Information Systems Management

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uism20

Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities

Christian Meske , Enrico Bunde , Johannes Schneider & Martin Gersch

To cite this article: Christian Meske , Enrico Bunde , Johannes Schneider & Martin Gersch (2020): Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities, Information Systems Management

To link to this article: https://doi.org/10.1080/10580530.2020.1849465

Published online: 08 Dec 2020.

INFORMATION SYSTEMS MANAGEMENT
https://doi.org/10.1080/10580530.2020.1849465

RESEARCH NOTE

Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities

Christian Meske[a], Enrico Bunde[a], Johannes Schneider[b], and Martin Gersch[c]

[a]Department of Information Systems, Freie Universität Berlin and Einstein Center Digital Future, Berlin, Germany; [b]Institute of Information Systems, University of Liechtenstein, Vaduz, Liechtenstein; [c]Department of Information Systems, Freie Universität Berlin, Einstein Center Digital Future as Well as Digital Entrepreneurship Hub, Berlin, Germany

ABSTRACT

Artificial Intelligence (AI) has diffused into many areas of our private and professional life. In this research note, we describe exemplary risks of black-box AI, the consequent need for explainability, and previous research on Explainable AI (XAI) in information systems research. Moreover, we discuss the origin of the term XAI, generalized XAI objectives, and stakeholder groups, as well as quality criteria of personalized explanations. We conclude with an outlook to future research on XAI.

KEYWORDS

Artificial Intelligence; explainability; accountability; transparency; trust; managing AI

Introduction

Artificial Intelligence (AI), a research area initiated in the 1950ies (Mccarthy et al., 2006), has received significant attention in science and practice. Global spending on AI systems is expected to more than double from 38 billion USD in 2019 to 98 billion USD by 2023 (Shirer & Daquila, 2019). Emphasizing on machine learning, and thereby connecting to what is meant by "intelligent", AI can be defined, for instance, as the "system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019, p. 15).

In combination with increasing IT-processing capabilities, especially machine learning approaches including artificial neural networks have led to a task performance of AI that has never been seen before. Hence, advanced technologies of today make increasingly use of 'bio-inspired paradigms' in order to effectively tackle complex real-world problems (Zolbanin et al., 2019). We still speak of such systems as "weak AI" or "narrow AI" – since they are only used for very specific tasks and, in contrast to "strong AI", are not universally applicable (Searle, 1980; Watson, 2017). However, today's algorithms already reached or even surpassed the task performance of humans in different domains. For example, corresponding applications outperformed professional human players in complex games such as Go and Poker (Blair & Saffidine, 2019; Silver et al., 2017) or proved to be more accurate in breast cancer detection (McKinney et al., 2020). In consequence, these advances in socio-technical systems will significantly affect the future of work (Dewey & Wilkens, 2019; Elbanna et al., 2020).

AI is thus increasingly applied in use cases with potentially severe consequences for humans. This holds true not only in medical diagnostics, but also in processes of job recruitment (Dastin, 2018), credit scoring (Wang et al., 2019), prediction of recidivism in drug courts (Zolbanin et al., 2019), or as autopilots in aviation (Garlick, 2017) and autonomous driving (Grigorescu et al., 2020). Furthermore, corresponding technology is more and more integrated into our everyday private lives in the form of intelligent agents like Google Home or Siri (Bruun & Duka, 2018). However, due to the growing complexity of underlying models and algorithms, AI appears as a "black box", because the internal learning processes as well as the resulting models are not completely comprehensible. This trade-off between performance and explainability can have a significant impact on individual beings, businesses, and society as a whole (Alt, 2018).

Research on information systems, so we argue, needs to respond to this challenge by fostering research on Explainable Artificial Intelligence (XAI), which to date has been mostly investigated with a method-oriented focus for developers in computer science. Yet, explainability is a prerequisite for fair, accountable, and trustworthy AI (Abdul et al., 2018; Fernandez et al., 2019; Miller, 2019), eventually affecting how we manage, use, and interact with it. For instance, the absence of

CONTACT Christian Meske christian.meske@fu-berlin.de Department of Information Systems, Freie Universität Berlin and Einstein Center Digital Future, Berlin 14195, Germany.

explainability implies that humans cannot conduct a risk or threat analysis, increasing the probability of undesirable behavior of the system. Further, our community's "collective research efforts should advance human welfare" (Malhotra et al., 2013, p. 1270), which may be jeopardized by such non-explainable and hence possibly uncontrollable AI. Also, as future automation and decision support systems will be increasingly based on complex algorithms, information systems may use machine learning more often as an additional method for scientific research.

In this *research note*, we will first discuss exemplary risks and the "dark side" of AI in Section 2, followed by a short overview of previous research on explainability in information systems in Section 3. In Section 4, we outline the terminology and origin as well as objectives and stakeholders of XAI, and list quality criteria of personalized explanations. In Section 5, we provide future research opportunities for behavioral as well as design science researchers, followed by a conclusion in Section 6.

