Brasse, Julia; Broder, Hanna Rebecca; Förster, Maximilian; Klier, Mathias; Sigler, Irina

Explainable artificial intelligence in information systems: A review of the status quo and future research directions

Electronic Markets

Provided in Cooperation with:
Springer Nature

Suggested Citation: Brasse, Julia; Broder, Hanna Rebecca; Förster, Maximilian; Klier, Mathias; Sigler, Irina (2023) : Explainable artificial intelligence in information systems: A review of the status quo and future research directions, Electronic Markets, ISSN 1422-8890, Springer, Berlin, Heidelberg, Vol. 33, Iss. 1, https://doi.org/10.1007/s12525-023-00644-5

This Version is available at:
https://hdl.handle.net/10419/311911

Electronic Markets (2023) 33:26
https://doi.org/10.1007/s12525-023-00644-5

RESEARCH PAPER

Explainable artificial intelligence in information systems: A review of the status quo and future research directions

Julia Brasse[1] · Hanna Rebecca Broder[1] · Maximilian Förster[1] · Mathias Klier[1] · Irina Sigler[1]

Received: 31 July 2022 / Accepted: 30 March 2023 / Published online: 27 May 2023

Abstract

The quest to open black box artificial intelligence (AI) systems evolved into an emerging phenomenon of global interest for academia, business, and society and brought about the rise of the research field of explainable artificial intelligence (XAI). With its pluralistic view, information systems (IS) research is predestined to contribute to this emerging field; thus, it is not surprising that the number of publications on XAI has been rising significantly in IS research. This paper aims to provide a comprehensive overview of XAI research in IS in general and electronic markets in particular using a structured literature review. Based on a literature search resulting in 180 research papers, this work provides an overview of the most receptive outlets, the development of the academic discussion, and the most relevant underlying concepts and methodologies. Furthermore, eight research areas with varying maturity in electronic markets are carved out. Finally, directions for a research agenda of XAI in IS are presented.

Keywords Explainable artificial intelligence · Explainable machine learning · Comprehensible artificial intelligence · Comprehensible machine learning · Literature review

JEL Classification M10

Introduction

Artificial intelligence (AI) is already ubiquitous at work and in everyday life: in the form of diverse technologies, such as natural language processing or image recognition (Abdul et al., 2018; Berente et al., 2021) and in various application domains, including electronic markets, finance, healthcare, human resources, public administration, and transport (Collins et al., 2021; Meske et al., 2020). The presence of AI will expand as about 70% of companies worldwide intend to adopt AI by 2030 (Bughin et al., 2018). Thereby, AI is expected to transform all aspects of society (Collins et al., 2021; Makridakis, 2017).

The current CEO of Alphabet Inc. anticipates AI to "have a more profound impact on humanity than fire, electricity and the internet" (Knowles, 2021). AI holds great potential through tremendous efficiency gains and novel information processing capabilities (Asatiani et al., 2021) and even surpasses human performance in specific tasks (Meske et al., 2022). For instance, AI has outperformed physicians in diagnosing breast cancer (e.g., McKinney et al., 2020). At the same time, the use of AI is associated with severe risks, particularly concerning managerial issues such as inscrutability, ethical issues including 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). Potential negative consequences of AI usage affect not only individuals and organizations, but society as a whole (Mirbabaie et al., 2022; Robert et al., 2020). For example,

Responsible Editor: Shahriar Akter

✉ Mathias Klier
mathias.klier@uni-ulm.de

Julia Brasse
julia.brasse@uni-ulm.de

Hanna Rebecca Broder
hanna.broder@uni-ulm.de

Maximilian Förster
maximilian.foerster@uni-ulm.de

Irina Sigler
irina.sigler@uni-ulm.de

[1] Institute of Business Analytics, University of Ulm, Helmholtzstraße 22, 89081 Ulm, Germany

an AI-based debt recovery program called "Robodebt" scheme unlawfully claimed almost $2 billion from more than 400,000 Australian citizens (Australian Broadcasting Corporation, 2022). There are growing concerns that using AI could exacerbate social or economic inequalities (Gianfrancesco et al., 2018). Examples include an AI-based recruiting engine used by Amazon.com Inc. which downgraded resumes from female in favor of male candidates (Gonzalez, 2018), an AI operated by Twitter Inc. to communicate with users who became verbally abusive, and an AI used by Google LLC which returned racist results in image searches (Yampolskiy, 2019).

The advancing capabilities of AI models contribute to their opacity, rendering their functioning and results uninterpretable to humans (Berente et al., 2021). Opacity can, one the one hand, lead to humans blindly relying on AI results and substituting their own judgment with potentially false decisions (Robert et al., 2020). On the other hand, the lack of interpretability may lead to reluctance to use AI. In the case of breast cancer diagnosis, AI-based decision support systems may fail to detect certain diseases, for instance, due to biased training data. Physicians exhibiting overreliance may fail to detect these errors; physicians that do not trust AI systems and refuse to use them may not benefit from the decision support.

Explainable AI (XAI) aims at both leveraging the potential and mitigating the risks of AI by increasing its explainability. XAI aims to empower human stakeholders to comprehend, appropriately trust, and effectively manage AI (Arrieta et al., 2020; Langer et al., 2021). In the example of breast cancer diagnosis, explainability can assist physicians in understanding the functioning and results of an AI-based decision support system. Thus, it may help them appropriately trust the system's decisions and detect its errors. Consequently, a partnership between physicians and AI might make better decisions than either physicians or AI individually. Efforts to increase the explainability of AI systems are emerging across various sectors of society. Companies strive to make their AI systems more comprehensible (e.g., Google, 2022; IBM, 2022). Regulators take action to demand accountability and transparency of AI-based decision processes. For instance, the European General Data Protection Regulation (GDPR) guarantees the "right to explanation" for those affected by algorithmic decisions (Selbst & Powles, 2017). The upcoming EU AI regulation requires human oversight—to interpret and contest AI systems' outcomes—in "high-risk" applications such as recruiting or creditworthiness evaluation (European Commission, 2021). XAI's economic and societal relevance attracts researchers' attention, which manifests in an increasing number of publications in recent years (Arrieta et al., 2020). For instance, XAI researchers work on revealing the functioning of specific AI-based applications, such as cancer diagnosis systems (Kumar et al., 2021) and malware prediction systems (Iadarola et al., 2021), to their users. Further, they investigate approaches to automatically generate explanations along AI decisions that can be applied independently from the underlying AI model. Exemplary use cases include credit risk assessment (Bastos & Matos, 2021) or fraud detection (Hardt et al., 2021). Information systems (IS) research is predestined to investigate and design AI explainability, as it views technology from individuals', organizations', and society's perspectives (Bauer et al., 2021).

Especially for an emerging research field such as XAI, a literature review can help to create "a firm foundation for advancing knowledge" (Webster & Watson, 2002, p. 13) and put forward the research's relevance and rigor (vom Brocke et al., 2009). We aim to provide deeper insights into this body of knowledge by conducting a structured literature review. The contribution is twofold: First, we provide a structured and comprehensive literature review of XAI research in IS. Second, we provide a future research agenda for XAI research in IS.

Our paper is structured as follows: In the following, we provide an overview of related work and outline our research questions. In the third section, we present the methodology, followed by the results in the fourth section. Finally, we carve out a future research agenda and present the contribution, implications, and limitations.

Theoretical background and related work

Theoretical foundations

Given that IS research investigates and shapes "how individuals, groups, organizations, and markets interact with IT" (Sidorova et al., 2008, p. 475), human-AI interaction is a crucial research topic for the discipline. In general, human-agent interaction occurs between an IT system and a user seeking to conduct a specific task in a given context (Rzepka & Berger, 2018). It is determined by the characteristics of the task, the context, the user, and the IT system (Rzepka & Berger, 2018). When the human counterpart is an AI system, specific characteristics of AI systems must be considered. Modern AI systems with continually evolving frontiers of emerging computing capabilities provide greater autonomy, more profound learning capacity, and higher inscrutability than previously studied IT systems (Baird & Maruping, 2021; Jiang et al., 2022). The rapid progress in AI is primarily contributed to the rise of machine learning (ML), which can be defined as the ability to learn specific tasks by constructing models based on processing data (Russell & Norvig, 2021). The autonomy and learning capacity of ML-based AI systems further reinforce inscrutability (Berente et al., 2021). Thus,

challenges arise to manage human-AI interaction with ever-increasing levels of AI autonomy, learning capacity, and inscrutability.

From a managerial perspective, inscrutability carries four interdependent emphases: opacity, transparency, explainability, and interpretability (Berente et al., 2021). First, opacity is a property of the AI system and refers to its complex nature, which impedes humans from understanding AI's underlying reasoning processes (Meske et al., 2020). Many AI systems are "black boxes," which means that the reasons for their outcomes remain obscure to humans—often not only to the users but also to the developers (Guidotti et al., 2019; Merry et al., 2021). A prominent example are neural networks. Second, transparency refers to the willingness to disclose (parts of) the AI system by the owners and is thus considered a strategic management issue (Granados et al., 2010). Third, explainability is a property of the AI system and refers to the system's ability to be understood by at least some parties, at least to a certain extent (Gregor & Benbasat, 1999). Finally, interpretability refers to the understandability of an AI system from human perspectives. An AI system with a certain degree of explainability might be adequately interpretable for one person but not necessarily for another (Berente et al., 2021). For instance, decision trees can become uninterpretable for some users as complexity increases (Mittelstadt et al., 2019).

Opacity significantly affects human-AI interaction: It prevents humans from scrutinizing or learning from an AI system's decision-making process (Arrieta et al., 2020). Confronted with an opaque system, humans cannot build appropriate trust; they often either blindly follow the system's decisions and recommendations or do not use the system (Herse et al., 2018; Rader & Gray, 2015). Thus, opacity constitutes an impediment to both human agency and AI adoption. The research field of XAI addresses the opacity of AI systems. XAI aims at approaches that make AI systems more explainable—sometimes also referred to as comprehensible (Doran et al., 2018)—by automatically generating explanations for their functioning and outcomes while maintaining the AI's high performance levels (Adadi & Berrada, 2018; Gregor & Benbasat, 1999). In day-to-day human interaction, "explanation is a social and iterative process between an explainer and an explainee" (Chromik & Butz, 2021, p. 1). This translates into the context of human-AI interaction, where explanations constitute human-understandable lines of reasoning for why an AI system connects a given input to a specific output (Abdul et al., 2018). Thus, explanations can address the opacity of AI systems and increase their interpretability from users' perspectives. Researchers emphasize that clarifying XAI's role can make significant contributions to the ongoing discussion of human-AI interaction (Sundar, 2020).

Terminological foundations

The XAI research discipline is driven by four key goals (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., 2019): First, to generate explanations that allow to *evaluate* an AI system and thus detect its flaws and prevent unwanted behavior (Adadi & Berrada, 2018; Gerlings et al., 2021; Meske et al., 2020; Wang et al., 2019). For instance, evaluation in this context is utilized to detect and prevent non-equitable treatment of marginalized communities (Arrieta et al., 2020). The second goal is to build explanations that help to *improve* an AI system. In this case, explanations can be used by developers to improve a model's accuracy by deepening their 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). Third, to provide explanations that *justify* an AI system's decisions by improving transparency and accountability (Adadi & Berrada, 2018; Gerlings et al., 2021; Meske et al., 2020; Wang et al., 2019). One prominent example highlighting the need to justify is based on the "right to explanation" for those affected by algorithmic decisions (cf., e.g., GDPR); another example concerns decisions made by a professional who follows an AI system's recommendation but remains accountable for the decision (Arrieta et al., 2020). Finally, to produce explanations that allow to *learn* from the system by unmasking unknown correlations that could indicate causal relationships in the underlying data (Adadi & Berrada, 2018; Langer et al., 2021; Meske et al., 2020). In a nutshell, XAI aims to evaluate, improve, justify, and learn from AI systems by building explanations for a system's functioning or its predictions (Abdul et al., 2018; DARPA, 2018).

To reach these goals, XAI research provides a wide array of approaches that can be grouped along two dimensions: scope of explainability and model dependency (Adadi & Berrada, 2018; Arrieta et al., 2020; Vilone & Longo, 2020). The scope of explainability can be global or local (Adadi & Berrada, 2018; Arrieta et al., 2020; Heuillet et al., 2021; Payrovnaziri et al., 2020; Vilone & Longo, 2020). A *global* explanation targets the functioning of the entire AI model. Using the example of credit line decisions, a global explanation might highlight the most relevant criteria that are exploited by the AI model to derive credit line decisions. *Local* explanations, on the other hand, focus on rationalizing an AI model's specific outcome. Returning to the example of credit line decisions, a local explanation might provide the most essential criteria for an individual denial or approval. The second dimension, dependency on the AI model, 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 specific AI models or model classes (Arrieta et al., 2020; Rawal et al., 2021), like neural networks (Montavon et al., 2018), as they consider internal components of the AI model (class), such as structural information. In turn, *model-agnostic* approaches disregard the models' 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).

Designing or choosing the best XAI approach for a given problem is equivalent to solving a "human-agent interaction problem" (Miller, 2019, p. 5). Thus, it is vital to consider an explanation's audience. Three major target groups are the focus of XAI research (Bertrand et al., 2022; Cooper, 2004; Ribera & Lapedriza, 2019; Wang et al., 2019). The first group comprises *developers* who build AI systems, i.e., data scientists, computer engineers, and researchers (Bertrand et al., 2022; Ribera & Lapedriza, 2019; Wang et al., 2019). To illustrate, using the example of credit line decisions, this is the team building the AI system or responsible for maintaining it. The second group contains *domain experts* who share expertise based on formal education or professional experience in the application field (Bertrand et al., 2022; Ribera & Lapedriza, 2019; Wang et al., 2019). In the case of credit line decisions, this would be the bank advisor accountable for the credit line decision. The final group, *lay users*, includes individuals who are affected by AI decisions (Bertrand et al., 2022; Cooper, 2004; Ribera & Lapedriza, 2019), e.g., the bank customer who was approved or denied a credit line based on an AI system's recommendation (Mittelstadt et al., 2019). Additionally, this third group includes lay users that interact with an AI, e.g., customers who explore credit line options with the help of an AI-based agent.

To investigate to what extent XAI approaches solve this "human-agent interaction problem," literature established a baseline of three different evaluation scenarios (Adadi & Berrada, 2018; Chromik & Schuessler, 2020; Doshi-Velez & Kim, 2018). *Functionally grounded evaluation*, as the first scenario, is employed to assess the technical feasibility of XAI approaches and explanations' characteristics employing proxy measures (Doshi-Velez & Kim, 2018), e.g., analyze an explanation's length to assess its complexity (Martens & Provost, 2014; Wachter et al., 2018). While functionally grounded evaluation omits user involvement, both the second and the third scenarios build on studies with humans (Doshi-Velez & Kim, 2018). The second scenario, *human-grounded evaluation*, aims to assess the quality of explanations by conducting studies with human subjects who are not necessarily the target users, e.g., students, performing simplified proxy-tasks (Doshi-Velez & Kim, 2018; Förster et al., 2020a). *Application-grounded evaluation*, as the third scenario, is based on real-world testing involving the intended users of an AI system and deployment in the actual application setting (Abdul et al., 2018). Reverting to the example of the credit line decisions, an application-grounded evaluation would be set in an actual bank environment, with actual bank advisors and/or customers as subjects, while human-grounded evaluation would allow for a simulated environment. Table 1 provides an overview of key concepts and definitions in XAI research, which we will draw on when analyzing the identified body of literature for providing a comprehensive literature review of XAI research in IS.

Existing literature reviews on XAI

Several literature reviews address the growing body of research in the field of XAI applying different foci and angles. While some of them aim at formalizing XAI (e.g., Adadi & Berrada, 2018), for example, by drawing together the body of knowledge on the nature and use of explanations from intelligent systems (Gregor & Benbasat, 1999), others provide taxonomies for XAI in decision support (Nunes and Jannach, 2017) or survey methods for explaining AI (e.g., Guidotti et al., 2019). Other literature reviews focus on specific (X)AI methods, such as rule-based models (e.g., Kliegr et al., 2021), neuro-fuzzy rule generation algorithms (e.g., Mitra & Hayashi, 2000), or neural networks (e.g., Heuillet et al., 2021), or review-specific explanation formats, like visual explanations (e.g., Zhang & Zhu, 2018). Another stream of literature reviews highlights user needs in XAI, for example, by reviewing design principles for user-friendly explanations (Chromik & Butz, 2021) or XAI user experience approaches (Ferreira & Monteiro, 2020).

