlings et al., 2021; Meske et al., 2020; Wang et al., 2019
Improve the system	Improve an AI system's accuracy by deepening the understanding of the AI system's functioning	Adadi & Berrada, 2018; Arrieta et al., 2020; Gilpin et al., 2018; Langer et al., 2021; Meske et al., 2020
Justify the system	Justify an AI system's decisions by improving transparency and accountability	Adadi & Berrada, 2018; Arrieta et al., 2020; Gerlings et al., 2021; Meske et al., 2020; Wang et al., 2019
Learn from the system	Learn from the AI system by identifying unknown correlations that could indicate causal relationships in the underlying data	Adadi & Berrada, 2018; Langer et al., 2021: Meske et al., 2020

Existing Literature Reviews on XAI

Several literature reviews refer to the growing body of research in the field of XAI using different foci and angles. While some of them aim to formalize XAI (For example, Adadi& Berrada, 2018), For example, using the nature and explanation of the cognitive system (Gregor &Benbasat, 1999), others provide a taxonomy; XAI in decision support (Nunes and Jannach, 2017) or research methods to interpret AI (eg Guidotti et al., 2019). Other literature reviews focus on different AI methods (X), such as rule-based models (e.g., Kliegr et al., 2021), neuro-fuzzy rule-making algorithms (e.g., Mitra & Hayashi, 2000), or neural networks (e.g., X). ,Heuillet et al., 2021) or specific annotation formats for reviews such as visual annotations (eg, Zhang & Zhu, 2018). Other literature review streams, user-friendly explanations (Chromik & Butz, 2021) or XAI user experience approaches (Ferreira & Monteiro, 2020) address user needs in XAI.

Other literature reviews conducted on XAI have focused on health (e.g., Amann et al., 2020; Chakrabartty& El-Ghayar, 2021; Payrownaziri et al., 2020; Tjoa & Guan, 2021), finance (e.g., Kute et al. , 2021). , 2021; Moscato et al., 2021) or vehicle (eg, Omeiza et al., 2021). For example, Amann et al. (2020) provided a comprehensive review of the role of AI interpretation in clinical practice to assess the implications of AI-based tools adopted in medicine. Omeiza et al. (2021) reviewed the XAI method in autonomous driving and provided a conceptual framework for the interpretation of autonomous vehicles. Other scholars apply XAI to related disciplines (e.g., Abdul et al., 2018; Miller, 2019). For example, Miller (2019) states in an often cited paper that XAI research can draw insights from the social sciences. The authors review articles in philosophy and psychology that examine how people define, create, choose, evaluate, and offer interpretations, and the cognitive and social norms that play a role. Thus, most literature reviews describe existing research gaps and focus on future research directions.

As mentioned above, the existing literature review covers several aspects of XAI research. However, to the best of our knowledge, none of them have provided a comprehensive literature review on XAI research in IS. Our literature review aims to address this gap.

Research Questions

Although computer scientists (Arrieta et al., 2020) have made significant progress in XAI, interest in this direction among AI scientists has grown rapidly in recent years (Meske et al., 2020). For example, there is an increase in the number of calls for Phones (for example, special issues about interpretable and responsible artificial intelligence in the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

electronic market, special issues about designing and managing human-AI interactions at the boundaries of information systems), according to the conference. track (For example, Minitrack on Artificial Intelligence Explained in Hawaii International Conference on Systems Engineering), and editorial (For example, editorial "Expl(AI) n This Me - Explanatory AI and Information Systems Research") in Engineering Information Systems). In their editorial, Bauer et al. (2021) emphasize that YD research focuses on XAI, given the multifaceted nature of the demands and the consequences of interpreting them from an individual and community perspective. In addition, in the research note summarizing the existing IS journal article, Meske et al. (2020) call for a resurgence of interpretive research in ED after an intensive study of explanations for a more transparent knowledge system. To our knowledge, there is no work that synthesizes XAI research in IS based on a structured and comprehensive literature review.

We conducted a structured and extensive literature review to provide a deeper understanding of the field of XAI research in the IS community. Our literature review addresses the following research questions (RQ):

RQ1: How can the academic discussion of XAI in the IS literature be characterized?

RQ2: What is the future direction of XAI research?

To answer the first research question, we aim to (i) identify IS publishing outlets that host XAI research, (ii) explain how the academic discussion of XAI in IS literature has developed over time, (iii) analyze key concepts and methodologies. of the academic discussion of XAI in the IS literature. direction and (iv) represents the most important XAI research area in the IS literature. To address the second research problem, we aim to focus on IS for the XAI research agenda.