Risks and dark sides of AI usage

Different risks exist regarding the use of AI systems. A major potential problem is "bias", which comes in different facets. In certain situations, humans have a tendency to over-rely on automated decision-making, called "automation bias", which can result in a potential failure to recognize errors in the black box (Goddard et al., 2012). As an example, medical doctors ignored their own diagnoses, even when they were correct, because their diagnosis was not recommended by the AI system (Friedman et al., 1999; Goddard et al., 2011). Furthermore, automation bias can foster the process of "deskilling", either because of the attrition of existing skills or due to the lack of skill development in general (Arnold & Sutton, 1998; Sutton et al., 2018). Such problems highlight the overall risk of inappropriate trust of humans toward AI (Herse et al., 2018).

Not only humans can have a bias but also the AI system itself. For instance, such systems can intentionally or unintentionally be biased toward wrongful output. Caliskan et al. (2017) point out, how text and web corpora in training data can contain human bias, leading to a machine learning model that is biased against race or gender, consequently establishing AI-based discrimination, racism, or sexism. Bias in the "real" world, and consequently in historical data, may therefore lead to statistical bias, which again can perpetuate bias in the real world (Parikh et al., 2019). For example, as shown in a recent review, Apple's face recognition systems failed to distinguish Asian users, Google's sentiment analyzer got homophobic and anti-Semitic, a predictive policing system disproportionately targeted minority neighborhoods, and a bot designed to converse with users on Twitter became verbally abusive (Yampolskiy, 2019, pp. 141–142). Moreover, AI may learn correlations that are not linked to causal relations in the real world (Lapuschkin et al., 2019). In Figure 1, such a classifier is depicted that learned to focus on a source tag, which was found for about 20% of images of horses in the training data. When the source tag was removed, the classifications changed accordingly. Hence, when the same source tag was implemented on an image of a car, the AI still classified it as a horse.

In another case, a machine learning model used the presence of a ruler on images for diagnosis of malignant skin tumors (Narla et al., 2018). The reason was that dermatologists tend to only mark lesions with a ruler that are a cause for concern to them, hence introducing bias to the training data set.

In addition, there is a "dark side" of AI based on misuse (Schneider et al., 2020; Xiao et al., 2020). We leave a digital footprint everywhere (Vidgen et al., 2017),

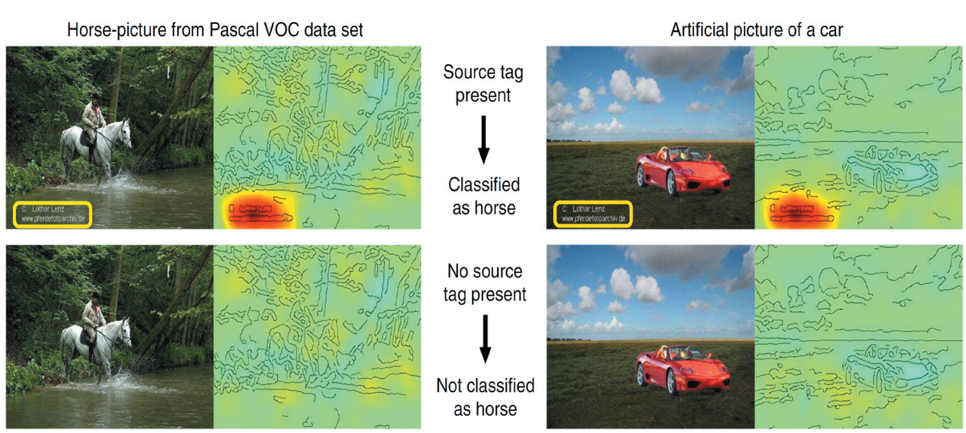

Figure 1. Explanations for AI-based classifications using Grad-CAM (Lapuschkin et al., 2019, p. 3).

through, for instance, online shopping, social media conversations, or usage of mobile navigation apps. While such data deluge has led to the proliferation of data analytics and AI for economic and business potential (Mikalef et al., 2020), it may also lead to a significant power imbalance and unwanted authority of private businesses (Zuboff, 2015) or public institutions alike (Brundage et al., 2018). Moreover, in so-called manipulative "adversarial attacks" only few pixels of an image need to be modified, which yet lead machine learning models to predict completely different classes (Su et al., 2019).

These exemplary risks highlight the need for explainable AI and control. In the following section, we will now provide an overview of how explainability has been investigated in information systems so far.

Explainability in information systems research

Investigating explainability is not completely new to the information systems community. With the rise of systems termed knowledge-based systems, expert systems or intelligent agents in the 1980ies and 1990ies, information systems research started to investigate the necessity for explanations to learn *about* and *from* the artifacts' reasoning. For instance, scholars discussed the potential impact of explanations on users' improved understanding about the system, consequently influencing the effectiveness and efficiency of judgmental decision making, as well as on the perception of the system's usefulness, ease of use, satisfaction, and trust (Dhaliwal & Benbasat, 1996; Mao & Benbasat, 2000; Ye & Johnson, 1995). It was found that novices had a higher and different need for explanations than experts, and that justifications of the system's actions or recommendations (*why*) are more requested than rule-oriented explanations of *how* the system reasoned (Mao & Benbasat, 2000; Ye & Johnson, 1995).