Another group of literature reviews on XAI focuses on specific application domains like healthcare (e.g., Amann et al., 2020; Chakrobartty & El-Gayar, 2021; Payrovnaziri et al., 2020; Tjoa & Guan, 2021), finance (e.g., Kute et al., 2021; Moscato et al., 2021), or transportation (e.g., Omeiza et al., 2021). For example, Amann et al. (2020) provide a comprehensive review of the role of AI explainability in clinical practice to derive an evaluation of what explainability means for the adoption of AI-based tools in medicine. Omeiza et al. (2021) survey XAI methods in autonomous driving and provide a conceptual framework for autonomous vehicle explainability. Other scholars apply XAI to adjacent disciplines (e.g., Abdul et al., 2018; Miller, 2019). For instance, in an often-cited paper, Miller (2019) argues that XAI research can build on insights from the social sciences. The author reviews papers from philosophy and psychology which study how people define, generate, select, evaluate, and present explanations and which cognitive biases and social norms play a role. Thereby, most literature reviews describe existing research gaps and point toward future research directions focusing on their specific view.

As outlined above, existing literature reviews cover various aspects of XAI research. However, to our best knowledge, none of them has provided a comprehensive literature review on XAI research in IS. Our literature review aims at addressing this gap.

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

Research questions

While considerable progress in XAI has already been made by computer scientists (Arrieta et al., 2020), interest in this field has increased rapidly among IS scholars in recent years (Meske et al., 2020). This is underpinned, for instance, by an increasing number of Calls for Papers (cf., e.g., Special Issue on Explainable and Responsible Artificial Intelligence in *Electronic Markets*, Special Issue on Designing and Managing Human-AI Interactions in *Information Systems Frontiers*), conference tracks (cf., e.g., Minitrack on Explainable Artificial Intelligence at *Hawaii International Conference on System Sciences*), and Editorials (cf., e.g., Editorial "Expl(AI)n It to Me – Explainable AI and Information Systems Research" in *Business & Information Systems Engineering*). In their Editorial, Bauer et al. (2021) emphasize that IS research is predestined to focus on XAI given the versatility of requirements and consequences of explainability from individuals' and society's perspectives. Moreover, in a research note summarizing existing IS journal articles, Meske et al. (2020) call for a resurgence of research on explainability in IS—after explanations for relatively transparent expert systems have been intensively investigated. To the best of our knowledge, no work exists synthesizing XAI research in IS based on a structured and comprehensive literature search.

To provide deeper insights into the research field of XAI in the IS community, we conduct a structured and comprehensive literature review. Our literature review addresses the following research questions (RQ):

RQ1: How can the academic discussion on XAI in the IS literature be characterized?
RQ2: Which are potential future XAI research areas in IS?

To address the first research question, we aim to (i) identify IS publication outlets that are receptive to XAI research, (ii) describe how the academic discussion on XAI in the IS literature developed over time, (iii) analyze the underlying concepts and methodological orientations of the academic discussion on XAI in the IS literature, and (iv) present the most critical XAI research areas in IS literature. To address the second research question, we aim to derive directions for a research agenda of XAI in IS.

Literature review approach

Relying on the previous discussions, we investigate how IS scholars conduct XAI research. We aim at not only summarizing but analyzing and critically examining the status quo 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 apply a comprehensive and replicable literature search strategy, which includes relevant journals and conferences, appropriate keywords, and an adequate time frame (vom Brocke et al., 2009). Bandara et al. (2011) propose two main steps: selecting the relevant sources to be searched (cf. Webster & Watson, 2002) and defining the search strategy in terms of time frame, search terms, and search fields (Cooper, 1988; Levy & Ellis, 2006). In order to systematically analyze the papers according to XAI theory and IS methodology, we added a third step and coded the articles with respect to relevant concepts in the literature (Beese et al., 2019; Jiang & Cameron, 2020).

Source selection

The literature search needs to include the field's leading journals known for their high quality and will thus publish the most relevant research contributions (Webster & Watson, 2002). The renowned Association for Information Systems (AIS), with members from approximately 100 countries, publishes the Senior Scholars' Basket of Journals, as well as the Special Interest Groups (SIG) Recommended Journals. In our search, we included the eight journals in the AIS Senior Scholars' Basket of Journals, and the 64 AIS SIG Recommended Journals. Because of their high quality, we considered all remaining journals in the AIS eLibrary (including Affiliated and Chapter Journals). In order to identify high-quality journals, different rankings are helpful (Akter et al., 2021b; Levy & Ellis, 2006; vom Brocke et al., 2009). We explicitly considered journals from three prominent rankings: First, journals from the Chartered Association of Business Schools (ABS)/Academic Journal Guide (AJG) 2021 (ranking tier 3/4/4* benchmark, category "Information Management"). Second, journals from the Journal Quality List of the Australian Business Deans Council (ABDC) (ranking tier A/A* benchmark, category "Information Systems"). Third, journals from the German Academic Association of Business Research VHB-JOURQUAL3 (ranking tier A + /A/B benchmark, category "Information Systems").

Moreover, it is recommended to include high-quality conference proceedings (Webster & Watson, 2002), especially when analyzing a relatively nascent and emerging research field such as XAI. Conferences are a venue for idea generation and support the development of new research agendas (Levy & Ellis, 2006; Probst et al., 2013). Thus, we included the major international IS conferences. More precisely, we considered the proceedings of the four AIS Conferences and the proceedings of the twelve AIS Affiliated Conferences. In addition, we ensured that all conferences from the VHB-JOURQUAL3 (ranking tier A + /A/B benchmark, category "Information Systems") are included.

This resulted in 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, 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.

Our literature search, which was performed in January 2022, resulted in 1724 papers. Papers were screened based on titles and abstracts, with researchers reading the full text where necessary. We excluded all papers that did not deal with XAI as defined above. More specifically, we excluded all papers that focus entirely on AI without the notion of explanations. For instance, we excluded papers on how humans can explain AI for other humans. Further, we excluded papers focusing on the explainability of "Good Old Fashioned AI" such as expert or rule-based systems (Meske et al., 2020, p. 6). In contrast to our understanding of AI, as defined in the introduction, this broader definition of AI also includes inherently interpretable systems, such as knowledge-based or expert systems, which do not face the same challenges of lacking transparency.

To determine our data set of relevant papers, three researchers coded independently from each other and discussed coding disagreements to reach consent. At least two researchers analyzed each paper. Interrater reliability measured by Cohen's Kappa was 0.82—"almost perfect agreement" (Landis & Koch, 1977, p. 165). This procedure led to a set of 154 papers, which then served as the basis for a backward (resulting in 32 papers) and forward search (resulting in 28 papers), as suggested by Webster and Watson (2002).

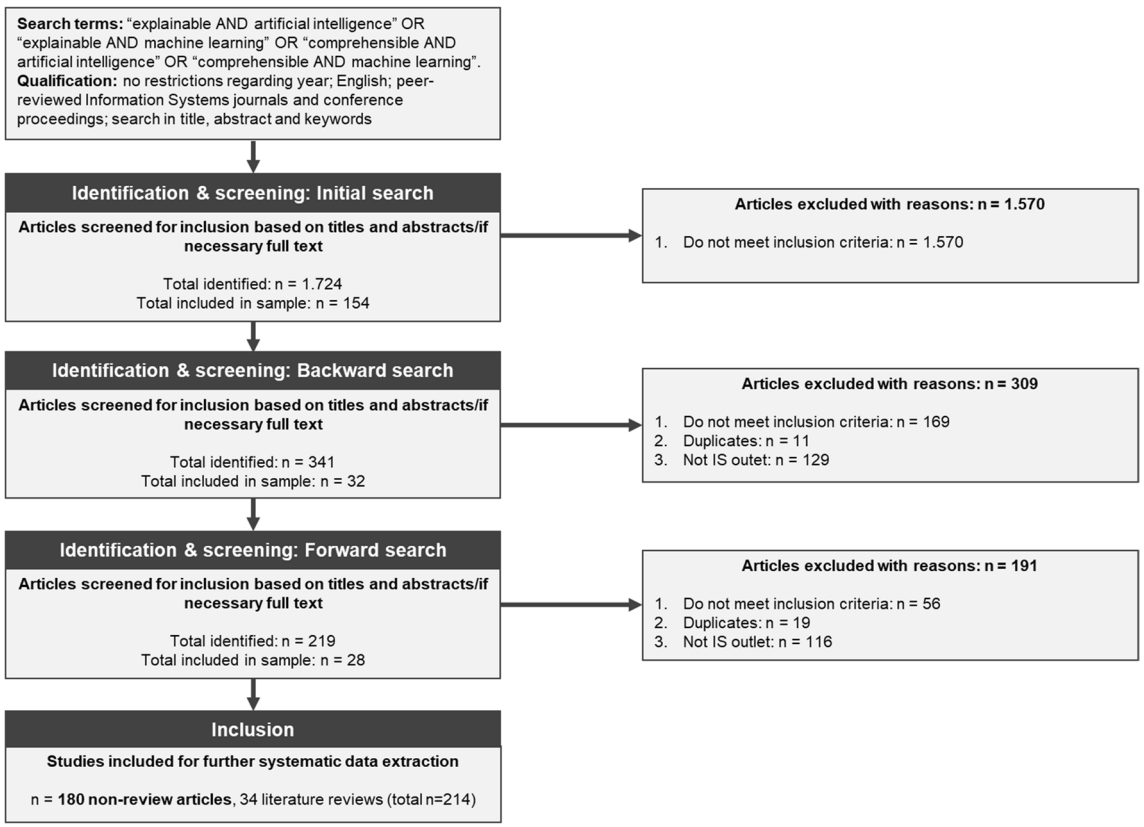


Fig. 1 Search strategy and screening process

We reached a final set of 214 papers that served as the basis for our subsequent analyses.

Analysis scheme and coding procedure

Our goal is to not only summarize but analyze and critically examine the status quo of XAI research in IS (Beese et al., 2019; Rowe, 2014). In order to do so, we first analyzed all 34 papers that solely provide an overview of current knowledge, i.e., literature reviews. We then coded the 180 remaining articles using an analysis scheme derived from existing literature (cf. Section "Terminological foundations"). More specifically, in our analysis, we differentiate relevant theoretical concepts in XAI research and central methodological concepts of IS research. Regarding relevant concepts of XAI literature, we distinguish an XAI approach's dependency on the AI model (Adadi & Berrada, 2018; Arrieta et al., 2020) and its scope of explainability (Adadi & Berrada, 2018; Arrieta et al., 2020; Payrovnaziri et al., 2020; Vilone & Longo, 2020) as well as explanation's target group (Ribera & Lapedriza, 2019; Wang et al., 2019) and goal (Meske et al., 2020). Regarding IS methodology, we distinguish the prevalent research paradigms, i.e., Design Science and Behavioral Science (Hevner et al., 2004). For Design Science contributions, we further specify the artifact type according to Hevner et al. (2004) and the evaluation type according to established evaluation scenarios for XAI approaches (Adadi & Berrada, 2018; Chromik & Schuessler, 2020; Doshi-Velez & Kim, 2018). This results in the following analysis scheme (Fig. 2):

Three researchers coded the 180 remaining articles according to the analysis scheme. Multiple labels per dimension were possible. For a subset of 100 articles, each article was coded by at least two researchers. Interrater reliability measured by Cohen's Kappa was 0.74, which is associated with "substantial agreement" (Landis & Koch, 1977, p. 165). In case of disagreement, the researchers reached a consensus based on discussion.

Results

This section is dedicated to our results. First, we analyze receptive IS publication outlets to XAI research. Second, we examine the development of the academic discussion on XAI in IS literature over time. Third, we analyze the academic discussion's underlying concepts and methodological orientation. Finally, we derive major XAI research areas.

Receptive IS outlets to XAI research

We analyzed which journals and conferences are receptive to XAI research. The results are helpful in three ways: they provide researchers and practitioners with potential outlets

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 Analysis scheme

where they can find related research, they assist researchers in identifying target outlets, and they offer insights for editors to what extent their outlet is actively involved in the academic discussion on the topic (Bandara et al., 2011). One hundred forty-one articles were published in journals, and 39 articles in conference proceedings. An overview of the number of publications per journal and per conference is included in the Appendix.

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

To examine the development of the academic discussion on XAI in IS literature over time, we evaluated the number of articles in conferences and journals per year (cf. Fig. 3). The amount of research increased over time, with the number of publications rising to 79 articles in 2021. Especially from 2019 onward, the number of published articles increased rapidly, with 79% of the studies appearing between 2019 and 2021. The rapid increase since 2019 is not attributed to particular calls for papers or individual conferences but due to a widely growing interest in XAI. In sum, the number of publications per year indicates that the nascent research field of XAI has been gaining significant attention from IS scholars in the last 3 years.

Characteristics of the academic discussion on XAI in IS literature

To examine the characteristics of the academic discussion on XAI in IS literature, we analyzed the dimensions of the research papers according to our analysis scheme, i.e., underlying XAI concepts and methodological orientation (cf. Fig. 4). Note that multiple answers or no answers per category were possible.

Most papers conceptually focus on XAI methods that generate explanations for specific AI systems, i.e., model-specific XAI methods (53%). In contrast, fewer papers deal with model-agnostic XAI methods, which can be used independently of the specific AI system (38%). The scope of explainability under investigation varies: Local explanations that focus on rationalizing an AI system's specific outcome are represented almost equally (55%) to global explanations that examine the functioning of the underlying AI model (57%). Thirty-three articles (18%) feature a combination of local and global explanations. First and foremost, explanations address domain experts (62%), followed by lay users (33%). The predominant goal of XAI is to justify an AI system's decisions (83%).

Regarding methodological orientation, IS research efforts concentrate on developing novel XAI artifacts (76%). Researchers mainly rely on the functionally grounded evaluation scenario (68 articles), which omits human involvement. Evaluation with users is relatively scarce, with 31 articles conducting human-grounded and nine papers performing an application-grounded evaluation. Compared to design-oriented research, behavioral science studies are rare (24%).

Analysis of XAI research areas in IS literature

To derive XAI research areas in IS literature, we identify patterns of homogenous groups of articles according to conceptual characteristics using cluster analysis. Cluster analysis is widely used in IS research as an analytical tool to classify and disentangle units in a specific context (Balijepally et al., 2011; Xiong et al., 2014) and to form homogenous groups of articles (Rissler et al., 2017; Xiong et al., 2014).