III. LITERATURE REVIEW APPROACH

Based on the previous discussion, we learn how scientists conduct XAI research. We not only summarize but also analyze and critically review the state of XAI research in IS (Rowe, 2014). This analysis requires a systematic and structured literature review (Bandara et al., 2011; Webster & Watson, 2002). In preparation, it is necessary to use a comprehensive and repeatable literature search strategy that includes relevant journals and conferences, relevant keywords and an adequate time frame (Brock et al., 2009). Bandara et al. . 2006). We added a third step to systematically analyze articles based on XAI theory and IS methodology, and code articles related to relevant concepts in the literature (Beese et al., 2019; Jiang & Cameron, 2020).

Source selection

Research literature should include leading journals that are known for their high quality so that the most important research contributions will be published (Webster & Watson, 2002). The Association for Popular Information Systems (AIS), with members from approximately 100 countries, publishes peer-reviewed journals as well as journals recommended by special interest groups (SIGs). In our search, we included eight journals in AIS Senior Basketball Scholars and 64 AIS SIG Recommended journals. We consider all journals in the AIS eLibrary (including Affiliated and Chapter journals) because of their high quality. Various ratings are useful for identifying high-quality journals (Actor et al., 2021b; Levy & Ellis, 2006; Brock et al., 2009). We are clearly considered a journal of three popular ratings: First, Association of Business Schools (ABS)/Academic Journal Guide (AJG) 2021 (rating 3/4/4 * level, Information Management category). Second, the journal from the Australian Business Deans' Council (ABDC) Journal Quality List (rating A/A* grade, Information System category). Third, the VHB-JOURQUAL3 journal of the German Academic Business Association (level A + / A / B, category "Information Systems").

In addition, inclusion of high-quality conference proceedings (Webster & Watson, 2002) is recommended, especially when analyzing emerging and emerging research areas such as XAI. Consultations are a place for the generation of ideas and the development of new research plans (Levy & Ellis, 2006; Probst et al., 2013). Therefore, we have included the main international conference ED. More specifically, we reviewed the proceedings of four AIS conferences and the proceedings of twelve AIS-related conferences. In addition, we ensured the inclusion of all conferences from VHB-JOURQUAL3 (rating level A + / A / B, category "Information Systems").

We ended up with 105 journals and 17 conferences as sources for our search.

Search strategy and results

The development of XAI as a research field started in the 1970s and gained momentum in the past 5 to 10 years (Adadi& Berrada, 2018; Mueller et al., 2019). In order to gain an overview of the development of XAI research in IS,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

we chose to not limit the literature search's time frame. To identify relevant publications, we conducted a search using different terms describing XAI via databases that contain the journals and conferences discussed above. Based on terms that are used synonymously to describe research in the field of XAI (cf. Section "Theoretical background and related work"), we determined the following search string to cover relevant articles: ("explainable" AND "artificial intelligence") OR ("explainable" AND "machine learning") OR ("comprehensible" AND "artificial intelligence") OR ("comprehensible" AND "machine learning"). We searched for these terms in the title, abstract, and keywords. Where a search in title, abstract, and keywords was impossible, we applied a full-text search. Please see Fig. 1 for an overview of our search and screening process.