Combining a cognitive effort perspective with cognitive learning theory and Toulmin's model of argumentation, further work emphasized on a detailed classification of explanations: Type I, trace or line of reasoning (which explain why certain decisions were or were not made), type II, justification or support (which justify the reasoning process by linking it to the "deep knowledge" from which it was derived), type III, control or strategic (which explain the system's control behavior and problem-solving strategy), and type IV, terminological (which supply definitional or terminological information) (Gregor & Benbasat, 1999, based on Chandrasekaran et al., 1989; Swartout & Smoliar, 1987). Explanations should be understandable for the user and easy to obtain, e.g., automatically, if this can be done unobtrusively. They should also be context-specific rather than generic (Gregor & Benbasat, 1999).

Subsequent work analyzed how natural language reports based on variable comparisons, which explain why a system suggests certain strategic decisions in situations of nuclear emergencies, help to evaluate the overall decision support system (Papamichail & French, 2005). It was furthermore shown, that long explanations with a conveyed strong confidence level and higher information value lead to an increased acceptance of interval forecasts compared to short explanations and conveyed weak confidence level with low information value (Gönül et al., 2006). Arnold et al. (2006) showed that users were more likely to adhere to recommendations of the KBS when an explanation facility was available, while choice patterns indicated that novices used feedforward explanations more than experts did, while experts mostly used feedback explanations. Further studies in the area of decision support systems indicate that tools, which have enhanced explanatory facilities and provide justifications at the end of the consultation process, lead to improved decision-process satisfaction and decision-advice transparency, subsequently leading to empowering effects like a higher sense of control and a lower perceived power distance (Li & Gregor, 2011). The authors also showed that personalization of explanations with a focus on a cognitive fit can increase the perceived explanation quality and hence explanation influence as well as perceived usefulness of the system (Li & Gregor, 2011).

Aforementioned systems, such as knowledge-based or expert systems, are referred to as symbolic AI, or Good Old Fashioned AI (GOFAI), since human knowledge was instructed through rules in a declarative form (Haugeland, 1985). With the turn of the millennium and discussions of "new-paradigm intelligent systems" (Gregor & Yu, 2002) like artificial neural networks, it was recognized, that the latter are typically neither capable to inherently declare the knowledge they contain, nor to explain the reasoning processes they go through. In that context, it was argued, that explanations could be obtained indirectly, e.g., through sensitivity analysis (Rahman et al., 1999), which derives conclusions from output variations caused by small changes of a particular input (Gregor & Yu, 2002). Besides only very few examples (e.g. Eiras-Franco et al., 2019; Giboney et al., 2015; Martens & Provost, 2014), since then most of the publications[1] on explainability of AI systems, or "Explainable Artificial Intelligence" (XAI), have been published outside of the information systems community, mostly in computer science. As one can see, the existing IS literature is very valuable but with its peak in the 1990ies and early 2000s also comparatively dated, which motivates our call for more IS research on the explainability of AI.

For a better understanding, in the following section we will first discuss the term XAI and its origin, XAI objectives, and stakeholders, as well as quality criteria of personalized explanations.

Explainable artificial intelligence

Terminology

Symbolic AI such as MYCIN, an expert system to diagnose and recommend treatment for bacteria-related infections in the 1970s (Fagan et al., 1980), was already able to explain its reasoning for diagnostic or instructional purposes. However, to the best of our knowledge, it took until 2002, when the term "Explainable Artificial Intelligence" was mentioned the first time as a side-note in a review of "Full Spectrum Command" (FSC, Brewster, 2002), a PC-based military simulation of tactical decision making. In this review of a preliminary beta version of FSC, which was still a GOFAI knowledge-based system, XAI referred to the feature that it "can tell the student exactly what it did and why" (Brewster, 2002, p. 8), consequently augmenting the instructor-facilitated after-action review. Two years later, FSC was presented by their developers in an article at the computer science conference on Innovative Applications of Artificial Intelligence, in which FSC was described as an "XAI System" for small-unit tactical behavior (Van Lent et al., 2004). In this paper, XAI systems were officially introduced and defined as systems that "present the user with an easily understood chain of reasoning from the user's order, through the system's knowledge and inference, to the resulting behavior" (Van Lent et al., 2004, p. 900).

A more current, machine learning-related and often-cited definition of XAI reads as follows: XAI aims to "produce explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately, trust, and effectively manage the emerging generation of artificially intelligent partners" (Gunning, 2017). However, there is no generally accepted definition for that term. It rather refers to "the movement, initiatives, and efforts made in response to AI transparency and trust concerns, *more than to a formal technical* concept" (Adadi & Berrada, 2018, p. 52140).