In our case, clustering is based on underlying XAI concepts and the methodological orientation of articles (cf. Fig. 4). To consider dimensions equally, we encoded articles

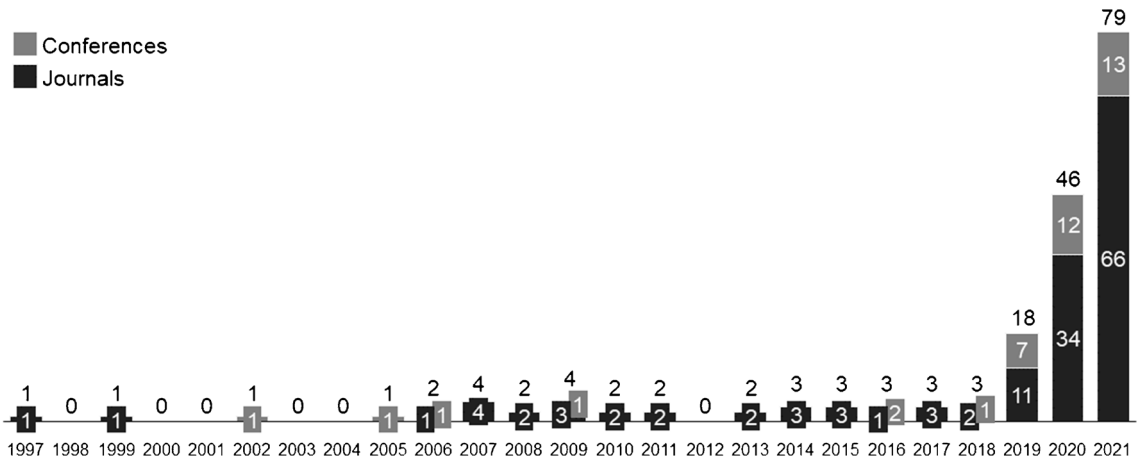
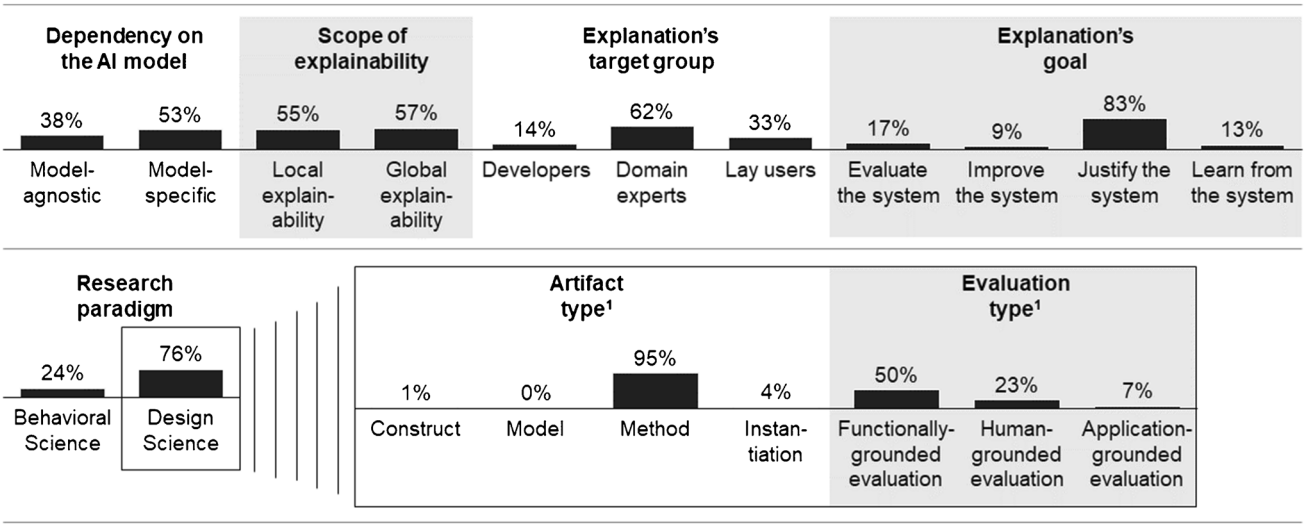


Fig. 3 Number of articles by year



1. Percentages refer to the 137 Design Science papers

Fig. 4 Characteristics of the academic discussion according to dimensions of the analysis scheme

as binary variables and normalized multiple answers per category. We applied the well-established agglomerative hierarchical clustering method using Euclidean distance measure as the similarity criterion and average linkage to group articles in clusters (Gronau & Moran, 2007). We chose this method as it does not form a predefined number of clusters but all possible clusters. To determine a reasonable number of clusters, we analyzed average silhouette scores (Shahapure & Nicholas, 2020). It resulted in eight clusters and two outliers with a positive average silhouette score (0.3), suggesting a solid clustering structure with an interpretable number of clusters.

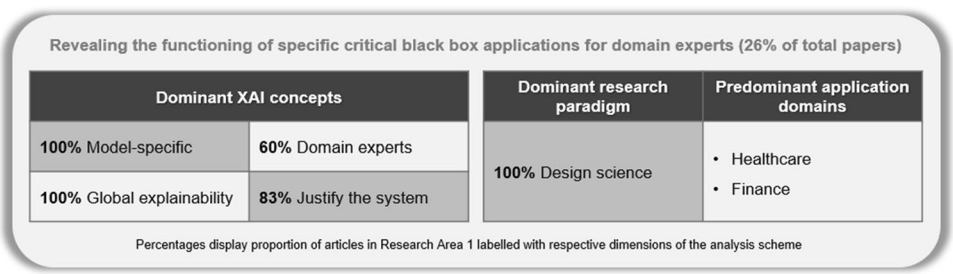
The clusters correspond to eight XAI research areas in IS literature, described in the following.

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

AI systems are increasingly applied in critical areas such as healthcare and finance, where there is a need for transparency in decision-making (He et al., 2006; Peñafiel et al., 2020; Pierrard et al., 2021). Transparency is meant to justify the usage of AI systems in such critical areas (Pessach et al., 2020). Research Area 1, which is among the largest with 47 papers (26%), aims at methods to reveal the functioning of specific critical black box applications to their users. For instance, XAI methods extract rules that reveal the functioning of an automatic diagnosis system to medical experts (Barakat et al., 2010; Seera & Lim, 2014) or, in the context of electronic markets, showcase central factors for loan approval on peer-to-peer lending platforms (Yang et al., 2021) (Fig. 5).

In critical application domains "where the cost of making a mistake is high" (Pierrard et al., 2021, p. 2), AI systems have the potential to serve as high-performant decision support systems—however, their lack of transparency constitutes a problem (e.g., Areosa & Torgo, 2019). To increase acceptance and adoption, researchers stress the need to justify their functioning to their users (Areosa & Torgo, 2019). For instance, medical practitioners not only need accurate predictions supporting their diagnosis but "would like to be convinced that the prediction is based

Fig. 5 Overview Research Area 1



on reasonable justifications" (Seera & Lim, 2014, p. 12). Thus, this research area aims at decision support systems that allow users to understand their functioning and predictive performance (Areosa & Torgo, 2019). To this end, explainable components are added to AI-based decision support systems for, e.g., diagnosis of diseases (Barakat et al., 2010; Singh et al., 2019; Stoean & Stoean, 2013), hiring decisions (Pessach et al., 2020), credit risk assessment (e.g., Florez-Lopez & Ramon-Jeronimo, 2015; Guo et al., 2021; Sachan et al., 2020), or fraud analysis in telecommunication networks (Irarrázaval et al., 2021). Studies in the healthcare domain identify that adding XAI methods for diagnosing diabetes increases medical accuracy and intelligibility by clinical practitioners (Barakat et al., 2010).

In Research Area 1, only very few articles develop XAI methods specifically for electronic markets or evaluate them in electronic markets. For instance, Nascita et al. (2021) develop a novel XAI approach for classifying traffic generated by mobile applications increasing the trustworthiness and interpretability of the AI system's outcomes. Grisci et al. (2021) evaluate their method for explaining neural networks on an online shopping dataset. They present a visual interpretation method that identifies which features are the most important for a neural network's prediction. While not explicitly designed for electronic markets, other methods might be transferable. Domain experts in electronic markets might benefit from global explanations, for instance, to improve supply chain management for B2B sales platforms or electronic purchasing systems.

Transparency of AI-based decision support systems is achieved by global explanations, which are supposed to reveal the functioning of the AI model as a whole rather than explain particular predictions (e.g., Areosa & Torgo, 2019; Pessach et al., 2020; Zeltner et al., 2021). Many approaches in Research Area 1 acquire a set of rules that approximate the functioning of an AI model (e.g., Aghaeipoor et al., 2021; Singh et al., 2019). For instance, researchers propose to produce explanatory rules in the form of decision trees from AI models to enable domain users such as medical practitioners to comprehend an AI system's prediction (Seera & Lim, 2014). More recently, approaches to approximate deep learning models with fuzzy rules have been pursued (e.g., Soares et al., 2021).

In an early paper, Taha and Ghosh (1999) emphasize the need to evaluate rule extraction approaches using fidelity, i.e., the capability to mimic the embedded knowledge in the underlying AI system. This is equivalent to functionally grounded evaluation, which is applied in many papers in Research Area 1 (62%). For instance, Soares et al. (2021) implement their rule extraction approach on several datasets and prove that it yields higher predictive accuracy than state-of-the-art approaches. Notably, only 6% of articles use users to evaluate explanations. For instance, Bresso et al. (2021) ask three pharmacology experts to evaluate whether extracted rules are explanatory for the AI system's outcomes, i.e., prognoses of adverse drug reactions. Irarrázaval et al. (2021) go further and perform an application-grounded evaluation. In a case study, they implement their explainable decision support system with a telecommunication provider and confirm that it helps reduce fraud losses. Thirty-four percent of papers demonstrate the technical feasibility of their methods and present how resulting explanations look like; however, they are not further evaluated.

Accordingly, a more robust evaluation, including users, may pave the way for future research in this research area, as suggested by Kim et al., (2020b). Other recurring themes of future research include the expansion of the developed ideas to other applications (Florez-Lopez & Ramon-Jeronimo, 2015; Sevastjanova et al., 2021). Finally, researchers often stress that explanations resulting from their approach are only one step toward a better understanding of the underlying AI system. Thus, it is essential to supplement and combine existing XAI approaches to help users gain a more comprehensive understanding (Murray et al., 2021).

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

The relatively small Research Area 2 consists of five papers (3%) and develops—similar to Research Area 1—methods to reveal the functioning of specific black box applications. Contrary to Research Area 1, which addresses domain experts, Research Area 2 focuses on explanations for developers. Explanations aim to provide insights into the functioning of opaque AI models to facilitate the development and implementation of AI systems (Martens et al., 2009) (Fig. 6).

Research Area 2 tackles the challenges of the growing complexity of AI models for developers: While predictions of more complex models often become more accurate, they also become less well understood by those implementing them (Eiras-Franco et al., 2019; Islam et al., 2020). Developers need information on how AI models process data and which patterns they discover to ensure that they are accurate and trustworthy (Eiras-Franco et al., 2019; Islam et al., 2020; Santana et al., 2007). Explanations can extract this information (Jakulin et al., 2005) and assist developers in validating a model before implementation, thereby improving its performance (Martens et al., 2009; Santana et al., 2007).

To this end, Research Area 2 develops model-specific XAI methods that generate global explanations and resemble those in Research Area 1. To illustrate, Martens et al. (2009) propose an approach to extract rules that represent the functioning of complex support vector machines (SVMs) and increase performance in predictive accuracy and comprehensibility. Eiras-Franco et al. (2019) propose an explainable

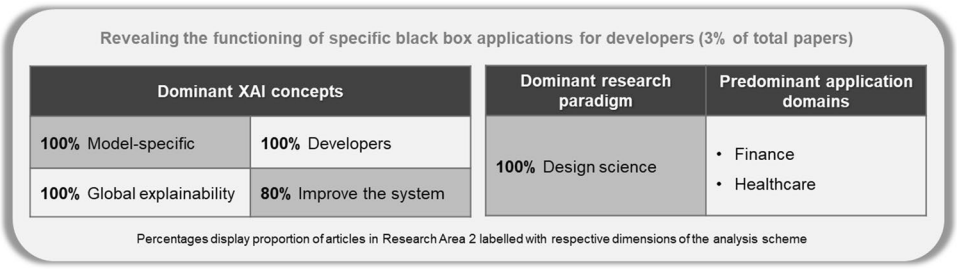


Fig. 6 Overview Research Area 2

method that improves both accuracy and explainability of predictions when describing interactions between two entities in a dyadic dataset. Due to the rather technical nature of the papers in Research Area 2, methods are not designed for or evaluated with electronic markets so far. However, XAI approaches in this research area might serve as a starting point to design novel XAI systems for digital platforms, for example, credit or sales platforms featuring AI systems.

Proof whether resulting explanations assist developers, as intended, is still pending. None of the papers in Research Area 2 includes an evaluation with humans. Sixty percent perform a functionally grounded evaluation. For instance, Martens et al. (2009) implement their rule extraction approach on several datasets and prove that it yields a performance increase in predictive accuracy compared to other rule extraction approaches.

The lack of evaluation with humans directly translates into a call for future research. In the next step, researchers should investigate the quality and efficacy of explanations from developers' perspectives. Moreover, in line with the rather technical focus of this research, improvements in the technical applicability of XAI methods, such as calculation speed, are suggested (Eiras-Franco et al., 2019).

Research Area 3: Explaining AI decisions of specific critical black box applications for domain experts

When utilizing complex AI systems as tools for decision-making, the reasons for particular AI outcomes often remain impenetrable to users. However, especially in critical application domains, AI decisions should not be acted upon blindly, as consequences can be severe (e.g., Gu et al., 2020; Su et al., 2021; Zhu et al., 2021). Thus, Research Area 3, encompassing 24 papers (13%), proposes XAI methods to generate explanations for particular outcomes of specific AI-based decision support systems. Decision support systems incorporating AI predictions and respective explanations serve to support domain experts in their daily work. Examples include anticipation of patient no-show behavior (Barrera Ferro et al., 2020), legal judgments (Zhong et al., 2019), and fault detection in industrial processes (Ragab et al., 2018). Some XAI methods are specifically designed for application in electronic markets, for example, mobile malware prediction (Iadarola et al., 2021), early risk detection in social media (Burdisso et al., 2019), and cost prediction for digital manufacturing platforms (Yoo & Kang, 2021) (Fig. 7).

Researchers commonly agree that AI-based decision support systems must be accompanied by explanations to effectively assist practitioners (e.g., Chatzimparmpas et al., 2020; Gu et al., 2020; Kwon et al., 2019). Thereby, explanations help practitioners better understand AI's reasoning, appropriately trust AI's recommendations, and take the best possible decisions (Hatwell et al., 2020; Hepenstal et al., 2021; Sun et al., 2021). Against this background, explanations are designed to be user-centric, i.e., to address the specific needs of certain (groups of) users. For instance, Barrera Ferro et al. (2020) propose a method to help healthcare professionals counteract low attendance behavior. Their XAI-based decision support system identifies variables explaining no-show probabilities. By adding explainability, the authors aim to prevent both practical and ethical issues when implementing the decision support system in a preventive medical care program for underserved communities in Columbia, identifying, e.g., income and local crime rates affect no-show probabilities.

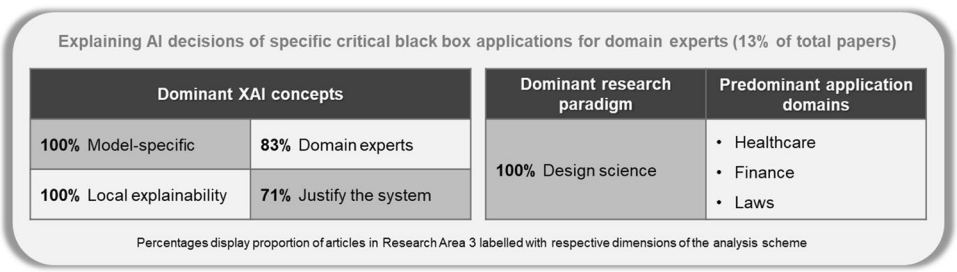


Fig. 7 Overview Research Area 3

To provide domain experts with explanations that meet their requirements, XAI methods to produce visual explanations along AI decisions are often employed: For instance, Gu et al. (2020) utilize an importance estimation network to produce visual interpretations for the diagnoses made by a classification network and demonstrate that the proposed method produces accurate diagnoses along fine-grained visual interpretations. Researchers argue that visualization allows users to easily and quickly observe patterns and test hypotheses (Kwon et al., 2019). Considering the drawbacks, visualizations of large and complex models such as random forests remain challenging (Neto & Paulovich, 2021).

Research Area 3 provides an above-average quota of evaluations with humans (33%). Majorly, researchers conduct user studies to assess the effectiveness of explanations (e.g., Chatzimparmpas et al., 2020; Neto & Paulovich, 2021; Zhao et al., 2019; Zhong et al., 2019). For example, Zhao et al. (2019) conduct a qualitative study with students and researchers to investigate the perceived effectiveness of an XAI-based decision support system in helping users understand random forest predictions in the context of financial scoring. Kumar et al. (2021) even go a step further and implement their XAI approaches in clinical practice to evaluate the trust level of oncologists working with a diagnosis system.

Existing research paves the way for three patterns with regard to future opportunities. First, researchers stress the need for other types of explainability to ensure a sufficient understanding of AI by users (Neto & Paulovich, 2021). Second, researchers propose to transfer XAI methods to different applications (Mensa et al., 2020). For instance, a novel XAI approach to design a conversational agent (Hepenstal et al., 2021) could also be applied in electronic markets. Third, whenever human evaluation is conducted in simulated scenarios with simplified tasks, there is a call to conduct application-grounded evaluation, such as field studies (Chatzimparmpas et al., 2020) and long-term studies (Kwon et al., 2019).

Research Area 4: Explaining AI decisions of specific black box applications for lay users

Similar to Research Area 3, Research Area 4, with seven papers (4%), focuses on model-specific XAI approaches to produce local explanations. While Research Area 3 targets AI users in a professional context, XAI approaches in Research Area 4 address lay people, such as users of a music platform seeking personalized recommendations (Kouki et al., 2020) or evaluating whether texts are similar in terms of meaning (Lopez-Gazpio et al., 2017). Thus, this research area is highly relevant for electronic markets (Fig. 8).