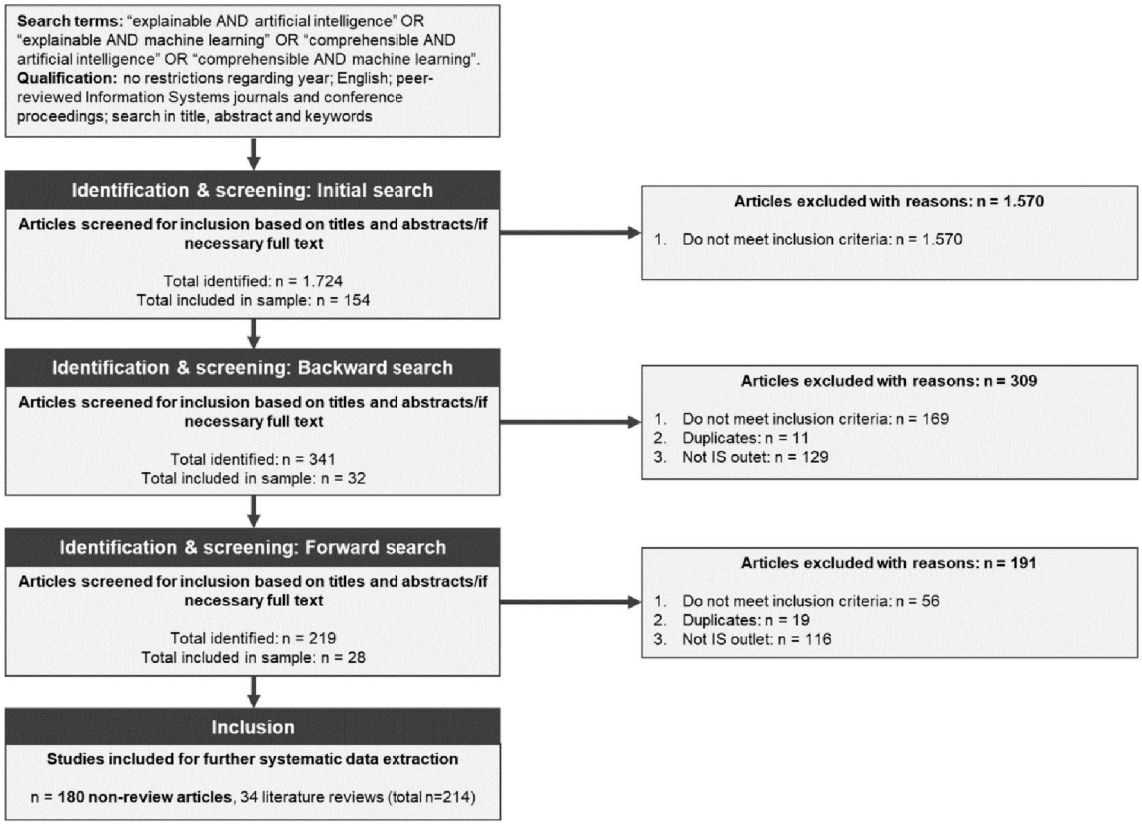


Fig. 1

Our January 2022 literature search yielded 1,724 papers. Papers were screened based on title and abstract, and the researcher read the full text when necessary. We excluded all papers that did not deal with XAI as defined above. More specifically, we exclude all papers focusing on AI without descriptive insights. For example, we put out a paper on how people can explain AI to other people. In addition, we publish papers that focus on the interpretation of "good old AI" such as experts or rule-based systems (Meske et al., 2020, p. 6). Unlike our understanding of AI, as defined in the introduction, this broad definition of AI includes self-explanatory systems such as knowledge-based or expert systems that do not suffer from problems such as lack of transparency.

Three researchers coded independently to determine a suitable set of documents and discussed coding disagreements to obtain consensus. At least two researchers analyzed each paper. Interrater reliability, measured by Cohen's Kappa, was 0.82- "approximate agreement" (Landis & Koch, 1977, p. 165). This process resulted in a set of 154 papers that served as a basis for going backward (32 papers) and forward (resulting in 28 papers) as suggested by Webster and Watson (2002). We reached a final set of 214 papers that served as the basis for further analyses.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

Analysis Scheme and Coding Procedure

Our goal is not only to summarize, but also to analyze and critique the state of XAI research in IS (Beese et al., 2019; Rowe, 2014). To do this, we first analyzed all the 34 papers that were provided only for the current knowledge review, which is the literature review. We then coded the remaining 180 articles using an analysis scheme derived from existing literature (see Terminological Framework section). More specifically, in our analysis, we distinguish between relevant theoretical concepts in XAI research and methodological concepts focused on IS research. Regarding the relevant concepts in the XAI literature, we distinguish between the dependence of the XAI approach on the AI model (Adadi& Berrada, 2018; Arrieta et al., 2020) and the explanatory scope (Adadi& Berrada, 2018; Arrieta et al., 2020; Payrownaziri et al. al., 2020; Vilone& Longo, 2020), as well as explanatory target groups (Ribera &Lapedriza, 2019; Wang et al., 2019) and targets (Meske et al., 2020). Regarding IS methodology, we distinguish between two general research paradigms: design science and behavioral science (Hevner et al., 2004). For the contribution of design science, we further define the types of artifacts according to Hevner et al. (2004) and the evaluation model based on the evaluation scenario defined for the XAI approach (Adadi& Berrada, 2018; Chromik & Schuessler, 2020; Doshi-Velez & Kim, 2018). This leads to the following analysis scheme (Figure 2):