In literature, the terms *explainability* and *interpretability* are often used synonymously. One way to describe potential differences is the following: if humans can directly make sense of a machine's reasoning and actions without additional explanations, we speak of interpretable machine learning or interpretable AI (Guidotti et al., 2018). Interpretability may therefore be seen as a passive characteristic of the artifact (Rudin, 2019). However, if humans need explanations as a proxy to understand the system's learning and reasoning processes, for example, because an artificial neural network is too complex, we speak of research on explainable AI (Adadi & Berrada, 2018).

In computer science, in which most of the research on XAI has been taking place, different instruments to explain an AI's inner working have been developed and categorized (Ras et al., 2018). Some of these methods allow to interpret a single prediction of a machine learning model, others allow to understand the whole model, leading to the differentiation between "local" and "global" explanations. The explanation output can be presented in the form of "feature attribution" (pointing out how data features supported or opposed a model's prediction, see also Figure 1 back in Section 2), "examples" (returning data instances as examples to explain the model's behavior), "model internals" (returning the model's internal representations, e.g., of the model's neurons) and "surrogate models" (returning an intrinsically interpretable, transparent model which approximates the target black-box model). Some XAI methods can be used for any machine learning model ("model-agnostic explanations"), others work only for e.g., neural networks ("model-specific explanations"). Certain XAI methods just work with textual input data, others only with tabular, visual, or audio data, and again others work with multiple inputs. For a detailed technical overview and categorization of existing XAI methods we refer to extensive surveys such as (Gilpin et al., 2018; Guidotti et al., 2018; Ras et al., 2018).

Objectives and stakeholders of explainable artificial intelligence

First, as our section on AI risks and failures highlights, it is important to build a sufficient understanding about the system's behavior to detect unknown vulnerabilities and flaws, for example, in order to avoid phenomena related to spurious correlations. As for that, so we argue, explainability is crucial for the human ability to *evaluate* the system (see Figure 2).

Second, especially from a developer's design perspective, understanding the inner workings of AI and consequent outcomes is vital to enhance the algorithm. Explainability can therefore support to increase the

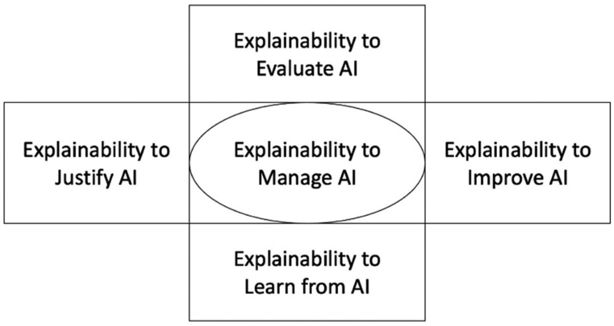

Figure 2. Generalized objectives of explainable artificial intelligence.

system's accuracy and value. Hence, *improvement* is an additional goal that can be achieved with the application of XAI methods (Gilpin et al., 2018).

Third, referring back to our discussion of knowledge-based systems, certain types of explanations provide information on why (or based on which knowledge) certain rules were programmed into the system, which represented "deep knowledge" (Chandrasekaran et al., 1989; Gregor & Benbasat, 1999). While there is no corresponding programmed knowledge in machine learning models, AI explanations could be used, for instance, to discover unknown correlations with causal relationships in data. We thus call it the goal of XAI to *learn* from the algorithm's working and results in order to gain deep knowledge.

Fourth, AI is increasingly used in critical situations which have potentially severe consequences for humans. Whether legislation, such as the General Data Protection Regulation (GDPR) in Europe, established a formal "right for explanation" (Goodman & Flaxman, 2017) is debatable, however, they are usually clear on the demand for accountability and transparency in automated decision processes, which lead to potential consequences that significantly affect the individual (European Union, 2016). Hence, to *justify*, as Adadi and Berrada (2018) call it, is an important goal of XAI.

Fifth, with a focus on implementation and usage, AI adds a level of novelty and complexity that goes beyond traditional IT and data applications, inserting new forms of material agency into organizational processes, potentially changing how work routines emerge and outcomes from work are produced (Berente et al., 2019; Rai et al., 2019). We hence argue that for tackling these challenges, we need explainability to evaluate, to improve, to learn, and to justify in order to achieve the overarching goal of to *manage* AI. Figure 2 summarizes the generalized objectives.

The generalized objectives of XAI manifest differently for various stakeholder groups. For instance, AI *Developers* focus on improving the algorithm's performance as well as on debugging and verification in order to pursue a structured engineering approach based on cause analysis instead of trial and error (Hohman et al., 2019). As such systems are increasingly used in critical situations, and depending on corresponding legislative circumstances, it may need certification. In consequence, there are AI *Regulators*, who need explanations in order to being able to test and certify the system.