Given that AI finds its way to many areas of everyday life, the relevance of providing lay users with tailored support when faced with AI systems increases (Wang et al., 2019). The "target of XAI [in Research Area 4] is an end user who depends on decisions, recommendations, or actions produced by an AI and therefore needs to understand the rationale for the system's decisions" (Kim et al., 2021, p. 2). Often, lay users, such as people affected by automated AI decisions or users of AI in daily life, are assumed to provide a relatively low level of AI literacy (Wang et al., 2019). Explanations shall help them to easily scrutinize AI decisions and confidently employ AI systems (Kim et al., 2021; Kouki et al., 2020). Like in Research Area 3, researchers predominantly develop approaches to generate explanations for particular outcomes of specific AI models. Most resulting explanations are visual (Kim et al., 2021, 2020a; Kouki et al., 2020; Wang et al., 2019).

Research Area 4 provides an above-average percentage of evaluation with (potential) users (57%) (Kim et al., 2021; Kouki et al., 2020; Lopez-Gazpio et al., 2017). For instance, Kim et al. (2021) experimented with undergraduate students using an XAI system for video search to evaluate the quality of explanations and their effect on users' level of trust. They find that the XAI system yields a comparable level of efficiency and accuracy as its black box counterpart if the user exhibits a high level of trust in the AI explanations. Lopez-Gazpio et al. (2017) conduct two user studies to show that users perform AI-supported text processing tasks better with access to explanations. Only one paper follows functionally grounded evaluation, using a Netflix dataset (Zhdanov et al., 2021), showing that explainability does not need to impact predictive performance negatively.

One commonly mentioned avenue for future research is to transfer XAI approaches—which are often developed for specific applications—to other contexts. For instance, an XAI approach designed for a medical diagnosis tool for lay users might also be beneficial when integrated into a fitness

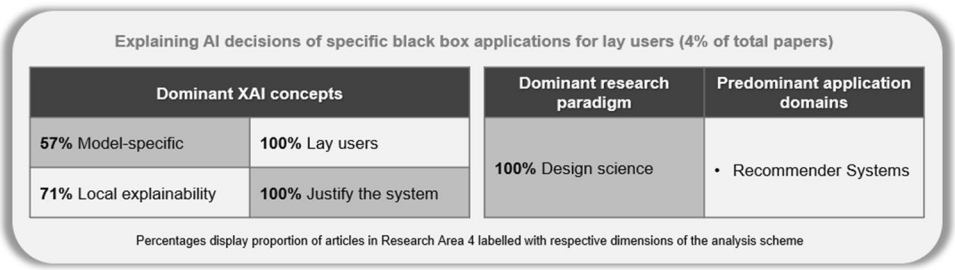


Fig. 8 Overview Research Area 4

app (Wang et al., 2019). While the authors formulate the need to investigate the effectiveness of explanations for lay users (Kouki et al., 2020), the lack of functionally grounded evaluation also translates into a call for a technical assessment and improvement of XAI approaches, such as computation time (Kim et al., 2020a).

Research Area 5: Explaining decisions and functioning of arbitrary black boxes

The ubiquitous nature of AI and its deployment in an increasing variety of applications is accompanied by a rising number of AI models. Consequently, the need for XAI approaches that can work independently from the underlying AI model arises (e.g., Ming et al., 2019). Research Area 5, among the most prominent research areas with 52 papers (29%), addresses this call and develops model-agnostic XAI approaches (Moreira et al., 2021). Many methods have already been applied for electronic markets, for example, for B2B sales forecasting (Bohanec et al., 2017) or prediction of Bitcoin prices (Giudici & Raffinetti, 2021) (Fig. 9).

Papers in Research Area 5 are also driven by the desire to make the outcomes and functioning of AI systems more understandable to users (Fernandez et al., 2019; Li et al., 2021; Ribeiro et al., 2016). First and foremost, explanations intend to assist users in appropriately trusting AI, i.e., critically reflecting on an AI system's decision instead of refusing to use it or blindly following it (Förster et al., 2020b). However, aiming to contribute to the explainability of arbitrary AI models, methods differ from Research Areas 1 to 4 in two ways.

First, methods are not designed to address specific needs in certain applications but aim to explain how and why models make their decisions in general (e.g., Blanco-Justicia et al., 2020). The target group are users of all "domains where ethical treatment of data is required" (Ming et al., 2019, p. 1), including domain experts (79%), such as managers or decision-makers (Bohanec et al., 2017) as well as lay users (38%), such as social media users supported by AI to detect hate speech (Bunde, 2021). In the latter example, researchers show that a dashboard showing and explaining whether a text contains hate is perceived as valuable by users, and that the XAI feature increased the perception of usefulness, ease of use, trustworthiness, and use intention of the artifact. Explanations are constructed to address the standard requirements of various AI users of different application domains. As a result, explanations are often accessible to a wider audience and help users with little AI experience understand, explore, and validate opaque systems (Ming et al., 2019). For example, for the identification of diseases on an automatic diagnosis platform for doctors and patients, building an understandable diagnostics flow for doctors and patients (Zhang et al., 2018). Second, the XAI methods are not designed to be technically tied to specific AI models, but to be applied to various AI models (Mehdiyev & Fettke, 2020, p. 4). Thus, XAI approaches in this area only access the inputs and outcomes without making architectural assumptions regarding the AI model (Ming et al., 2019).

Most papers in Research Area 5 focus on local explanations (73%). A well-known local method is LIME which identifies important features for particular AI predictions by learning easy-to-interpret models locally around the inputs (Ribeiro et al., 2016). Researchers stress that explanations should be human-friendly to facilitate human understanding of the reasons for AI decisions (e.g., Cheng et al., 2021). For instance, Blanco-Justicia et al. (2020) aim at human-comprehensible explanations by limiting the depth of decision trees that approximate the AI model's functioning. Many researchers focus on methods to generate counterfactual explanations, which align with how humans construct explanations themselves (Cheng et al., 2021; Fernandez et al., 2019; Förster et al., 2021). Counterfactual explanations point out why the AI system yields a particular outcome instead of another similarly perceivable one.

The focus of Research Area 5 lies on the XAI methods themselves rather than specific applications. Accordingly, researchers choose relevant but exemplary use cases to evaluate their proposed XAI methods, such as the prediction of credit risk (Bastos & Matos, 2021), churn prediction (Lukyanenko et al., 2020), or mortality in intensive care units (Kline et al., 2020). To demonstrate versatile applicability, researchers often implement their approaches on a range of datasets from different domains including applications in electronic markets such as fraud detection (Hardt et al., 2021) or news-story classification for online advertisements, which helps improve data quality and model performance

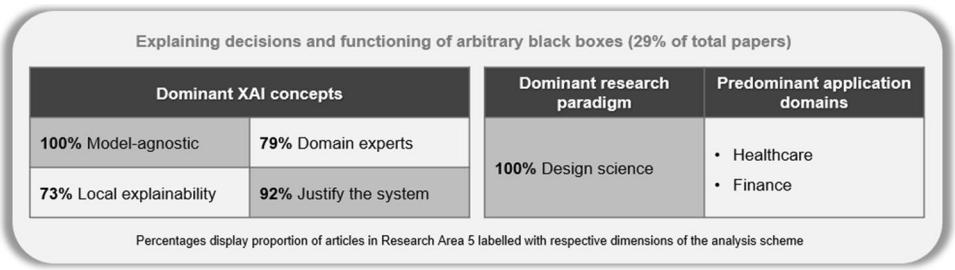


Fig. 9 Overview Research Area 5

(Martens & Provost, 2014). XAI approaches in Research Area 5 could beyond be applied to electronic markets—for example, an XAI dashboard consolidating a large amount of data necessary for child welfare screening is also considered helpful for different data-intensive online platforms (Zytek et al., 2021).

Like in Research Areas 1 and 2, most papers conduct functionally grounded evaluation (52%). However, as repeatedly stated by the authors in this research area, XAI methods are designed to assist humans in building appropriate trust (e.g., Bunde, 2021; van der Waa et al., 2020). Accordingly, in recent years, papers include evaluations with users (46%) (Abdul et al., 2020; Hardt et al., 2021; Ming et al., 2019). User studies serve, for instance, to assess perceived characteristics of explanations (Förster et al., 2020b, 2021) or to compare the utility of different explanations for decision-making (van der Waa et al., 2020). Researchers often resort to simplified tasks with subjects being students (Štrumbelj & Kononenko, 2014) or recruited via platforms like Amazon Mechanical Turk (van der Waa et al., 2020).

As evaluation is often conducted in somewhat artificial settings, researchers propose to evaluate model-agnostic XAI methods in realistic or real settings, for instance, through field experiments (Bohanec et al., 2017; Förster et al., 2020b, 2021; Giudici & Raffinetti, 2021). Other recurring themes for future research include the expansion of the ideas to other application domains (e.g., Spinner et al., 2020; Zytek et al., 2021). Finally, further empirical research is requested to identify required modifications of existing XAI approaches and specific requirements that can serve as a starting point for the design of novel XAI methods (Moradi & Samwald, 2021).

Research Area 6: Investigating the impact of explanations on lay users

There is a substantial body of literature developing XAI methods to automatically generate explanations (cf. Research Areas 1 to 5); however, insights on the role of explainability in human-AI interaction are somewhat rare (Ha et al., 2022; Narayanan et al., 2018; Schmidt et al., 2020). Against this background, this research area with 22 articles (12%) empirically investigates user experience and user behavior in response to explanations, such as understanding of and trust in the underlying AI system (Dodge et al., 2018; Shin, 2021a; van der Waa et al., 2021). The focus lies on lay users as an explanation's target group of (100%). Many papers investigate XAI for electronic market applications—for example, recommendation of online news articles (Shin, 2021a), intelligent tutoring (Conati et al., 2021), or credit risk assessment (Moscato et al., 2021) (Fig. 10).

Researchers stress the importance of involving users to derive how explanations should be designed (Wanner et al., 2020b). Articles in this research area pursue two goals: (i) generating insights on how explanations affect the interaction between users and AI and (ii) deriving requirements for adequate explanations. More concretely, researchers investigate lay user experience and lay user behavior, such as trust (Alam & Mueller, 2021; Burkart et al., 2021; Conati et al., 2021; Hamm et al., 2021; Jussupow et al., 2021; Schmidt et al., 2020; Shin, 2021a, 2021b), understanding (Lim et al., 2009; Shen et al., 2020; Shin, 2021a, 2021b; van der Waa et al., 2021), perception (Fleiß et al., 2020; Ha et al., 2022; Jussupow et al., 2021; Shin, 2021a), and task performance (van der Waa et al., 2021). Lay users considered are, for instance, potential job candidates interacting with conversational agents in recruiting processes (Fleiß et al., 2020) or diabetes patients interacting with a decision support system to determine the correct dosage of insulin (van der Waa et al., 2021). Based on their findings, researchers contribute knowledge on how practical explanations can be designed (Dodge et al., 2018; Förster et al., 2020a; Wanner et al. 2020b). Most of these findings are valid for electronic markets, such as AI-led moderation for eSports communities (Kou & Gui, 2020) or patient platforms with AI as the first point of contact (Alam & Mueller, 2021). The authors of the latter study find that visual and example-based explanations had a significantly better impact on patient satisfaction and trust than text-based explanations or no explanations at all.

A reoccurring study design to investigate user experience and behavior is a controlled experiment with human subjects performing simplified tasks (Lim et al., 2009). For example, Burkart et al. (2021) investigate users' willingness to adapt their initial prediction in response to four treatments with different degrees of explainability. Surprisingly, in their

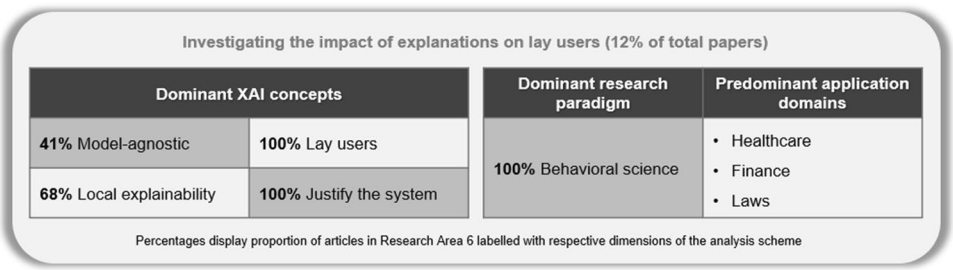


Fig. 10 Overview Research Area 6

specific study, all participants improved their predictions after receiving advice, regardless of whether it featured an explanation. Likewise, Jussupow et al. (2021) investigate users' trust in a biased AI system depending on whether explanations are provided or not. They find that users with low awareness of gender biases perceive a gender-biased AI system that features explanations as trustworthy, as it is more transparent than a system without explanations. Focusing on user experience, Shen et al. (2020) examine users' subjective preferences for different degrees of explainability. Only a few papers build their work on existing theories. For instance, Hamm et al. (2021) adapt the technology acceptance model to examine the role of explainability on user behavior.

The results in Research Area 6 reveal that explanations indeed affect user experience and user behavior. Most papers propose a positive effect on human-AI interaction, such as an increase of users' trust in the AI system (Lim et al., 2009) or intention to reuse the system (Conati et al., 2021). However, some studies indicate a contrary effect, i.e., participants supported by an AI-based decision support tool for text classification reported reduced trust in response to increased transparency (Schmidt et al., 2020). Beyond, the findings of this research area inform how explanations should be built to be effective. For instance, Burkart et al. (2021) found that while local and global explanations help improve participants' decisions, local explanations are used more often. The findings by Förster et al. (2020a) indicate that concreteness, coherence, and relevance are decisive characteristics of local explanations and should guide the development of novel XAI methods. Overall, researchers conclude that user involvement is indispensable to assess if researchers' assumptions on explanations hold (Shin, 2021a; van der Waa et al., 2021).

Results from this research area mainly stem from experiments with recruited participants for simplified tasks, such as students (Alam & Mueller, 2021). Paving the way for future research, researchers stress the importance of verifying findings with real users performing actual tasks (Shen et al., 2020). Furthermore, there is a call for longitudinal studies considering that users' characteristics and attitudes might change over time (Shin, 2021a). Finally, while first progress is made to consider mediating factors predicting the influence of explainability (e.g., Shin, 2021a), most works do not tie their studies to theories; thus, there is a call for developing and testing theories (Hamm et al., 2021).

Research Area 7: Investigating the impact of explanations on domain experts

Most XAI methods are designed to assist domain experts in interacting with AI-based decision support systems. To better understand how explainability influences user experience and user behavior in this regard, Research Area 7 includes 17 empirical papers (9%) with a focus on domain experts, such as doctors (Ganeshkumar et al., 2021; Kim et al., 2020b) or decision-makers in credit scoring (Huysmans et al., 2011). Compared to Research Area 6, fewer papers investigate the impact of explanations in electronic market applications. Examples include an AI-based scheduling platform for healthcare professionals (Schlicker et al., 2021) and an AI web application for patient analysis and risk prediction (Fang et al., 2021) (Fig. 11).

Researchers argue that while there is agreement on the need to increase the explainability of critical AI applications, insights on how different explanation types affect the interaction of domain experts with AI is rare (Liao et al., 2020). This research area aims to understand the impact of explainability concerning user experience and user behavior in the context of AI-based decision support systems (Chakraborty et al., 2021; Elshawi et al., 2019; Liao et al., 2020; Martens et al., 2007). Similar to Research Area 6, findings aim to provide knowledge on how to design adequate explanations, however, with a focus on domain experts (Liao et al., 2020; Wanner et al., 2020a).

A reccurring research approach is to conduct experiments investigating the impact of explainability on users' decision-making with AI. In a pioneering paper, Huysmans et al. (2011) examine how different degrees of explainability affect AI system comprehensibility in a laboratory experiment. They find that decision tables perform significantly better than decision trees, propositional rules, and oblique rules with regard to accuracy, response time, answer confidence, and ease of use. Moreover, researchers conduct interviews to assess user needs for explainability in critical AI applications (Liao et al., 2020).