Category	XAI conceptual dimensions			
Dependency on the AI model Adadi and Berrada 2018; Arrieta et al. 2019	**1.** Model-agnostic	**2.** Model-specific		
Scope of explainability Adadi and Berrada 2018; Arrieta et al. 2019; Payrovnaziri et al. 2020; Vilone and Longo 2020	**1.** Local explainability	**2.** Global explainability		
Explanation's target group Ribera and Lapedriza 2019; Wang et al. 2019	**1.** Developers	**2.** Domain experts	3. Lay users	
Explanation's goal Adadi and Berrada 2018; Meske et al. 2020	**1.** Evaluate the system	**2.** Improve the system	3. Justify the system	4. Learn from the system

Category	IS methodological dimensions			
Research paradigm Hevner et al. 2004	**1.** Behavioral Science	**2.** Design Science		
Artifact type Hevner et al. 2004	**1.** Construct	**2.** Model	3. Method	4. Instantiation
Evaluation type Adadi and Berrada 2018; Chromik and Schussler 2020; Doshi-Velez and Kim 2018	**1.** Functionally-grounded evaluation	**2.** Human-grounded evaluation	3. Application-grounded evaluation	

Fig. 2

Three researchers coded the remaining 180 articles according to the analytical scheme. Multiple parameters are possible for scaling. For a sample of 100 articles, each article was coded by at least two researchers. Interrater reliability, measured by Cohen's kappa, was 0.74, indicating "significant agreement" (Landis & Koch, 1977, p. 165). In case of disagreement, the researcher reached a consensus through discussion.

IV. RESULTS

This section is devoted to our results. First, we analyzed IS publishing shops interested in XAI research. Second, we examine the development of the academic discussion of XAI in the IS literature over time. Third, we analyze the basic concepts and methodological framework of the academic debate. Finally, we get the basic search direction XAI.

Receptive IS outlets to XAI research

We analyzed which journals and conferences accepted XAI research. The results are useful in three ways: they provide researchers and practitioners a point from which to find relevant research, they help researchers set targets, and they show editors how actively their papers contribute to academic debate; topic (Bandara et al., 2011). Forty-one articles were published in journals and 39 in conference proceedings. An overview of the number of journal and conference publications is presented in the Appendix.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

Development of the academic discussion on XAI in IS literature over time

To investigate the development of the academic discussion about XAI in the IS literature over time, we assessed the number of articles in annual conferences and journals (see Figure 3). The volume of research has increased over time, with the number of publications reaching 79 articles in 2021. The number of articles published especially from 2019 increased rapidly, with 79% of research appearing between 2019 and 2021. The rapid increase since 2019 is due to the expansion of interest in XAI rather than calling for papers or individual conference. In summary, the number of publications per year shows that the emerging research field of XAI has attracted the attention of AI scientists for the past 3 years.