In an organizational context, there are "*AI Managers*" who, for example, need explanations to supervise and control the algorithm, its usage and assure its compliance. Those who apply a given system, called "*AI Users*", are rather interested in explainability features to understand and compare the artifact's reasoning with his or her own reasoning, in order to analyze its validity and reliability, or to determine influential factors for a specific prediction (e.g., doctors). Eventually, so we argue, there are *Individuals affected by AI-based decisions* (e.g., patients) caused by AI users or even by autonomous ruling, who may have an interest in explainability to evaluate the fairness of a given AI-based decision. The following Figure 3 provides an overview of potential stakeholder groups and their exemplary interests in explainability of AI.

Members between different and within the same stakeholder groups can have varying backgrounds regarding training, experience, and demographic characteristics. This can lead to different needs for AI explanations as well as their perceptions as, e.g., being useful. Thus, based on personal traits and in combination with their task-related interest in transparency, explanations need to be personalized (Kühl et al., 2019; Schneider & Handali, 2019). Corresponding quality criteria of personalized explanations will be described in the following section.

Quality criteria of personalized explanations

There are different factors that determine the quality of explanations, which in addition can be perceived differently by the various XAI stakeholder groups. As described in Section 3, explanations should, amongst others, be understandable for the individual user, easy to get, context-specific rather than generic, with a conveyed strong confidence level and high information value, and personalized to the explainee (Gönül et al., 2006; Gregor & Benbasat, 1999; Li & Gregor, 2011). In the following, we provide a list of overarching quality criteria for personalized explanations based on and extended from (Schneider & Handali, 2019).

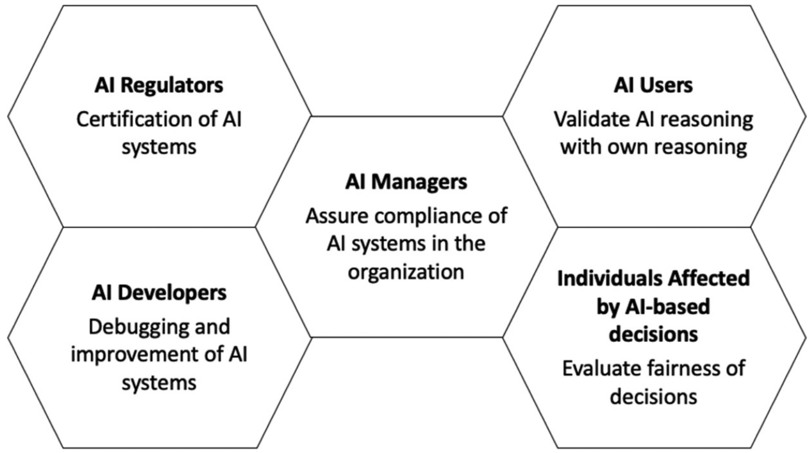

Figure 3. Stakeholder groups of explainable artificial intelligence.

Fidelity describes, to which extend a black-box accurately matches the input-output mapping of a given model (Guidotti et al., 2018; Ras et al., 2018). *Generalizability* refers to the range of models which the XAI technique can explain or be applied to, whereby a high generalizability increases the usefulness of the explanation technique (Ras et al., 2018). *Explanatory power* refers to the scope of questions that can be answered: explanations that allow to understand the general model behavior have more explanatory power compared to explanation of specific predictions only (Ras et al., 2018; Ribeiro et al., 2016). *Interpretability* describes to which extend an explanation is understandable for humans (Guidotti et al., 2018).

Comprehensibility refers to the capacity of an explanation to aid a human user in performing a task, while *plausibility* can be understood as a measure regarding the acceptance of the explanatory content (Fürnkranz et al., 2020). *Effort* addresses the (ideally few) resources needed in order to understand or interpret an explanation (Schneider & Handali, 2019). *Privacy* should prevent the risk that (meta)data, for instance, in the course of XAI personalization, can be used to draw conclusions about the person or its behavior (Radaelli et al., 2015). *Fairness* refers to the goal that explanations should be egalitarian, e.g., in terms of the quality presented to different groups of explainees (Binns, 2018; Kusner et al., 2017). Figure 4 summarizes the quality criteria for personalized explanations.

Findings from the social sciences can help to tailor the design of XAI more precisely to the requirements of the various stakeholders, for example, individually accepted indicators of trustworthiness for services with predominant credence qualities (Böhmann et al., 2014; Kasnakoglu, 2016; Lynch & Schuler, 1990; Matzner et al., 2018; Wood & Schulman, 2019).

Figure 4. Quality criteria for personalized explanations.

Further research opportunities

Explainability is described as being as old as the topic of AI itself rather than being a problem that arises through AI (Holzinger et al. 2019). In the early days of AI research, the models often consisted of reasoning methods, which were logical and symbolic, resulting in limited performance, scalability, and applicability. However, such kind of AI systems delivered a basis for explanations as they performed some sort of logical inference on symbols that were readable for humans. In contrast, the AI systems of today are more complex why explainability is more challenging. Hence, research on XAI and computer-aided verification "needs to keep pace with applied AI research in order to close the research gaps that could hinder operational deployment." (Kistan et al., 2018, p. 1). We argue that this does not only

Table 1. Summary of potential research opportunities and contributions.