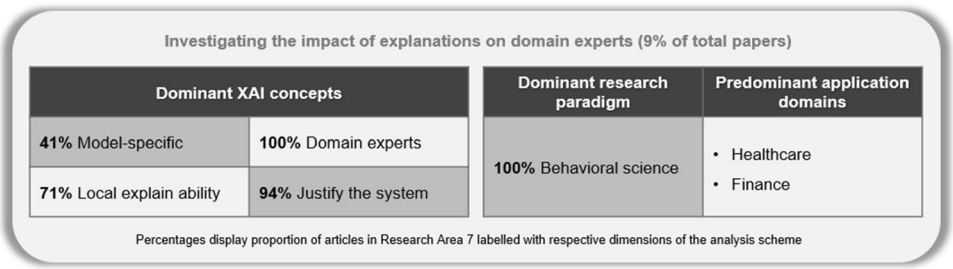


Fig. 11 Overview Research Area 7

Overall, findings from these studies indicate that explainability can positively influence user experience and user behavior of domain experts. The findings by Huysmans et al. (2011) outlined above suggest that explainability in the form of decision tables can lead to faster decisions while increasing answer confidence. Additionally, findings inform how explanations should be designed and applied to yield specific effects. For example, Elshawi et al. (2019) reveal that local explanations are suitable for medical diagnoses to foster users' understanding while global explanations increase users' understanding of the entire AI model. Although this research area proves the benefit of XAI for domain experts, practitioners still struggle with the gaps between existing XAI algorithmic work and the aspiration to create human-consumable explanations (Liao et al., 2020).

While existing studies show that types of explanations, such as local and global explanations, vary in effectiveness on users' system understanding, future research may deepen these insights and investigate other concepts, such as concreteness and coherence. Furthermore, researchers stress the importance of further investigating how users' characteristics moderate explanations' influence on user experience and user behavior (Bruijn et al., 2021). Expert users of electronic markets are not the focus of research attention yet. Finally, while most researchers focus on the impact of explanations on users' perceptions and intentions, there is a call for research on actual behavior (Bayer et al., 2021).

Research Area 8: Investigating employment of XAI in practice

In contrast to Research Areas 6 and 7, which comprise empirical studies to investigate user experience and user behavior, Research Area 8 focuses on technical and managerial aspects of XAI in practice. For instance, researchers conduct case studies to examine scalability (Sharma et al., 2020) and trade-offs of XAI in practice (Tabankov & Möhlmann, 2021). The four papers (2%), which all were published between 2019 and 2021, represent the smallest research area. Findings predominantly address developers (100%) and managers who want to implement XAI in organizations (Sharma et al., 2020) (Fig. 12).

The motivation for this research area is a scarce understanding of organizational and technical challenges practitioners face when implementing explanations for AI (Hong et al., 2020). Researchers agree that this might hinder XAI from addressing critical real-world needs. Against this background, empirical studies aim to generate insights into how XAI can be successfully employed in organizations (Hong et al., 2020; Tabankov & Möhlmann, 2021).

To this end, Hong et al. (2020) conduct semi-structured interviews with industry practitioners to examine the role of explainability when developers plan, build, and use AI models. One important finding is the high practical relevance of scalability and integrability of XAI methods—which has not yet been the focus of existing research. Building on these insights, Sharma et al. (2020) evaluate the performance of XAI methods with respect to technical aspects in an electronic market–related case study, i.e., anomaly detection for cloud-computing platforms. Findings reveal that the computation time of tree-based XAI methods should be improved to enable the large-scale application. Tabankov and Möhlmann (2021), with their case study, take a managerial perspective and investigate trade-offs between explainability and accuracy of XAI for in-flight services. Findings suggest that compromises and limitations for both sides have to be weighed during the implementation process.

Insights from this research area pave the way for future research: First, when developing novel XAI methods, researchers should consider technical aspects, first and foremost, scalability (Hong et al., 2020). This is especially relevant for electronic market applications, which often need to adapt to sudden user growth. Second, more empirical research on XAI from an organizational and managerial perspective is needed. In particular, further research might provide deeper insights into whether and to what extent explainability is needed to achieve organizational goals (Tabankov & Möhlmann, 2021). Third, there is a call for insights into the demands of XAI developers (Hong et al., 2020).

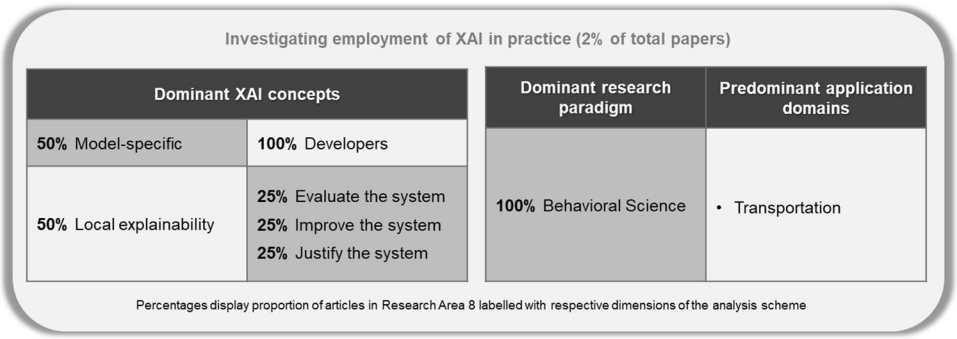


Fig. 12 Overview Research Area 8

Synthesis of XAI research areas in IS literature

In sum, based on theoretical concepts of XAI research and methodological concepts of IS research, a cluster analysis reveals eight major XAI research areas in IS literature (cf. Fig. 13, Appendix).

Five research areas (76% of all papers in our corpus) deal with developing novel XAI approaches. This body of literature can be further differentiated depending on the underlying XAI concepts, first and foremost dependency on the AI model and scope of explainability, as well as whom explanations address. Research Area 1 and Research Area 2 both focus on model-specific XAI approaches to generate global explanations for expert audiences—domain experts in Research Area 1, and developers in Research Area 2. Research Area 3 and Research Area 4 entail largely local explanations for specific AI models that address domain experts and lay users, respectively. Research Area 5 features model-agnostic approaches. Overall, the primary purpose of explanations is to justify the (decisions of) AI systems (Research Areas 1, 3, 4, and 5).

The remaining three research areas comprise fewer articles (24%) focusing on behavioral science research. Note that in our case, the term "behavioral science" not only refers to studies that build and justify theory, for instance, in deriving and testing hypotheses but, more generally, includes research that aims at generating empirical insights. Indeed, only a few XAI papers in IS derive and test hypotheses. Empirical research in our corpus can be distinguished by its focus on specific target groups. While Research Area 6 focuses on lay users, Research Area 7 deals with users with domain knowledge. Research Area 8 focuses on developers.

Discussion and conclusion

We conducted a systematic and structured review of research on XAI in IS literature. This section outlines opportunities for future research that may yield interesting insights into the field but have not been covered so far. Subsequently, we describe our work's contribution, implications, and limitations.

Future research agenda

Our synthesis reveals five overarching future research directions related to XAI research in IS, which, along with a related future research agenda, are outlined below: (1) refine the understanding of XAI user needs, (2) reach a more comprehensive understanding of AI, (3) perform a more diverse mix of XAI evaluation, (4) solidify theoretical foundations on the role of XAI for human-AI interaction, and (5) increase and improve the application to electronic market needs. Note that the future research directions and future research agenda are by no means exhaustive but intend to highlight and illustrate potential avenues that seem particularly promising.

Future Research Direction 1: Refine the understanding of XAI user needs

XAI research is criticized for not focusing on user needs, which is a prerequisite for the effectiveness of explanations (cf. Herse et al., 2018; Meske et al., 2020). Indeed, as argued in many papers in the different research areas identified, there is still a gap between the research's focus on novel algorithms and the aspiration to create human-consumable explanations (e.g., Liao et al., 2020; Seera & Lim, 2014). Areosa and Torgo

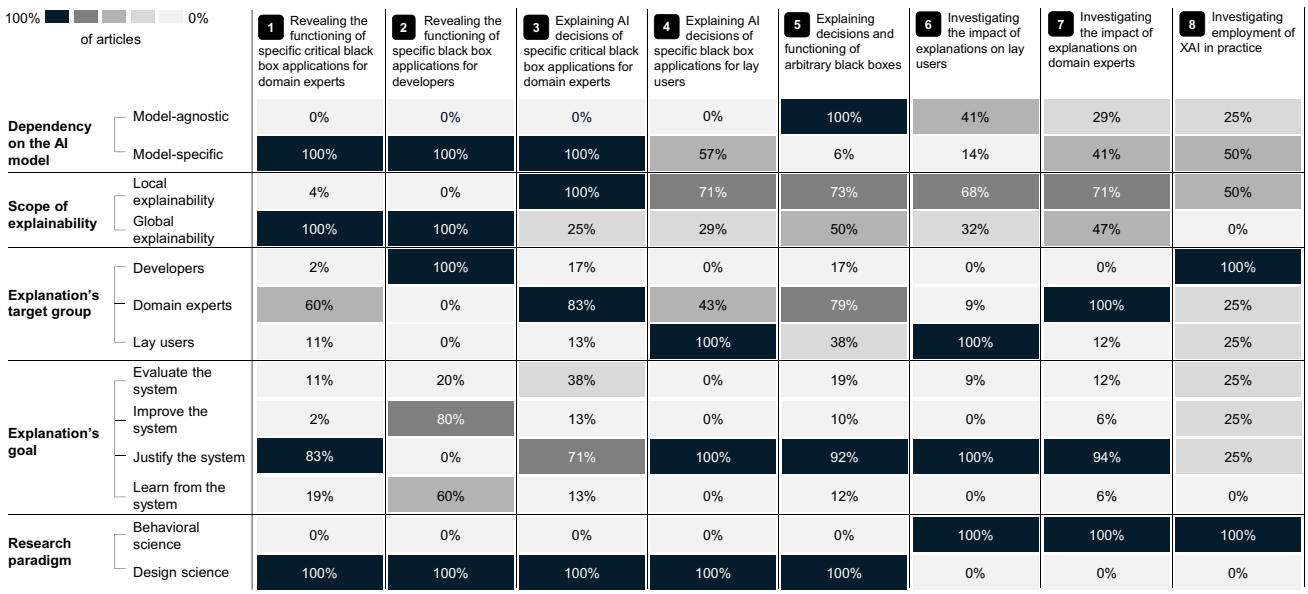


Fig. 13 Synthesis of XAI research areas in IS literature

(2019) stress the necessity to provide insights into the type of usage and information XAI tools bring to end users. As one of the foci in IS research is the design of user-centric and interactive technologies, IS research is predestined to put the user at the center of attention and make explanations understandable (Bauer et al., 2021). While six of the eight research areas focus on broader user groups, i.e., lay users, domain experts, or developers, only a few studies base the design of XAI approaches on specified target users and their needs (e.g., medical experts with different level of domain knowledge). This shortcoming has already been raised in studies that call for a more user-specific design of XAI solutions (cf. Abdul et al., 2018; Miller, 2019). However, only a few studies have implemented user-specific designs so far. For instance, Barda et al. (2020) propose an XAI approach that produces explanations for predictions based on a pediatric intensive care unit's mortality risk model. It considers user-specific explanation and information goals, which vary according to the clinical role (e.g., nurses and physicians). Further empirical insights highlight the necessity for the user-specific design of explanations, as XAI can only create human agency and appropriate trust if it considers the specific user needs (Dodge et al., 2018; Elshawi et al., 2019).

We identify several research opportunities to pave the way for a refined understanding of XAI user needs: First, more empirical research might sharpen insights into how different types of explanations affect the behavior and experience of various user groups and which effects different explanation types might have on these groups—for example, medical practitioners (e.g., Seera & Lim, 2014). Second, future research could refine the differentiation between developers, domain experts, and lay users, as other user characteristics besides expertise might play a central role (e.g., Cui et al., 2019). For instance, the user's knowledge structure, beliefs, interests, expectations, preferences, and personality could be considered (Miller et al., 2017). Third, the conjunction of user characteristics and the purpose of explanations could be analyzed, especially given that the purpose of explanations depends on the context and user type (Liao et al., 2020). Fourth, future research could put more emphasis on investigating the concrete XAI needs of developers, which would benefit from explainability (cf. Kim et al., 2021) but are so far seldomly addressed. This is underlined by the fact that in Research Area 2 ("Revealing the functioning of specific black box applications for developers"), the only research area focusing on developers, none of the papers evaluates its concepts with actual developers.

Future Research Direction 2: Reach a more comprehensive understanding of AI

While a plethora of techniques produce various types of explanations, only a few researchers combine different XAI approaches with the aim of a comprehensive understanding of AI. The overarching goal of XAI is to make AI systems and their outcomes understandable to humans, especially important when AI supports decision-making in critical areas such as healthcare and finance (Pessach et al., 2020). Single (types of) explanations are often insufficient to reach the ambitious goal of comprehensive user understanding. Many researchers underpin that their approaches are only one step toward a better understanding of the underlying AI systems (e.g., Moradi & Samwald, 2021; Neto & Paulovich, 2021). However, the question of how to synthesize different research efforts to get closer to a comprehensive understanding of AI systems has received little research attention. Especially in Research Area 1 ("Revealing the functioning of specific critical black box applications for domain experts") and Research Area 3 ("Explaining AI decisions of specific critical black box applications for domain experts"), both of which focus on domain experts, researchers identify the need for further explanation types to ensure that users can reach a more comprehensive understanding of AI (e.g., Murray et al., 2021; Neto & Paulovich, 2021).

Against this backdrop, promising future research opportunities arise: First, it could be beneficial to investigate the combination of different types of explanations which might complement each other for user understanding, e.g., local and global explanations, a call made in many of the analyzed papers (cf. Burkart et al., 2021; Elshawi et al., 2019; Mombini et al., 2021). So far, efforts on developing novel approaches mainly concentrate on either type, with only 18% of the papers combining local and global interpretability (e.g., Burkart et al., 2021; Elshawi et al., 2019). Second, a stronger focus on user interfaces might serve as an auspicious starting point for a more complete understanding of AI. For example, interactivity would allow users to explore an algorithm's behavior, and XAI approaches to adapt explanations to users' needs (Cheng et al., 2019). Ming et al. (2019) provide the first promising attempts in this direction, developing an interactive visualization technique to help users with little AI expertise understand, explore, and validate predictive models. Third, personalized explanations taking into account users' mental models and the application domain can foster understanding (Schneider & Handali, 2019). Kouki et al. (2020) are among the first to study the problem of generating and visualizing personalized explanations for recommender systems.

Future Research Direction 3: Perform a more diverse mix of XAI evaluation

Our analysis shows that existing IS literature on XAI exposes a one-sided tendency toward the functional evaluation of XAI approaches. Seminal design science contributions emphasize the need for rigor in evaluating IT artifacts, including

functional evaluations but also "the complications of human and social difficulties of adoption and use" (Venable et al., 2016, p. 82). While the latter plays a significant role in the context of XAI, 71% of the articles that develop XAI approaches in our corpus neglect evaluation with (potential) users. Only 6% combine functional evaluation with user evaluation. Thus, existing research runs the risk of inaccurate insights derived from unduly simplified evaluation scenarios (Wang et al., 2019). In almost all research areas, papers identify a better mix of evaluation methods as one of the most important directions for future research (e.g., Chatzimparmpas et al., 2020; Kim et al., 2020b).

Proposed avenues for further research are closely linked to a call for a more diverse mix of different kinds of evaluations (cf. Venable et al., 2016). First, XAI approaches should be more frequently evaluated with humans (cf. human-grounded evaluation) to take into account human risks associated with novel XAI approaches. For example, many papers in Research Area 1 ("Revealing the functioning of specific critical black box applications for domain experts") call for a more robust evaluation, including human users (e.g., Areosa & Torgo, 2019; Kim et al., 2020b). Second, there should be a stronger focus on evaluation with real users in real settings (cf. application-grounded evaluation) to assess the utility, quality, and efficacy of novel approaches in real-life scenarios. This point is stressed by several papers in Research Area 3 ("Explaining AI decisions of specific critical black box applications for domain experts") (e.g., Chatzimparmpas et al., 2020; Kwon et al., 2019) and Research Area 6 ("Investigating the impact of explanations on lay users") (e.g., Shen et al., 2020; Shin, 2021a). Third, novel evaluation strategies might be investigated that combine functionally and human-grounded evaluation to consolidate the benefits of both, i.e., the possibility of a robust comparison of competing XAI approaches at relatively low cost and the consideration of social intricacies.