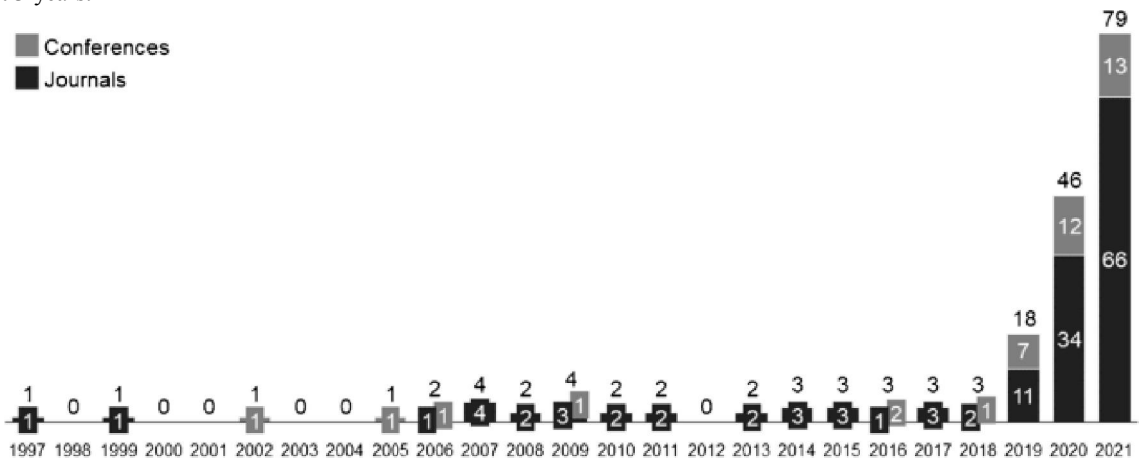
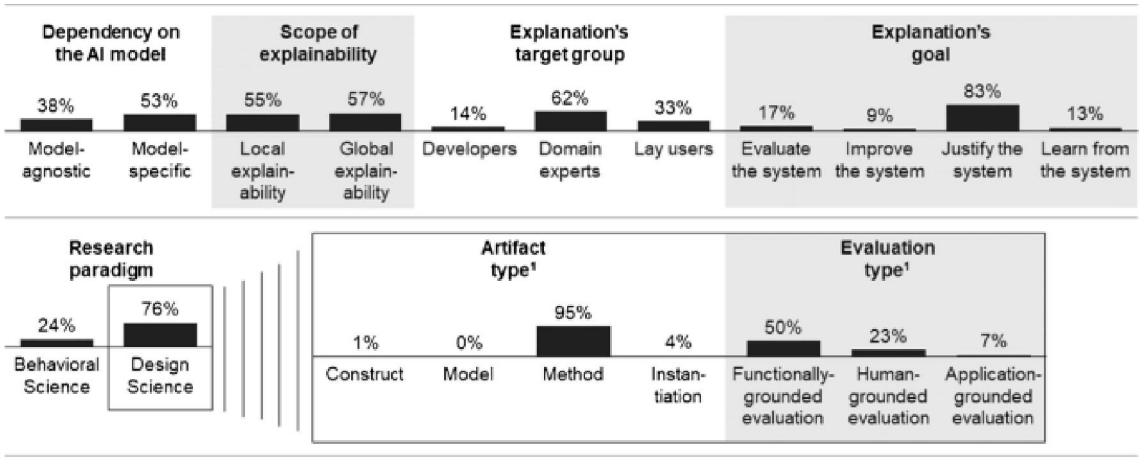


Fig. 3

Characteristics of the academic discussion on XAI in IS literature

In order to study the characteristics of the academic discussion about XAI in the IS literature, we analyzed the dimensions of research studies based on the research agenda, that is, the concept of XAI and the methodological direction (cf. Figure 4). Note that there cannot be multiple answers or responses per category.



1. Percentages refer to the 137 Design Science papers

Fig. 4

Most papers focus on XAI methods that generate annotations for specific AI systems, such as model-specific XAI methods (53%). In contrast, more papers deal with model-agnostic XAI methods that can be used independently of specific AI frameworks (38%). The scope of explanations investigated is diverse: local explanations aimed at rationalizing specific results of AI systems are similar (55%) to global explanations that check the performance of

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

existing AI models. Thirty-three articles (18%) contain a combination of local and global interpretations. First, comments are addressed to domain experts (62%), followed by users (33%). The main purpose of XAI is to justify the decisions of AI systems (83%).

In terms of methodology, R&D research focuses on developing novel XAI (76%). Researchers mainly rely on functionally based assessment scenarios that do not include human involvement (68 articles). Evaluation with users is few, with 31 papers providing human-based evaluation and nine papers providing program-based evaluation. Compared to design-oriented research, behavioral science research is rare (24%).

Analysis of XAI research areas in IS literature

To capture XAI research trends in the IS literature, we identified patterns of similar groups of articles based on conceptual features using cluster analysis. Cluster analysis has been widely used as an analytical tool to classify and group chapters in a specific context (Balijepally et al., 2011; Xiong et al., 2014) and create groups of similar articles (Rissler et al., 2017); Xiong et al., 2014).

In our case, clustering is based on the concept of XAI and the methodological direction of the article (see Figure 4). We coded articles as binary variables and normalized multiple responses per category to account for the same dimensions. We use the well-established agglomerative hierarchical clustering method (Gronau & Moran, 2007), using the Euclidean distance measure as a measure of similarity and the average correlation between group articles within the group. We chose this method because it generates all possible groups rather than predefined groups. We analyzed the average silhouette score to determine the number of groups (Shahapure& Nicholas, 2020). Finally, eight clusters and two outliers with a positive mean silhouette score (0.3) suggest a strong cluster structure with multiple clusters.

Groups corresponding to the eight XAI research directions in the IS literature are described below.