Research stream	Research question	Research contribution
Behavioral Science	How do AI explanations influence the users' and mangers' cognitive perception of the AI?	Knowledge about how explainability may be an important variable in existing theories about human perception of the world and IS artifacts (e.g., affordance theory, mental model theory, sensemaking, UTAUT, and others).
	How do explanations influence employees' compliance behavior and work practices?	Knowledge on how AI explanations support IT governance.
	How do explanations help to detect bias in managerial decision making?	Knowledge on how a higher degree of AI transparency leads to a better understanding of potentially undesired practices in the organizational offline world, which found their way into the data sets (e.g., when it comes to racial or gender bias).
	Under which circumstances do explanations support or inhibit individual's trust toward the AI?	Knowledge on how different levels of expertise and personality traits like risk aversion elicit different reactions to AI explanations.
	How can explanations fulfill task-related needs of the different XAI stakeholders?	Knowledge on when and how explanations should be presented to users in order to increase task performance.
	What are adequate metrics to evaluate AI explanations?	Knowledge on the dimensions that are relevant for explanations to be effective; differentiate "good" from "bad" explanations.
	How do explanations influence (de)skilling of employees?	Knowledge on how explanations help to maintain or increase user qualification and self-efficacy regarding AI usage.
Design Science	How can the technical advancements of computer science (e.g., XAI instruments) be integrated with advancements of information systems (e.g., theorizing and categorization of explanations)?	Bring together knowledge and methodical expertise of different disciplines in order to accelerate and improve XAI research across research communities.
	Which features in explanations support the evaluation of an AI's ethicality and morality?	Derive an understanding of how an AI's state of ethicality and morality can be evaluated and which information need to be provided via explanations.
	How can the transdisciplinary design of AI explainability across different stakeholders look like?	Conceptualization of a standardized design process for fair, accountable and transparent AI, that take the needs of different stakeholders into account.
	What are design principles on how to build explainable AI systems that allow for a stakeholder- and domain-specific personalization?	Knowledge of technical possibilities to allow for a flexible adaptation of explanations by users (based on their task-specific needs and level of expertise).
	How should mechanisms of push and pull information through explanations look like?	Knowledge on when the system needs to push information on its reasoning or emerging risks, and how the user can be enabled to individually pull explanations (which includes different regulatory needs for explainability of AI according to its criticality).
	How can the analysis of XAI feature usage help to improve the design and hence quality of AI explanations?	Knowledge on how the manual or automatic analysis of AI usage data improve the understanding of the users' information needs and hence AI explanations.
	How should explanation interfaces in the context of interactive machine learning be designed, in order to improve the AI system based on a users' feedback to its reasoning?	Improving our understanding on the role of explanations in the context of Human-in-the-loop (HITL) interactions between users and AI.

refer to the development of new XAI methods but also requires a socio-technical perspective. There are hence various opportunities for further investigations on the topic of explainability in information systems, of which we outline examples in Table 1.

Conclusion

AI has diffused into many areas of our private and professional life. It hence influences how we live and work. Moreover, it is increasingly used in critical situations with potentially severe consequences for individual human beings, businesses, and the society as a whole. In consequence, new ethical questions arise that challenge necessary compromises between an open development of AI-based innovations and regulations based on societal consensus (EU Commission, 2019; Jobin et al., 2019). Research on explainability, so we argue, is an important factor to support such compromises. In the last 70 years, there have been several AI "summers" (Grudin, 2019). As our brief review on explainability in information systems highlights, there has also been an "explainability summer" in the 1990ies and an "explainability winter" since the dawn of the new millennium. At the moment, witnessing another raise of attention for AI, we therefore call for a second summer of explainability research in information systems. In summary, it can be concluded that XAI is a central issue for information systems research, which opens up a multitude of interesting but also challenging questions to investigate.

Note

1. We acknowledge that there have been recent XAI publications on IS conferences. However, in this section, we only focus on articles in IS journals.

Notes on contributors

Christian Meske is Assistant Professor at the Department of Information Systems, Freie Universität Berlin, and board member of the Einstein Center Digital Future (Berlin), Germany. His research on digital transformation and collaboration has been published in journals such as Business & Information Systems Engineering, Business Process Management Journal, Communications of the Association for Information Systems, Information Systems Frontiers, Information Systems Management, Journal of Enterprise Information Management, or Journal of the Association for Information Science and Technology. Amongst others, he has been recognized with the AIS Best Information Systems Publication of the Year Award and ICIS Paper-a-Thon Award.

Enrico Bunde is a research assistant and PhD at the Department of Information Systems at Freie Universität Berlin (Germany). There he is member of the research group "Digital Transformation and Strategic Information Management", and focuses on explainable artificial intelligence and decision support systems. His work has been published or accepted for publication at conferences such as the International Conference on Information Systems, Hawaii International Conference on System Sciences, or International Conference on Artificial Intelligence in Human-Computer Interaction.