Future Research Direction 4: Solidify theoretical foundations on the role of XAI for human-AI interaction

Our examination shows that XAI in IS research is predominantly not very theory-rich. While broad efforts to develop novel artifacts exist, only few papers (24%) explicitly focus on contributions to theory by conducting empirical research. These studies generate first exciting insights into how explainability may affect the experience and behavior of AI users (cf. Research Areas 6 and 7); however, only 13 papers explicitly tie their research to theory. The following IS theories have been used to investigate XAI in our literature corpus: Activity Theory (Kou & Gui, 2020), Agency Theory (Wanner et al., 2020a), Attribution Theory (Ha et al., 2022; Schlicker et al., 2021), Cognitive Fit Theory (Huysmans et al., 2011), Elaboration Likelihood Model/Heuristic Systematic Model (Shin, 2021a, 2021b; Springer & Whittaker, 2020), Information Boundary Theory (Yan & Xu, 2021), Information Foraging Theory (Dodge et al., 2018), Information Processing Theory (Sultana & Nemati, 2021), Psychological Contract Violation (Jussupow et al., 2021), Technology Acceptance Model/Theory of Planned Behavior/Theory of Reasoned Action (Bayer et al., 2021; Wanner et al., 2020a), Theory of Swift Trust (Yan & Xu, 2021), and Transaction Cost Theory (Wanner et al., 2020a). Mainly, cognitive theories are employed. As the human side of explanations is both social and cognitive, literature points out that explainability in the context of human-AI interaction should be viewed through a cognitive and a social lens (Berente et al., 2021; Malle, 2006). The extant studies pave the way for a diverse and meaningful XAI research agenda. It is crucial to add theoretical lenses (Wang et al., 2019), to deepen the understanding of the role of XAI for human-AI interaction. Extant literature stresses the need to further develop and test theories, for example, concerning the relationship between XAI and use behavior (Hamm et al., 2021).

Pursuing this avenue, first, we call to supplement insights based on cognitive theories by investigating XAI through a social lens. Second, it might be helpful not only to include and test IS theories but also theories from disciplines such as social sciences, management, and computer science. XAI is multidisciplinary by nature with people, information technology, and organizational contexts being intertwined. For instance, the social sciences might be promising to model user experience and behavior as they aim to understand how humans behave when explaining to each other (Miller, 2019). Third, as extant empirical studies are mostly limited to one-time interactions between humans and XAI, more research on the long-term influence of explanations is needed. For instance, the question of how explanations may sustainably change users' mental models and behavior should gain more attention. Papers in our body of literature also call for longitudinal studies considering that users' characteristics and attitudes might change over time (Shin, 2021a). Fourth, the organizational perspective on XAI is mainly neglected. Existing literature examines AI's influence on the competitiveness of companies (e.g., Rana et al., 2022). For different organizations, AI has become an essential source of decision support (Arrieta et al., 2020); thus, XAI is of utmost importance for bias mitigation (Akter et al., 2021a; Zhang et al., 2022). Therefore, it would be beneficial to examine the role of XAI from an organizational perspective as well.

Future Research Direction 5: Increase and improve the application to electronic market needs

The literature review shows that only a minority of extant studies aim at solving electronic market-related challenges (e.g., Burdisso et al., 2019; Irarrázaval et al.,

2021). Among business applications, XAI is especially relevant for electronic markets, as trust is paramount in all buyer-seller relationships (Bauer et al., 2020; Marella et al., 2020). Promising first studies on XAI in electronic markets focus on recurring use cases, for example, recommender systems in entertainment (e.g., Zhdanov et al., 2021), patient platforms in healthcare (e.g., van der Waa et al., 2021), and credit platforms in finance (e.g., Moscato et al., 2021). Given that electronic markets are increasingly augmented with AI-based systems and their complex nature is often an obstacle (Adam et al., 2021; Thiebes et al., 2021), electronic markets provide large potential for XAI research. To illustrate, the benefit of XAI could be explored for AI-based communication with customers on company platforms or AI-augmented enterprise IS for domain experts in supply chain or customer relationship management. While the benefits of XAI in electronic markets become obvious, an XAI research agenda with a focus on the needs of electronic markets might, in turn, benefit from diverse cases, including a variety of users.

There are three possible pathways in which researchers could address this issue and improve the application to electronic markets: First, existing XAI approaches could be transferred to and investigated in electronic markets. For instance, an XAI approach for conversational agents (Hepenstal et al., 2021) could be applied in electronic markets, for example, in the context of B2C sales platforms or for customer support. Second, given the strong interaction of people and technology in electronic markets (cf. Thiebes et al., 2021), it is pivotal to gain a better understanding of users' needs regarding the explainability of AI in electronic markets, for example, users of music platforms (Kouki et al., 2020), news websites (Shin, 2021a), or streaming platforms (Zhdanov et al., 2021) seeking personalized recommendations. Third, researchers could develop novel XAI methods and user interfaces that specifically meet electronic market needs, for instance, the ability to work with large amounts of data and provide interactive interfaces for business and private users. Table 2 summarizes the future research directions and opportunities outlined above.

Contribution

The contribution of our study is twofold. First, we provide a structured and comprehensive literature review of XAI research in IS. A literature review is especially important for a young and emerging research field like XAI, as it "uncover[s] the sources relevant to a topic under study" (vom Brocke et al., 2009, p. 13) and "creates a firm foundation for advancing knowledge" (Webster & Watson, 2002, p. 13). XAI draws from various scientific disciplines such as computer science, social sciences, and IS. While existing research already views XAI through the lenses of adjacent disciplines like social sciences (e.g., Miller, 2019), we accumulate the state of knowledge on XAI from the IS perspective. With its multiperspective view, IS research is predestined to investigate and design the explainability of AI. In turn, XAI can significantly contribute to the ongoing discussion of human-AI interaction in the IS

Table 2 Future research agenda

Future research directions	Future research opportunities
1: Refine the understanding of XAI user needs	• Pursue empirical research to sharpen understanding of how explanations affect behavior and experience of user groups • Refine differentiation between user groups for a more complete understanding of XAI end-user characteristics • Analyze the conjunction of XAI user characteristics and the purpose of explanations • Investigate the needs of developers in the context of XAI
2: Reach a more comprehensive understanding of AI	• Investigate the combination of different types of explanations • Investigate user interfaces with a focus on interactivity • Pursue personalized explanations taking users' mental models into account
3: Perform a more diverse mix of XAI evaluation	• Pursue evaluations with human users • Pursue evaluations with real users in real-life scenarios • Combine functionally and human-grounded evaluation
4: Solidify theoretical foundations on the role of XAI for human-AI interaction	• Investigate XAI through a social lens • Pursue interdisciplinary approaches, e.g., employ theories from the social sciences • Research the long-term influence of explanations, e.g., on users' mental models • Examine the role of XAI from an organizational perspective
5: Increase and improve the application to electronic market needs	• Transfer existing XAI approaches to electronic markets • Investigate user needs regarding the explainability of AI in electronic markets • Design XAI approaches that meet specific electronic market requirements

discipline. Compared to existing works on XAI in IS (e.g., Meske et al., 2020), our study is the first to synthesize XAI research in IS based on a structured and comprehensive literature search. The structured and comprehensive literature search reveals 180 research articles published in IS journals and conference proceedings. From 2019 onward, the number of published articles increased rapidly, resulting in 79% of the articles published between 2019 and 2021. Model-specific XAI methods (53%) are more often in focus than model-agnostic XAI methods (38%). Most articles address domain experts as the target group (62%) and focus on the justification of AI systems' decisions as XAI goal (83%). Extant IS research efforts concentrate on developing novel XAI artifacts (76%); however, only 23% of the proposed artifacts are evaluated with humans. A minority of studies aim at building and justifying theories or generating empirical insights (24%). Building on established XAI concepts and methodological orientation in IS, we are the first to derive XAI research areas in IS. Extant XAI research in IS can be synthesized in eight research areas: (1) Revealing the functioning of specific critical black box applications for domain experts (26% of papers), (2) Revealing the functioning of specific black box applications for developers (3% of papers), (3) Explaining AI decisions of specific critical black box applications for domain experts (13% of papers), (4) Explaining AI decisions of specific black box applications for lay users (4% of papers), (5) Explaining decisions and functioning of arbitrary black boxes (29% of papers), (6) Investigating the impact of explanations on lay users (12% of papers), (7) Investigating the impact of explanations on domain experts (9% of papers), (8) Investigating employment of XAI in practice (2% of papers).

Second, we provide a future research agenda for XAI research in IS. The research agenda comprises promising avenues for future research raised in existing contributions or derived from our synthesis. From an IS perspective, the following directions for future research might provide exciting insights into the field of XAI but have not yet been covered sufficiently: (1) Refine the understanding of XAI user needs, (2) Reach a more comprehensive understanding of AI, (3) Perform a more diverse mix of XAI evaluation, (4) Solidify theoretical foundations on the role of XAI for human-AI interaction, (5) Increase and improve the application to electronic market needs. These research directions reflect the imbalance of existing IS research with respect to methodological orientation, which so far focuses on designing novel XAI artifacts and rather neglects to generate empirical insights and develop theory.

Implications

Our findings have implications for different stakeholders of XAI research. IS researchers might benefit from our findings in three different ways. First, the accumulated knowledge helps novice researchers find access to XAI research in IS and assists more experienced researchers in situating their own work in the academic discussion. Second, the presented state of knowledge as well as the future research agenda can inspire researchers to identify research themes that might be of interest to future work. Third, our findings on XAI-receptive publication outlets may assist researchers in identifying potential outlets for their work. Furthermore, editors and reviewers are supported in assessing whether the research under review has sufficiently referenced the existing body of knowledge on XAI in IS and to what extent articles under review are innovative in this field. Finally, given that IS research predominantly addresses business needs (Hevner et al., 2004), our findings are particularly suitable for helping practitioners to make use of the accumulated knowledge on XAI.

Limitations

The findings of this paper have to be seen in light of some limitations. Although we conducted a broad and structured literature search, there exists the possibility that not all relevant articles were identified, due to three reasons. First, while we covered all major IS journals and conferences, the number of sources selected for our literature search is nevertheless limited. Second, although we thoroughly deducted the search terms based on existing XAI literature, additional terms might have revealed further relevant papers. We tried to mitigate this issue by conducting a forward and backward search. Third, by focusing on opaque AI systems, we excluded papers that deal with the explainability of inherently transparent systems, such as rule-based expert systems. Apart from this, by utilizing a quantitative clustering approach to identify research areas, our results do not represent the only possible solution to synthesize existing IS knowledge on XAI. However, our methodology yields a broad, transparent, and replicable overview of XAI research in IS. We hope our findings will help researchers and practitioners gain a thorough overview and better understanding of the body of IS literature on XAI and stimulate further research in this fascinating field.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12525-023-00644-5.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The data that support the findings of this study are available from the authors upon reasonable request.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Barrera Ferro, D., Brailsford, S., Bravo, C., & Smith, H. (2020). Improving healthcare access management by predicting patient no-show behaviour. *Decision Support Systems*, *138*(113398). https://doi.org/10.1016/j.dss.2020.113398

Bastos, J. A., & Matos, S. M. (2021). Explainable models of credit losses. *European Journal of Operational Research, 301*(1), 386–394. https://doi.org/10.1016/j.ejor.2021.11.009

Bauer, I., Zavolokina, L., & Schwabe, G. (2020). Is there a market for trusted car data? *Electronic Markets, 30*(2), 211–225. https://doi.org/10.1007/s12525-019-00368-5

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

Bayer, S., Gimpel, H., & Markgraf, M. (2021). The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*, 1–29. https://doi.org/10.1080/12460125.2021.1958505

Beese, J., Haki, M. K., Aier, S., & Winter, R. (2019). Simulation-based research in information systems. *Business & Information Systems Engineering, 61*(4), 503–521. https://doi.org/10.1007/s12599-018-0529-1

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly, 45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect XAI-assisted decision-making: A systematic review. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 78–91). https://hal.telecom-paris.fr/hal-03684457

Blanco-Justicia, A., Domingo-Ferrer, J., Martinez, S., & Sanchez, D. (2020). Machine learning explainability via microaggregation and shallow decision trees. *Knowledge-Based Systems, 194*(5), 105532. https://doi.org/10.1016/j.knosys.2020.105532

Bohanec, M., Kljajić Borštnar, M., & Robnik-Šikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications, 71*(0957–4174), 416–428. https://doi.org/10.1016/j.eswa.2016.11.010

Bresso, E., Monnin, P., Bousquet, C., Calvier, F.-E., Ndiaye, N.-C., Petitpain, N., Smaïl-Tabbone, M., & Coulet, A. (2021). Investigating ADR mechanisms with explainable AI: A feasibility study with knowledge graph mining. *BMC Medical Informatics and Decision Making, 21*(1), 1–14. https://doi.org/10.1186/s12911-021-01518-6

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). *Notes from the AI frontier: Modeling the impact of AI on the world economy*. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

Bunde, E. (2021). AI-assisted and explainable hate speech detection for social media moderators – A design science approach. *Proceedings of the 2021 Annual Hawaii International Conference on System Sciences (HICSS)* (pp. 1264–1274). https://doi.org/10.24251/HICSS.2021.154

Burdisso, S. G., Errecalde, M., & Montes-y-Gómez, M. (2019). A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications, 133*, 182–197. https://doi.org/10.1016/j.eswa.2019.05.023

Burkart, N., Robert, S., & Huber, M. F. (2021). Are you sure? Prediction revision in automated decision-making. *Expert Systems, 38*(1), e12577. https://doi.org/10.1111/exsy.12577

Chakraborty, D., Başağaoğlu, H., & Winterle, J. (2021). Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Systems with Applications*, *170*(114498). https://doi.org/10.1016/j.eswa.2020.114498

Chakrobartty, S., & El-Gayar, O. (2021). Explainable artificial intelligence in the medical domain: a systematic review. *AMCIS 2021 Proceedings* (p. 1). https://scholar.dsu.edu/cgi/viewcontent.cgi?article=1265&context=bispapers

Chatzimparmpas, A., Martins, R. M., & Kerren, A. (2020). T-viSNE: Interactive assessment and interpretation of t-SNE projections. *IEEE Transactions on Visualization and Computer Graphics, 26*(8), 2696–2714. https://doi.org/10.1109/TVCG.2020.2986996

Cheng, F., Ming, Y., & Qu, H. (2021). Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics, 27*(2), 1438–1447. https://doi.org/10.1109/TVCG.2020.3030342

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–12). https://doi.org/10.1145/3290605.3300789

Chromik, M., & Butz, A. (2021). Human-XAI interaction: A review and design principles for explanation user interfaces. *2021 IFIP Conference on Human-Computer Interaction (INTERACT)* (pp. 619–640). https://doi.org/10.1007/978-3-030-85616-8_36

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 (ExSS-ATEC'20)* (p. 7). Cagliari, Italy. https://ceur-ws.org/Vol-2582/paper9.pdf

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management, 60*, 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence, 298*, 1–23. https://doi.org/10.1016/j.artint.2021.103503

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society, 1*(1), 104–126. https://doi.org/10.1007/BF03177550

Cooper, A. (2004). *The inmates are running the asylum. Why high-tech products drive us crazy and how to restore the sanity* (2nd ed.). Sams Publishing.