Research Area 1: Revealing the functioning of specific critical black box applications for domain experts

AI systems are increasingly used in critical areas such as healthcare and finance, where transparency is required in decision-making (He et al., 2006; Peñafiel et al., 2020; Pierrard et al., 2021). Transparency is intended to justify the use of AI systems in such critical areas (Pessach et al., 2020). Research area 1, with 47 papers (26%) and among the largest, focuses on ways to detect black box application functionality that is very important to users. For example, the XAI method extracts rules that show the functionality of automatic diagnostic systems to medical professionals (Barakat et al., 2010; Seera & Lim, 2014) or provide key factors for peer-to-peer credit approval in electronic marketplaces. peer-to-peer lending platform (Etang et al., 2021) (Figure 5).

Revealing the functioning of specific critical black box applications for domain experts (26% of total papers)

Dominant XAI concepts		Dominant research paradigm	Predominant application domains
100% Model-specific	**60%** Domain experts	**100%** Design science	• Healthcare • Finance
100% Global explainability	**83%** Justify the system		

Percentages display proportion of articles in Research Area 1 labelled with respective dimensions of the analysis scheme

Fig. 5

In applications where "error costs are high" (Pierrard et al., 2021, p. 2), AI systems can serve as high-performance decision support systems, but the lack of transparency is a problem (eg Areosa &Torgo, 2019). To increase acceptance and adoption, researchers emphasize the need to justify its functionality to users (Areosa & Trgo, 2019). For example, doctors do not need precise assumptions to support their diagnosis, but rather "want to be sure that these assumptions are based on a reasonable basis" (Seera & Lim, 2014, p. 12). Therefore, this line of research focuses on decision support systems that allow users to understand their performance and predict results (Areosa &Torgo, 2019). For this purpose, AI-based decision support systems include interpretation components for disease diagnosis (Barakat et al., 2010; Singh et al., 2019; Stoean&Stoean, 2013), decision making (Pessach et al., 2020), risk assessment credit (for example, Florez-Lopez & Ramon-Jeronimo, 2015; Guo et al., 2021; Sachan et al., 2020) or fraud analysis in telecommunications

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

networks (Irarrázaval et al., 2021). Research in the healthcare industry has found that incorporating the XAI method to diagnose diabetes improves accuracy and medical understanding by doctors (Barakat et al., 2010).

Research Area 1 is only developing XAI methods for electronic markets or evaluating electronic markets. For example, Nascita et al. (2021) developed a new XAI approach to improve the reliability and interpretability of the results of the AI system for traffic classification generated by mobile applications. Grisci et al. (2021) evaluated neural network annotation methods in online shopping databases. They provide a visual interpretation method that determines which features are most important for neural network prediction. Other methods may be used if not expressly designed for electronic markets. Domain experts in e-markets can use global insights to improve supply chain management for B2B sales platforms or e-procurement systems.

The transparency of AI-based decision support systems is achieved through a global explanation that reveals how the AI model works as a whole, rather than explaining specific assumptions (eg, Areosa & Trgo, 2019; Pessach et al., 2020; Zeltner. et al., 2021). In Research Area 1, many approaches have a set of rules that approximate the performance of AI models (eg, Aghaeipoor et al., 2021; Singh et al., 2019). For example, researchers recommend that AI practitioners develop rule annotations in the form of decision trees from AI models to improve the understanding of predictive AI systems (Seera & Lim, 2014). Recently, approaches to deep learning models with deep rules have been implemented (eg, Soares et al., 2021).

In the first paper, Taha and Ghosh (1999) emphasized the need to evaluate the rule extraction approach using fidelity, which is the ability to simulate knowledge embedded in the state AI system. This is the same as the functionally based assessment used in most jobs in Research Area 1 (62%). For example, Soares et al. (2021) implemented a rule extraction approach on multiple databases and provided higher prediction accuracy than state-of-the-art approaches. In particular, only 6% of articles use users to rate comments. For example, Bresso et al. .Irarrazaval et al. (2021) go further and provide a practical assessment. For example, they implemented an interpretable decision support system with a telecommunications provider and claimed that it helped reduce fraud losses. Thirty-four percent of the papers demonstrated the technical capabilities of their methods and how interpretations were made; however, it was not evaluated again.

Accordingly, a more robust assessment including users can pave the way for future research in this area, as recommended by Kim et al. (2020b). Other recurring themes for future research include extending the developed ideas to other applications (Florez-Lopez & Ramon-Jeronimo, 2015; Sevastjanova et al., 2021). Finally, researchers often emphasize that the interpretations that emerge from their approach are only one step toward better understanding the underlying AI system. Therefore, it is important to complement and integrate the existing XAI approaches to help users for a broader understanding (Murray et al., 2021).