Johannes Schneider holds a tenure-track position as Assistant Professor in Data Science at the Institute of Information Systems at the University of Liechtenstein. His research has been published within and outside the Information Systems community, including journals such as Journal of the ACM, the ACM Transactions on Knowledge Discovery from Data, IEEE Transactions on Software Engineering, the journal of Theoretical Computer Science and the International Journal of Information Management.

Martin Gersch is a full professor of Business Administration, Information and Organization at the School of Business & Economics of Freie Universität Berlin (Germany) and there one of the founders of the Department of Information Systems. He serves also, amongst others, as Principle Investigator on Digital Transformation at the Einstein Center Digital Future and as a mentor for startup teams in the digital economy. His research has been published in journals such as Journal of Management Studies, Electronic Markets, Business & Information Systems Engineering, Journal of Business Process Management, Organization Studies and interdisciplinary also e.g., in Journal as BMC Health Services Research or Journal of the Intensive Care Society.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems. *Proceedings of the 2018 CHI conference on human factors in computing systems - CHI '18*, Montréal, Canada,1–18. https://doi.org/10.1145/3173574.3174156

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

Alt, R. (2018). Electronic markets and current general research. *Electronic Markets*, *28*(2), 123–128. https://doi.org/10.1007/s12525-018-0299-0

Arnold, V., Clark, N., Collier, P. A., Leech, S. A., & Sutton, S. G. (2006). The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *MIS Quarterly: Management Information Systems*, *30*(1), 79. https://doi.org/10.2307/25148718

Arnold, V., & Sutton, S. G. (1998). The theory of technology dominance: Understanding the impact of intelligent decision maker´s judgments. *Advances in Accounting Behavioral Research*, *1*(3), 175–194.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2019). Call for papers MISQ special issue on managing AI. *MIS Quarterly*, 1–5. https://misq.org/skin/frontend/default/misq/pdf/CurrentCalls/ManagingAI.pdf

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Maschine Learning Research*, *81*, 1–11. https://arxiv.org/abs/1712.03586

Blair, A., & Saffidine, A. (2019). AI surpasses humans at six-player poker. *Science*, *365*(6456), 864–865. https://doi.org/10.1126/science.aay7774

Böhmann, T., Leimeister, J. M., & Möslein, K. (2014). Service-systems-engineering. *WIRTSCHAFTSINFORMATIK*, *56*(2), 83–90. https://doi.org/10.1007/s11576-014-0406-6

Brewster, F. W., II. (2002). Using tactical decision exercises to study tactics. *Military Review*, *82*(6), 3–9. https://www.semanticscholar.org/paper/Using-Tactical-Decision-Exercises-to-Study-Tactics-Brewster/acc2892aa434e4a743e7638f769b06be0ee0639d

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., & Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. 1–101.

Bruun, E. P. G., & Duka, A. (2018). Artificial intelligence, jobs and the future of work: Racing with the machines. *Basic Income Studies*, *13*(2), 1–15. https://doi.org/10.1515/bis-2018-0018

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Chandrasekaran, B., Tanner, M. C., & Josephson, J. R. (1989). Explaining control strategies in problem solving. *IEEE Expert*, *4*(1), 9–15. https://doi.org/10.1109/64.21896

European Comission, High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. Retrieved June 5, 20209, from https://ai.bsa.org/wp-content/uploads/2019/09/AIHLEG_EthicsGuidelinesforTrustworthyAI-ENpdf.pdf

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secretai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dewey, M., & Wilkens, U. (2019). The bionic radiologist: Avoiding blurry pictures and providing greater insights.

Npj Digital Medicine, *2*(1), 65. https://doi.org/10.1038/s41746-019-0142-9

Dhaliwal, J. S., & Benbasat, I. (1996). The use and effects of knowledge-based system explanations: Theoretical foundations and a framework for empirical evaluation. *Information Systems Research*, *7*(3), 342–362. https://doi.org/10.1287/isre.7.3.342

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning, arXiv preprint*, 1–13. https://arxiv.org/abs/1702.08608

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

Elbanna, A., Dwivedi, Y., Bunker, D., & Wastell, D. (2020). The search for smartness in working, living and organising: Beyond the 'technomagic. *Information Systems Frontiers*, *22* (2), 275–280. https://doi.org/10.1007/s10796-020-10013-8

Fagan, L. M., Shortliffe, E. H., & Buchanan, B. G. (1980). COMPUTER-BASED MEDICAL DECISION MAKING: FROM MYCIN TO VM. *Automedica*.