Cui, X., Lee, J. M., & Hsieh, J. P. A. (2019). An integrative 3C evaluation framework for explainable artificial intelligence. *Proceedings of the twenty-fifth Americas conference on information systems (AMCIS),* Cancun, 2019. https://aisel.aisnet.org/amcis2019/ai_semantic_for_intelligent_info_systems/ai_semantic_for_intelligent_info_systems/10

DARPA. (2018). *Explainable artificial intelligence*. https://www.darpa.mil/program/explainable-artificial-intelligence. Accessed 02 Feb 2023

de Bruijn, H., Warnier, M., & Janssen, M. (2021). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly, 39*(2), 101666. https://doi.org/10.1016/j.giq.2021.101666

de Santana, Á. L., Francês, C. R., Rocha, C. A., Carvalho, S. V., Vijaykumar, N. L., Rego, L. P., & Costa, J. C. (2007). Strategies for improving the modeling and interpretability of Bayesian networks. *Data & Knowledge Engineering, 63*, 91–107. https://doi.org/10.1016/j.datak.2006.10.005

Dodge, J., Penney, S., Hilderbrand, C., Anderson, A., & Burnett, M. (2018). How the experts do it: Assessing and explaining agent behaviors in real-time strategy games. *Proceedings of the 36th International Conference on Human Factors in Computing Systems (CHI)* (pp. 1–12). Association for Computing. https://doi.org/10.1145/3173574.3174136

Doran, D., Schulz, S., & Besold, T. R. (2018). What does explainable AI really mean? A new conceptualization of perspectives. In T. R. Besold & O. Kutz (Chairs), *Proceedings of the first international workshop on comprehensibility and explanation in AI and ML 2017*. https://ceur-ws.org/Vol-2071/CExAIIA_2017_paper_2.pdf

Doshi-Velez, F., & Kim, B. (2018). Considerations for evaluation and generalization in interpretable machine learning. In *Explainable and Interpretable Models in Computer Vision and Machine Learning* (pp. 3–17). Springer. https://doi.org/10.1007/978-3-319-98131-4_1

Eiras-Franco, C., Guijarro-Berdiñas, B., Alonso-Betanzos, A., & Bahamonde, A. (2019). A scalable decision-tree-based method to explain interactions in dyadic data. *Decision Support Systems*, *127*(113141). https://doi.org/10.1016/j.dss.2019.113141

Elshawi, R., Al-Mallah, M. H., & Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, *19*(146). https://doi.org/10.1186/s12911-019-0874-0

European Commission (Ed.). (2021). *Regulation of the European Parliament and of the Council: Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN. Accessed 02 Feb 2023

Fang, H. S. A., Tan, N. C., Tan, W. Y., Oei, R. W., Lee, M. L., & Hsu, W. (2021). Patient similarity analytics for explainable clinical risk prediction. *BMC Medical Informatics and Decision Making, 21*(1), 1–12. https://doi.org/10.1186/s12911-021-01566-y

Fernandez, C., Provost, F., & Han, X. (2019). Counterfactual explanations for data-driven decisions. *Proceedings of the fortieth international conference on information systems (ICIS).* https://aisel.aisnet.org/icis2019/data_science/data_science/8

Ferreira, J. J., & Monteiro, M. S. (2020). What are people doing about XAI user experience? A survey on AI explainability research and practice. *2020 International Conference on Human-Computer*

Interaction (HCII) (pp. 56–73). https://doi.org/10.1007/978-3-030-49760-6_4

Fleiß, J., Bäck, E., & Thalmann, S. (2020). Explainability and the intention to use AI-based conversational agents. An empirical investigation for the case of recruiting. *CEUR Workshop Proceedings (CEUR-WS.Org)* (vol 2796, pp. 1–5). https://ceur-ws.org/Vol-2796/xi-ml-2020_fleiss.pdf

Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. *Expert Systems with Applications, 42*(13), 5737–5753. https://doi.org/10.1016/j.eswa.2015.02.042

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020a). Evaluating explainable artificial intelligence – what users really appreciate. *Proceedings of the 2020 European Conference on Information Systems (ECIS). A Virtual AIS Conference.* https://web.archive.org/web/20220803134652id_/https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1194&context=ecis2020_rp

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020b). Fostering human agency: a process for the design of user-centric XAI systems. *In Proceedings of the Forty-First International Conference on Information Systems (ICIS). A Virtual AIS Conference*. https://aisel.aisnet.org/icis2020/hci_artintel/hci_artintel/12

Förster, M., Hühn, P., Klier, M., & Kluge, K. (2021). Capturing users' reality: a novel approach to generate coherent counterfactual explanations. *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS). A Virtual AIS Conference*. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/947e7f6b-c7b0-4dba-afcc-95c4edef0a27/content

Ganeshkumar, M., Ravi, V., Sowmya, V., Gopalakrishnan, E. A., & Soman, K. P. (2021). Explainable deep learning-based approach for multilabel classification of electrocardiogram. *IEEE Transactions on Engineering Management*, 1–13. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9537612&casa_token=6VeV8vXBRT0AAAAA:cVhYpdlNbD1BgRH_9GBDQofEVy38quzW6zs3v3doJzJ2Fx2MP02wy0YqLcoAeC8y2GekDshY0bg&tag=1

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 (HICSS)* (pp. 1284–1293). https://doi.org/10.48550/arXiv.2012.01007

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine, 178*(11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 80–89). https://doi.org/10.48550/arXiv.1806.00069

Giudici, P., & Raffinetti, E. (2021). Shapley-Lorenz eXplainable Artificial Intelligence. *Expert Systems with Applications*, *167*(114104). https://doi.org/10.1016/j.eswa.2020.114104

Gonzalez, G. (2018). *How Amazon accidentally invented a sexist hiring algorithm: A company experiment to use artificial intelligence in hiring inadvertently favored male candidates.* https://www.inc.com/guadalupe-gonzalez/amazon-artificial-intelligence-ai-hiring-tool-hr.html

Google (Ed.). (2022). *Explainable AI*. https://cloud.google.com/explainable-ai. Accessed 02 Feb 2023

Granados, N., Gupta, A., & Kauffman, R. J. (2010). Information transparency in business-to-consumer markets: Concepts, framework, and research agenda. *Information Systems Research, 21*(2), 207–226. https://doi.org/10.1287/isre.1090.0249

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly, 23*(4), 497–530. https://doi.org/10.2307/249487

Grisci, B. I., Krause, M. J., & Dorn, M. (2021). Relevance aggregation for neural networks interpretability and knowledge discovery on tabular data. *Information Sciences, 559*, 111–129. https://doi.org/10.1016/j.ins.2021.01.052

Gronau, I., & Moran, S. (2007). Optimal implementations of UPGMA and other common clustering algorithms. *Information Processing Letters, 104*(6), 205–210. https://doi.org/10.1016/j.ipl.2007.07.002

Gu, D., Li, Y., Jiang, F., Wen, Z., Liu, S., Shi, W., Lu, G., & Zhou, C. (2020). VINet: A visually interpretable image diagnosis network. *IEEE Transactions on Multimedia, 22*(7), 1720–1729. https://doi.org/10.1109/TMM.2020.2971170

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42. https://doi.org/10.1145/3236009

Guo, M., Xu, Z., Zhang, Q., Liao, X., & Liu, J. (2021). Deciphering feature effects on decision-making in ordinal regression problems: An explainable ordinal factorization model. *ACM Transactions on Knowledge Discovery from Data (TKDD), 16*(3), 1–26. https://doi.org/10.1145/3487048

Ha, T., Sah, Y. J., Park, Y., & Lee, S. (2022). Examining the effects of power status of an explainable artificial intelligence system on users' perceptions. *Behaviour & Information Technology, 41*(5), 946–958. https://doi.org/10.1080/0144929X.2020.1846789

Hamm, P., Wittmann, H. F., & Klesel, M. (2021). Explain it to me and I will use it: A proposal on the impact of explainable AI on use behavior. *ICIS 2021 Proceedings, 9*, 1–9.

Hardt, M., Chen, X., Cheng, X., Donini, M., Gelman, J., Gollaprolu, S., He, J., Larroy, P., Liu, X., McCarthy, N., Rathi, A., Rees, S., Siva, A., Tsai, E., Vasist, K., Yilmaz, P., Zafar, M. B., Das, S., Haas, K., Hill, T., Kenthapadi, K. (2021). Amazon SageMaker clarify: machine learning bias detection and explainability in the cloud. In *2021 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 2974–2983). https://arxiv.org/pdf/2109.03285.pdf

Hatwell, J., Gaber, M. M., & Atif Azad, R. M. (2020). Ada-WHIPS: Explaining AdaBoost classification with applications in the health sciences. *BMC Medical Informatics and Decision Making, 20*(250), 1–25. https://doi.org/10.1186/s12911-020-01201-2

He, J., Hu, H.-J., Harrison, R., Tai, P. C., & Pan, Y. (2006). Transmembrane segments prediction and understanding using support vector machine and decision tree. *Expert Systems with Applications, 30*, 64–72. https://doi.org/10.1016/j.eswa.2005.09.045

Hepenstal, S., Zhang, L., Kodagoda, N., Wong, B., & l. w. (2021). Developing conversational agents for use in criminal investigations. *ACM Transactions on Interactive Intelligent Systems (TiiS), 11*(3–4), 1–35. https://doi.org/10.1145/3444369

Herse, S., Vitale, J., Tonkin, M., Ebrahimian, D., Ojha, S., Johnston, B., Judge, W., & Williams, M. (2018). Do you trust me, blindly? Factors influencing trust towards a robot recommender system. *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. https://ieeexplore.ieee.org/document/8525581/

Heuillet, A., Couthouis, F., & Díaz-Rodríguez, N. (2021). Explainability in deep reinforcement learning. *Knowledge-Based Systems, 214*, 106685. https://doi.org/10.1016/j.knosys.2020.106685

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75–105. https://doi.org/10.2307/25148625

Hong, S. R., Hullman, J., & Bertini, E. (2020). Human factors in model interpretability: Industry practices, challenges, and

needs. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1, Article 68). https://doi.org/10.1145/3392878

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems, 51*(1), 141–154. https://doi.org/10.1016/j.dss.2010.12.003

Iadarola, G., Martinelli, F., Mercaldo, F., & Santone, A. (2021). Towards an interpretable deep learning model for mobile malware detection and family identification. *Computers & Security, 105*, 1–15. https://doi.org/10.1016/j.cose.2021.102198

IBM (Ed.). (2022). *IBM Watson OpenScale - Overview*. https://www.ibm.com/docs/en/cloud-paks/cp-data/3.5.0?topic=services-watson-openscale

Irarrázaval, M. E., Maldonado, S., Pérez, J., & Vairetti, C. (2021). Telecom traffic pumping analytics via explainable data science. *Decision Support Systems, 150*, 1–14. https://doi.org/10.1016/j.dss.2021.113559

Islam, M. A., Anderson, D. T., Pinar, A., Havens, T. C., Scott, G., & Keller, J. M. (2020). Enabling explainable fusion in deep learning with fuzzy integral neural networks. *IEEE Transactions on Fuzzy Systems, 28*(7), 1291–1300. https://doi.org/10.1109/TFUZZ.2019.2917124

Jakulin, A., Možina, M., Demšar, J., Bratko, I., & Zupan, B. (2005). Nomograms for visualizing support vector machines. *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD)* (pp. 108–117). https://doi.org/10.1145/1081870.1081886

Jiang, J., & Cameron, A.-F. (2020). IT-enabled self-monitoring for chronic disease self-management: An interdisciplinary review. *MIS Quarterly, 44*(1), 451–508. https://doi.org/10.25300/MISQ/2020/15108

Jiang, J., Karran, A. J., Coursaris, C. K., Léger, P. M., & Beringer, J. (2022). A situation awareness perspective on human-AI interaction: Tensions and opportunities. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2022.2093863

Jussupow, E., Meza Martínez, M. A., Mädche, A., & Heinzl, A. (2021). Is this system biased? – How users react to gender bias in an explainable AI System. *Proceedings of the 42nd International Conference on Information Systems (ICIS)* (pp. 1–17). https://aisel.aisnet.org/icis2021/hci_robot/hci_robot/11

Kim, C., Lin, X., Collins, C., Taylor, G. W., & Amer, M. R. (2021). Learn, generate, rank, explain: A case study of visual explanation by generative machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS), 11*(3–4), 1–34.

Kim, B., Park, J., & Suh, J. (2020a). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, *134*(113302). https://doi.org/10.1016/j.dss.2020.113302

Kim, J., Lee, S., Hwang, E., Ryu, K. S., Jeong, H., Lee, J. W., Hwangbo, Y., Choi, K. S., & Cha, H. S. (2020b). Limitations of deep learning attention mechanisms in clinical research: Empirical case study based on the Korean diabetic disease setting. *Journal of Medical Internet Research*, *22*(12). https://doi.org/10.2196/18418

Kliegr, T., Bahník, Š, & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence, 295*, 103458. https://doi.org/10.1016/j.artint.2021.103458

Kline, A., Kline, T., Shakeri Hossein Abad, Z., & Lee, J. (2020). Using item response theory for explainable machine learning in predicting mortality in the intensive care unit: Case-based approach. *Journal of Medical Internet Research*, *22*(9). https://doi.org/10.2196/20268

Knowles, T. (2021). *AI will have a bigger impact than fire, says Google boss Sundar Pichai*. https://www.thetimes.co.uk/article/ai-will-have-a-bigger-impact-than-fire-says-google-boss-sundar-pichai-rk8bdst7r

Kou, Y., & Gui, X. (2020). Mediating community-AI interaction through situated explanation. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2, Article 102). https://doi.org/10.1145/3415173

Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2020). Generating and understanding personalized explanations in hybrid recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS), 10*(4), 1–40.

Kumar, A., Manikandan, R., Kose, U., Gupta, D., & Satapathy, S. C. (2021). Doctor's dilemma: Evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 17*(3s), 1–26.

Kute, D. V., Pradhan, B., Shukla, N., & Alamri, A. (2021). Deep learning and explainable artificial intelligence techniques applied for detecting money laundering – A critical review. *IEEE Access, 9*, 82300–82317.

Kwon, B. C., Choi, M.-J., Kim, J. T., Choi, E., Kim, Y. B., Kwon, S., Sun, J., & Choo, J. (2019). Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1). https://doi.org/10.1109/TVCG.2018.2865027

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Seeing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*. https://doi.org/10.1016/j.artint.2021.103473

Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science*, *9*. https://doi.org/10.28945/479

Li, J., Shi, H., & Hwang, K. S. (2021). An explainable ensemble feedforward method with Gaussian convolutional filter. *Knowledge-Based Systems*, *225*. https://doi.org/10.1016/j.knosys.2021.107103

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–15) https://doi.org/10.1145/3313831.3376590

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Proceedings of the 2009 SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 2119–2128). https://doi.org/10.1145/1518701.1519023

Lopez-Gazpio, I., Maritxalar, M., Gonzalez-Agirre, A., Rigau, G., Uria, L., & Agirre, E. (2017). Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems, 119*, 186–199. https://doi.org/10.1016/j.knosys.2016.12.013

Lukyanenko, R., Castellanos, A., Storey, V. C., Castillo, A., Tremblay, M. C., & Parsons, J. (2020). Superimposition: Augmenting machine learning outputs with conceptual models for explainable AI. In G. Grossmann & S. Ram (Eds.), *Lecture notes in computer science. Advances in conceptual modeling* (pp. 26–34). Springer International Publishing. https://doi.org/10.1007/978-3-030-65847-2_3

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures, 90*, 46–60. https://doi.org/10.1016/j.futures.2017.03.006

Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press.

Marella, V., Upreti, B., Merikivi, J., & Tuunainen, V. K. (2020). Understanding the creation of trust in cryptocurrencies: The case of Bitcoin. *Electronic Markets, 30*(2), 259–271. https://doi.org/10.1007/s12525-019-00392-5

Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. *MIS Quarterly, 38*(1), 73–99. https://doi.org/10.25300/MISQ/2014/38.1.04

Martens, D., Baesens, B., & van Gestel, T. (2009). Decompositional rule extraction from support vector machines by active learning. *IEEE Transactions on Knowledge and Data Engineering, 21*(2), 178–191. https://doi.org/10.1109/TKDE.2008.131

Martens, D., Baesens, B., van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *SSRN Electronic Journal*. Advance online publication.https://doi.org/10.2139/ssrn.878283

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J. & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577* (7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Mehdiyev, N., & Fettke, P. (2020). Prescriptive process analytics with deep learning and explainable artificial intelligence. *Proceedings of the 28th European Conference on Information Systems (ECIS)*. An Online AIS Conference. https://aisel.aisnet.org/ecis2020_rp/122

Mensa, E., Colla, D., Dalmasso, M., Giustini, M., Mamo, C., Pitidis, A., & Radicioni, D. P. (2020). Violence detection explanation via semantic roles embeddings. *BMC Medical Informatics and Decision Making*, *20*(263). https://doi.org/10.1186/s12911-020-01237-4

Merry, M., Riddle, P., & Warren, J. (2021). A mental models approach for defining explainable artificial intelligence. *BMC Medical Informatics and Decision Making, 21*(1), 1–12. https://doi.org/10.1186/s12911-021-01703-7

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2020). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management, 39*(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Meske, C., Abedin, B., Klier, M., & Rabhi, F. (2022). Explainable and responsible artificial intelligence. *Electronic Markets, 32*(4), 2103–2106. https://doi.org/10.1007/s12525-022-00607-2

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

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *ArXiv*. arXiv:1712.00547. https://arxiv.org/pdf/1712.00547.pdf

Ming, Y., Huamin, Qu., & Bertini, E. (2019). RuleMatrix: Visualizing and understanding classifiers with rules. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 342–352. https://doi.org/10.1109/TVCG.2018.2864812

Mirbabaie, M., Brendel, A. B., & Hofeditz, L. (2022). Ethics and AI in information systems research. *Communications of the Association for Information Systems, 50*(1), 38. https://doi.org/10.17705/1CAIS.05034

Mitra, S., & Hayashi, Y. (2000). Neuro-fuzzy rule generation: Survey in soft computing framework. *IEEE Transactions on Neural Networks, 11*(3), 748–768. https://doi.org/10.1109/72.846746

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT)* (pp. 279–288). https://doi.org/10.1145/3287560.3287574

Mombini, H., Tulu, B., Strong, D., Agu, E. O., Lindsay, C., Loretz, L., Pedersen, P., & Dunn, R. (2021). An explainable machine learning model for chronic wound management decisions. *AMCIS 2021 Proceedings, 18*, 1–10.