Research Area 2: Revealing the functioning of specific black box applications for developers

The smaller research area 2 consists of five papers (3%) and is similar to research area 1 to explore the performance of specific black-box applications. Unlike Research Area 1, which addresses domain experts, Research Area 2 focuses on information for developers. The annotations aim to provide information about the workings of transparent AI models to facilitate the development and implementation of AI systems (Martens et al., 2009) (Figure 6).

Revealing the functioning of specific black box applications for developers (3% of total papers)

Dominant XAI concepts		Dominant research paradigm	Predominant application domains
100% Model-specific	**100%** Developers	**100%** Design science	• Finance • Healthcare
100% Global explainability	**80%** Improve the system		

Percentages display proportion of articles in Research Area 2 labelled with respective dimensions of the analysis scheme

Fig. 6

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

Research area 2 addresses the challenges of the complexity of the growth of AI models for developers: Although the predictions of more complex models are often more accurate, they are poorly understood by those who apply them (Eiras-Franco et al., 2019; Islam et al. ., 2020). Developers need information about how AI models process data and what patterns they reveal to ensure they are accurate and reliable (Eiras-Franco et al., 2019; Islam et al., 2020; Santana et al., 2007). Annotations can capture this information (Jaqulin et al., 2005) and help developers validate models before implementation, thereby improving performance (Martens et al., 2009; Santana et al., 2007).

For this purpose, Research Area 2 created XAI methods that create global annotations and are similar to those in Research Area. For example, Martens et al. (2009) proposed a rule extraction approach that demonstrates the performance of complex support vector machine (SVM) and improves prediction accuracy and clarity. Eiras-Franco et al. (2019) proposed an explanatory method that improves the accuracy and interpretability of predictions when describing the interaction between two entities in a dyadic database. Due to the technical nature of the 2 research papers, the method has not been developed or evaluated for electronic markets to date. However, the XAI approach from this area of research can be a starting point for designing new XAI systems for digital platforms, for example, credit or sales platforms that include AI systems.

It remains to be seen whether annotations effectively help developers as intended. None of the papers in research area 2 include human evaluation. Sixty percent evaluate functionally. For example, Martens et al. (2009) applied the rule extraction approach in several databases and proved that the prediction accuracy improves performance compared to other rule extraction approaches.

The lack of human evaluation is a direct call for future research. In the next step, the researcher should examine the quality and effectiveness of the explanation from the perspective of the developer. Then, according to the technical direction of this research, it is proposed to improve the technical usability of the XAI method, such as computing speed (Eiras-Franco et al., 2019).

V. DISCUSSION AND CONCLUSION

We conducted a systematic and structured review of XAI studies in the IS literature. This section presents opportunities for future research that may provide interesting insights into this yet underexplored area. Finally, we describe the contributions, results, and limitations of our work.

Future research agenda

Our synthesis shows the direction of future research related to XAI research in IS, together with the future research agenda outlined below: (1) refine the understanding of XAI user needs, (2) develop a broad understanding of AI; , (3) implement a more diverse form of XAI evaluation, (4) strengthen the theoretical framework for the role of XAI for human-AI interaction, and (5) improve and improve the use of electronic market needs. Keep in mind that future research directions and future research plans are not exhaustive, and are intended to indicate and indicate potential avenues that may seem promising.

Future Research Direction : Refine the understanding of XAI user needs

XAI research has been criticized for not focusing on user needs, a necessary condition for the effectiveness of interpretation (Herse et al., 2018; Meske et al., 2020). Indeed, there is a gap between the research focus on new algorithms and the desire to create human-consumable explanations, as discussed in many articles in different research fields (eg Liao et al., 2020; Seera & Lim, 2014). Areosa and Trgo (2019) emphasize the need to provide insight into the type of use and information that XAI tools bring to end users. Since one of the main focuses of IS research is user-centered and interactive technology design, IS research will be user-centered and interpretive (Bauer et al., 2021). While six of the eight research areas focus on a wider user group, namely users, domain experts, or developers, only a few basic studies design the XAI approach on specific target users and their needs (eg, medical professionals from various levels). domain knowledge). This shortcoming has been highlighted in studies that call for the design of more user-friendly XAI solutions (Abdul et al., 2018; Miller, 2019). However, so far only a few studies have carried out user-specific design. For example, Barda et al. (2020) proposed an XAI approach that interprets predictions based on risk models of pediatric intensive care unit mortality. This addresses the user-specific interpretation and purpose of

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

information that varies according to their clinical role (eg nurses and doctors). Additional empirical insights point to the need for user-interpretation design, as XAI can only generate human agency and appropriate trust if it considers user needs (Dodge et al., 2018; Elshawi et al., 2019).