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal, 73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Moradi, M., & Samwald, M. (2021). Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications*, *165*(113941). https://doi.org/10.1016/j.eswa.2020.113941

Moreira, C., Chou, Y.-L., Velmurugan, M., Ouyang, C., Sindhgatta, R., & Bruza, P. (2021). LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models. *Decision Support Systems, 150*, 1–16. https://doi.org/10.1016/j.dss.2021.113561

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications, 165*, 1–8. https://doi.org/10.1016/j.eswa.2020.113986

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *ArXiv*. https://arxiv.org/pdf/1902.01876

Murray, B. J., Islam, M. A., Pinar, A. J., Anderson, D. T., Scott, G. J., Havens, T. C., & Keller, J. M. (2021). Explainable AI for the Choquet integral. *IEEE Transactions on Emerging Topics in Computational Intelligence, 5*(4), 520–529. https://doi.org/10.1109/TETCI.2020.3005682

Narayanan, M., Chen, E., He, J, Kim, B, Gershman, S., & Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *ArXiv,* 1802.00682. https://doi.org/10.48550/arXiv.1802.00682

Nascita, A., Montieri, A., Aceto, G., Ciuonzo, D., Persico, V., & Pescapé, A. (2021). XAI meets mobile traffic classification: Understanding and improving multimodal deep learning architectures. *IEEE Transactions on Network and Service Management, 18*(4), 4225–4246. https://doi.org/10.1109/TNSM.2021.3098157

Neto, M. P., & Paulovich, F. V. (2021). Explainable matrix - visualization for global and local interpretability of random forest classification ensembles. *IEEE Transactions on Visualization and Computer Graphics, 27*(2), 1427–1437. https://doi.org/10.1109/TVCG.2020.3030354

Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction, 27*(3), 393–444. https://doi.org/10.1007/s11257-017-9195-0

Omeiza, D., Webb, H., Jirotka, M., & Kunze, L. (2021). Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems, 23*(8), 10142–10162. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9616449&casa_token=pCkvj82hzqwAAAAA:yYPZ8qTUP7U8tLQj793sviDzuwLewzQZCvBPza4SHtG_P-eSlpp0Te5X9aF1OuVt35wT6EMfP1w&tag=1

Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., Liu, X., & He, Z. (2020). Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. *Journal of the American Medical*

Informatics Association: JAMIA, 27(7), 1173–1185. https://doi.org/10.1093/jamia/ocaa053

Peñafiel, S., Baloian, N., Sanson, H., & Pino, J. A. (2020). Applying Dempster-Shafer theory for developing a flexible, accurate and interpretable classifier. *Expert Systems with Applications, 148*(113262), 1–12. https://doi.org/10.1016/j.eswa.2020.113262

Pessach, D., Singer, G., Avrahami, D., Chalutz Ben-Gal, H., Shmueli, E., & Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, *134*(113290). https://doi.org/10.1016/j.dss.2020.113290

Pierrard, R., Poli, J.-P., & Hudelot, C. (2021). Spatial relation learning for explainable image classification and annotation in critical applications. *Artificial Intelligence*, *292*(103434). https://doi.org/10.1016/j.artint.2020.103434

Probst, F., Grosswiele, L., & Pfleger, R. (2013). Who will lead and who will follow: Identifying Influential Users in Online Social Networks. *Business & Information Systems Engineering, 5*(3), 179–193. https://doi.org/10.1007/s12599-013-0263-7

Rader, E., & Gray, R. (2015). Understanding user beliefs about algorithmic curation in the Facebook news feed. *Proceedings of the 33rd International Conference on Human Factors in Computing Systems (CHI)* (pp. 173–182). https://doi.org/10.1145/2702123.2702174

Ragab, A., El-Koujok, M., Poulin, B., Amazouz, M., & Yacout, S. (2018). Fault diagnosis in industrial chemical processes using interpretable patterns based on Logical Analysis of Data. *Expert Systems with Applications, 95*, 368–383. https://doi.org/10.1016/j.eswa.2017.11.045

Rana, N. P., Chatterjee, S., Dwivedi, Y. K., & Akter, S. (2022). Understanding dark side of artificial intelligence (AI) integrated business analytics: Assessing firm's operational inefficiency and competitiveness. *European Journal of Information Systems, 31*(3), 364–387. https://doi.org/10.1080/0960085X.2021.1955628

Rawal, A., McCoy, J., Rawat, D., Sadler, B., & Amant, R. (2021). Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives. *IEEE Transactions on Artificial Intelligence, 1*(01), 1–1. https://doi.org/10.1109/TAI.2021.3133846

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? A proposal of user-centered explainable AI. In C. Trattner, D. Parra, & N. Riche (Chairs), *Joint Proceedings of the ACM IUI 2019 Workshops*. http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf

Rissler, R., Nadj, M., Adam, M., & Maedche, A. (2017). Towards an integrative theoretical Framework of IT-Mediated Interruptions. *Proceedings of the 25th European Conference on Information Systems (ECIS)*. http://aisel.aisnet.org/ecis2017_rp/125

Robert, L. P., Bansal, G., & Lütge, C. (2020). ICIS 2019 SIGHCI Workshop Panel Report: Human– computer interaction challenges and opportunities for fair, trustworthy and ethical artificial intelligence. *AIS Transactions on Human-Computer Interaction, 12*(2), 96–108. https://doi.org/10.17705/1thci.00130

Rowe, F. (2014). What literature review is not: Diversity, boundaries and recommendations. *European Journal of Information Systems, 23*(3), 241–255. https://doi.org/10.1057/ejis.2014.7

Russell, S., & Norvig, P. (2021). *Artificial intelligenc: A modern approach (4th)*. Pearson.

Rzepka, C., & Berger, B. (2018). User interaction with AI-enabled systems: A systematic review of IS research. *Proceedings of the Thirty-Nine International Conference on Information Systems (ICIS)*. https://aisel.aisnet.org/icis2018/general/Presentations/7

Sachan, S., Yang, J.-B., Xu, D.-L., Benavides, D. E., & Li, Y. (2020). An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications, 144*(113100), 1–49. https://doi.org/10.1016/j.eswa.2019.113100

Schlicker, N., Langer, M., Ötting, S. K., Baum, K., König, C. J., & Wallach, D. (2021). What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior, 122*, 1–16. https://doi.org/10.1016/j.chb.2021.106837

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*. Advance online publication. https://doi.org/10.1080/12460125.2020.1819094

Schneider, J., & Handali, J. P. (2019). Personalized explanation for machine learning: a conceptualization. *Proceedings of the Twenty-Seventh European Conference on Information Systems (ECIS 2019)*. Stockholm-Uppsala, Sweden. https://arxiv.org/ftp/arxiv/papers/1901/1901.00770.pdf

Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. *Expert Systems with Applications, 41*(5), 2239–2249. https://doi.org/10.1016/j.eswa.2013.09.022

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), 233–242. https://doi.org/10.1093/idpl/ipx022

Sevastjanova, R., Jentner, W., Sperrle, F., Kehlbeck, R., Bernard, J., & El-Assady, M. (2021). QuestionComb: A gamification approach for the visual explanation of linguistic phenomena through interactive labeling. *ACM Transactions on Interactive Intelligent Systems (TiiS), 11*(3–4), 1–38.

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 747–748). https://doi.org/10.1109/DSAA49011.2020.00096

Sharma, P., Mirzan, S. R., Bhandari, A., Pimpley, A., Eswaran, A., Srinivasan, S., & Shao, L. (2020). Evaluating tree explanation methods for anomaly reasoning: A case study of SHAP TreeExplainer and TreeInterpreter. In G. Grossmann & S. Ram (Eds.), *Lecture notes in computer science. Advances in conceptual modeling* (pp. 35–45). Springer International Publishing. https://doi.org/10.1007/978-3-030-65847-2_4

Shen, H., Jin, H., Cabrera, Á. A., Perer, A., Zhu, H., & Hong, J. I. (2020). Designing alternative representations of confusion matrices to support non-expert public understanding of algorithm performance. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), 1–22. https://doi.org/10.1145/3415224

Shin, D. (2021a). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*(102551). https://doi.org/10.1016/j.ijhcs.2020.102551

Shin, D. (2021b). Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm. *Journal of Information Science*, 1–14. https://doi.org/10.1177/0165551520985495

Sidorova, A., Evangelopoulos, N., Valacich, J. S., & Ramakrishnan, T. (2008). Uncovering the intellectual core of the information systems discipline. *MIS Quarterly*, 467–482. https://www.jstor.org/stable/25148852

Singh, N., Singh, P., & Bhagat, D. (2019). A rule extraction approach from support vector machines for diagnosing hypertension among diabetics. *Expert Systems with Applications, 130*, 188–205. https://doi.org/10.1016/j.eswa.2019.04.029

Soares, E., Angelov, P. P., Costa, B., Castro, M. P. G., Nageshrao, S., & Filev, D. (2021). Explaining deep learning models through rule-based approximation and visualization. *IEEE Transactions*

on Fuzzy Systems, 29(8), 2399–2407. https://doi.org/10.1109/TFUZZ.2020.2999776

Spinner, T., Schlegel, U., Schafer, H., & El-Assady, M. (2020). Explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics, 26*(1), 1064–1074. https://doi.org/10.1109/TVCG.2019.2934629

Springer, A., & Whittaker, S. (2020). Progressive disclosure: When, why, and how do users want algorithmic transparency information? *ACM Transactions on Interactive Intelligent Systems (TiiS), 10*(4), 1–32. https://doi.org/10.1145/3374218

Stoean, R., & Stoean, C. (2013). Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. *Expert Systems with Applications, 40*, 2677–2686. https://doi.org/10.1016/j.eswa.2012.11.007

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems, 41*, 647–665. https://doi.org/10.1007/s10115-013-0679-x

Su, G., Lin, B., Luo, W., Yin, J., Deng, S., Gao, H., & Xu, R. (2021). Hypomimia recognition in Parkinson's disease with semantic features. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *17*(3), 1–20. https://doi.org/10.1145/3476778

Sultana, T., & Nemati, H. (2021). Impact of explainable AI and task complexity on human-machine symbiosis. *Proceedings of the Twenty-Seventh Americas Conference on Information Systems (AMCIS)*. https://aisel.aisnet.org/amcis2021/sig_hci/sig_hci/20

Sun, C., Dui, H., & Li, H. (2021). Interpretable time-aware and co-occurrence-aware network for medical prediction. *BMC Medical Informatics and Decision Making, 21*(1), 1–12.

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication, 25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Tabankov, S. S., & Möhlmann, M. (2021). Artificial intelligence for in-flight services: How the Lufthansa group managed explainability and accuracy concerns. *Proceedings of the International Conference on Information Systems (ICIS), 16*, 1–9.

Taha, I. A., & Ghosh, J. (1999). Symbolic interpretation of artificial neural networks. *IEEE Transactions on Knowledge and Data Engineering, 11*(3), 448–463. https://doi.org/10.1109/69.774103

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

van der Waa, J., Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2020). Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*, *144*(102493). https://doi.org/10.1016/j.ijhcs.2020.102493

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: A systematic review. *ArXiv*. https://arxiv.org/pdf/2006.00093

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*(103404). https://doi.org/10.1016/j.artint.2020.103404

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems, 25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

vom Brocke, J., Simons, A., Niehaves, B [Bjoern], Niehaves, B [Bjorn], Reimer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. *ECIS 2009 Proceedings*(161). http://aisel.aisnet.org/ecis2009/161

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. *Proceedings of the 2019 Conference on Human Factors in Computing Systems (CHI)*. http://dl.acm.org/citation.cfm?doid=3290605.3300831

Wanner, J., Heinrich, K., Janiesch, C., & Zschech, P. (2020a). How much AI do you require decision factors for adopting AI technology. *Proceedings of the Forty-First International Conference on Information Systems (ICIS)*. https://aisel.aisnet.org/icis2020/implement_adopt/implement_adopt/10

Wanner, J., Herm, L. V., & Janiesch, C. (2020b). How much is the black box? The value of explainability in machine learning models. *ECIS 2020 Research-in-Progress*. https://aisel.aisnet.org/ecis2020_rip/85

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly, 26*(2), xiii–xxiii.

Xiong, J., Qureshi, S., & Najjar, L. (2014). A cluster analysis of research in information technology for global development: Where to from here? *Proceedings of the SIG GlobDev Seventh Annual Workshop*. https://aisel.aisnet.org/globdev2014/1

Yampolskiy, R. V. (2019). Predicting future AI failures from historic examples. *Foresight, 21*(1), 138–152. https://doi.org/10.1108/FS-04-2018-0034

Yan, A., & Xu, D. (2021). AI for depression treatment: Addressing the paradox of privacy and trust with empathy, accountability, and explainability. *Proceedings of the Fourty-Second International Conference on Information Systems (ICIS)*. https://aisel.aisnet.org/icis2021/is_health/is_health/15/

Yang, Z., Zhang, A., & Sudjianto, A. (2021). Enhancing explainability of neural networks through architecture constraints. *IEEE Transactions on Neural Networks and Learning Systems, 32*(6), 2610–2621. https://doi.org/10.1109/TNNLS.2020.3007259

Yoo, S., & Kang, N. (2021). Explainable artificial intelligence for manufacturing cost estimation and machining feature visualization. *Expert Systems with Applications, 183*, 1–14. https://doi.org/10.1016/j.eswa.2021.115430

Zeltner, D., Schmid, B., Csiszár, G., & Csiszár, O. (2021). Squashing activation unctions in benchmark tests: Towards a more eXplainable Artificial Intelligence using continuous-valued logic. *Knowledge-Based Systems*, *218*. https://doi.org/10.1016/j.knosys.2021.106779

Zhang, Q. S., & Zhu, S. C. (2018). Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering, 19*(1), 27–39. https://doi.org/10.1631/FITEE.1700808

Zhang, K., Liu, X., Liu, F., He, L., Zhang, L., Yang, Y., Li, W., Wang, S., Liu, L., Liu, Z., Wu, X., & Lin, H. (2018). An interpretable and expandable deep learning diagnostic system for multiple ocular diseases: Qualitative study. *Journal of Medical Internet Research, 20*(11), 1–13. https://doi.org/10.2196/11144

Zhang, C. A., Cho, S., & Vasarhelyi, M. (2022). Explainable Artificial Intelligence (XAI) in auditing. *International Journal of Accounting Information Systems, 46*, 100572. https://doi.org/10.1016/j.accinf.2022.100572

Zhao, X., Wu, Y., Lee, D. L., & Cui, W. (2019). Iforest: Interpreting random forests via visual analytics. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 407–416. https://doi.org/10.1109/TVCG.2018.2864475

Zhdanov, D., Bhattacharjee, S., & Bragin, M. (2021). Incorporating FAT and privacy aware AI modeling approaches into business

decision making frameworks. *Decision Support Systems, 155*, 1–12. https://doi.org/10.1016/j.dss.2021.113715

Zhong, Q., Fan, X., Luo, X., & Toni, F. (2019). An explainable multi-attribute decision model based on argumentation. *Expert Systems with Applications, 117*, 42–61. https://doi.org/10.1016/j.eswa.2018.09.038

Zhu, C., Chen, Z., Zhao, R., Wang, J., & Yan, R. (2021). Decoupled feature-temporal CNN: Explaining deep learning-based machine health monitoring. *IEEE Transactions on Instrumentation and Measurement, 70*, 1–13. https://doi.org/10.1109/TIM.2021.3084310

Zytek, A., Liu, D., Vaithianathan, R., & Veeramachaneni, K. (2021). Sibyl: Explaining machine learning models for high-stakes decision making. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI)* (pp. 1–6). https://doi.org/10.1145/3411763.3451743