We identify several research opportunities to pave the way for a better understanding of the needs of XAI users: First, more empirical research can improve understanding of how different explanations affect the behavior and experiences of different user groups, and how different types of explanations can affect them. this group, for example, doctors (eg, Seera & Lim, 2014). Second, future research may further refine the distinction between developers, domain experts, and users, as user characteristics other than experience may play a key role (e.g., Cui et al., 2019). For example, the user's knowledge base, beliefs, interests, expectations, preferences, and personality can be taken into account (Miller et al., 2017). Third, it is possible to analyze the characteristics of users and the purpose of comments, especially considering that the purpose of comments depends on the context and type of user (Liao et al., 2020). Fourth, future research can be more focused on examining the specific XAI needs of developers, which will benefit from interpretation (cf. Kim et al., 2021), but which has rarely been addressed so far. Research in Research Area 2 (Uncovering Developer-Specific Black Box Application Functionality), the only developer-focused research area, was not evaluated with actual developers in any of the papers.

REFERENCES

[1]. Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., &Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–18). http://dl.acm.org/citation.cfm?doid=3173574.3174156

[2]. Abdul, A., Weth, C. von der, Kankanhalli, M., & Lim, B. Y. (2020). COGAM: Measuring and moderating cognitive load in machine learning model explanations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–14). https://doi.org/10.1145/3313831.3376615

[3]. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[4]. Adam, M., Wessel, M., &Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets, 31*(2), 427–445. https://doi.org/10.1007/s12525-020-00414-7

[5]. Aghaeipoor, F., Javidi, M. M., & Fernandez, A. (2021). IFC-BD: An interpretable fuzzy classifier for boosting explainable artificial intelligence in big data. *IEEE Transactions on Fuzzy Systems.* Advance online publication.https://doi.org/10.1109/TFUZZ.2021.3049911

[6]. Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management, 60*, 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387

[7]. Akter, S., Hossain, M. A., Lu, Q. S., & Shams, S. R. (2021b). Big data-driven strategic orientation in international marketing. *International Marketing Review, 38*(5), 927–947. https://doi.org/10.1108/IMR-11-2020-0256

[8]. Alam, L., & Mueller, S. (2021). Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics and Decision Making, 21*(1), 1–15. https://doi.org/10.1186/s12911-021-01542-6

[9]. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*(1), 1–9. https://doi.org/10.1186/s12911-020-01332-6

[10]. Areosa, I., &Torgo, L. (2019). Visual interpretation of regression error. In P. Moura Oliveira, P. Novais, & L. P. Reis (Eds.), *Lecture notes in computer science. Progress in artificial intelligence* (pp. 473–485). Springer International Publishing. https://doi.org/10.1007/978-3-030-30244-3_39

[11]. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI):

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, January 2024

Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[12]. Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T. & Salovaara, A. (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems, 22*(2). https://aisel.aisnet.org/jais/vol22/iss2/8

[13]. Australian Broadcasting Corporation. (2022). *Robodebt inquiry: Royal commission on unlawful debt scheme begins*. ABC News. https://www.youtube.com/results?search_query=robodebt+royal+commission. Accessed 02 Feb 2023

[14]. Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly*, *45*(1). https://doi.org/10.25300/MISQ/2021/15882

[15]. Balijepally, V., Mangalaraj, G., & Iyengar, K. (2011). Are we wielding this hammer correctly? A reflective review of the application of cluster analysis in information systems research. *Journal of the Association for Information Systems, 12*(5), 375–413. https://doi.org/10.17705/1jais.00266

[16]. Bandara, W., Miskon, S., &Fielt, E. (2011). A systematic, tool-supported method for conducting literature reviews in information systems. *Proceedings of the 19th European Conference on Information Systems (ECIS 2011)* (p. 221). Helsinki, Finland. https://eprints.qut.edu.au/42184/1/42184c.pdf

[17]. Barakat, N. H., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Transactions on Information Technology in Biomedicine, 14*(4), 1114–1120. https://doi.org/10.1109/TITB.2009.2039485

[18]. Barda, A. J., Horvat, C. M., & Hochheiser, H. (2020). A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Medical Informatics and Decision Making, 20*(1), 1–16. https://doi.org/10.1186/s12911-020-01276-x