Fernandez, A., Herrera, F., Cordon, O., Jose Del Jesus, M., & Marcelloni, F. (2019). Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational Intelligence Magazine*, *14*(1), 69–81. https://doi.org/10.1109/MCI.2018.2881645

Friedman, C. P., Elstein, A. S., Wolf, F. M., Murphy, G. C., Franz, T. M., Heckerling, P. S., Fine, P. L., Miller, T. M., & Abraham, V. (1999). Enhancement of clinicians' diagnostic reasoning by computer-based consultation. *JAMA*, *282*(19), 1851–1856. https://doi.org/10.1001/jama.282.19.1851

Fürnkranz, J., Kliegr, T., & Paulheim, H. (2020). On cognitive preferences and the plausibility of rule-based models. Machine Learning, 109, 853–898. https://doi.org/10.1007/s10994-019-05856–5 4 109 doi:10.1007/s10994-019-05856-5

Garlick, B. (2017). *Flying smarter: AI & machine learning in aviation autopilot systems*. Stanford University.

Giboney, J. S., Brown, S. A., Lowry, P. B., & Nunamaker, J. F. (2015). User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit. *Decision Support Systems*, *72*, 1–10. https://doi.org/10.1016/j.dss.2015.02.005

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

Goddard, K., Roudsari, A., & Wyatt, J. C. (2011). Automation bias - a hidden issue for clinical decision support system use. *Studies in Health Technology and Informatics*, *164*, 17–22. https://doi.org/10.3233/978-1-60750-709-3-17

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association?: JAMIA*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Gönül, M. S., Önkal, D., & Lawrence, M. (2006). The effects of structural characteristics of explanations on use of a DSS. *Decision Support Systems*, *42*(3), 1481–1493. https://doi.org/10.1016/j.dss.2005.12.003

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a right to explanation. *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gregor, S., & Yu, X. (2002). Exploring the explanatory capabilities of intelligent system technologies. In V. Dimitrov & V. Korotkich (Eds.), *Fuzzy Logic* (pp. 288–300). Physica-Verlag HD.

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

Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, *37*(3), 362–386. https://doi.org/10.1002/rob.21918

Grudin, J. (2019). AI summers' do not take jobs. *Communications of the ACM*, *59*(2), 8–9. https://dl.acm.org/doi/fullHtml/10.1145/2871050

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Gunning, D. (2017). *Explainable artificial intelligence (XAI)*. DARPA Program Update November. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Haugeland, J. (1985). *Artificial intelligence: The very idea*. Massachusetts Institute of Technology, MIT PRESS.

Herse, S., Vitale, J., Tonkin, M., Ebrahimian, D., Ojha, S., Johnston, B., Judge, W., & Williams, M. A. (2018). Do you trust me, blindly? Factors influencing trust towards a robot recommender system. *RO-MAN 2018-27th IEEE International Symposium on Robot and Human Interactive Communication*. https://doi.org/10.1109/ROMAN.2018.8525581

Hohman, F., Kahng, M., Pienta, R., & Chau, D. H. (2019). Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, *25*(8), 2674–2693. https://doi.org/10.1109/TVCG.2018.2843369

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, *9*(4), 1–13. https://doi.org/10.1002/widm.1312

Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial intelligence: The global landscape of ethics guidelines. *Nature Maschine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kasnakoglu, B. T. (2016). Antecedents and consequences of co-creation in credence-based service contexts. *The Service Industries Journal*, *36*(1–2), 1–20. https://doi.org/10.1080/02642069.2016.1138472

Kistan, T., Gardi, A., & Sabatini, R. (2018). Machine learning and cognitive ergonomics in air traffic management: Recent developments and considerations for certification. *Aerospace*, *5*(4), 103. https://doi.org/10.3390/aerospace5040103

Kühl, N., Lobana, J., & Meske, C. (2019). Do you comply with AI? - Personalized explanations of learning algorithms and their impact on employees' compliance behavior. *40th International Conference on Information Systems (ICIS)*, 1–6. https://arxiv.org/pdf/2002.08777.pdf

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4066–4076). Curran Associates, Inc.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1096), 1–8. https://doi.org/10.1038/s41467-019-08987-4

Li, M., & Gregor, S. (2011). Outcomes of effective explanations: Empowering citizens through online advice. *Decision Support Systems*, *52*(1), 119–132. https://doi.org/10.1016/j.dss.2011.06.001

Lynch, J., & Schuler, D. (1990). Consumer evaluation of the quality of hospital services from an economics of information perspective. *Journal of Health Care Marketing*, *10*(2), 16–22. https://pubmed.ncbi.nlm.nih.gov/10105192/

Malhotra, A., Melville, N. P., & Watson, R. T. (2013). Spurring Impactful Research on Information Systems for Environmental Sustainability. *MIS Quarterly*, *37*(4), 1265–1274. https://doi.org/10.1002/mrdd

Mao, J.-Y., & Benbasat, I. (2000). The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *Journal of Management Information Systems*, *17*(2), 153–179. https://doi.org/10.1080/07421222.2000.11045646

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

Matzner, M., Büttgen, M., Demirkan, H., Spohrer, J., Alter, S., Fritzsche, A., Ng, I. C. L., Jonas, J. M., Martinez, V., Möslein, K. M., & Neely, A. (2018). Digital transformation in service management. *Journal of Service Management Research*, *2*(2), 3–21. https://doi.org/10.15358/2511-8676-2018-2-3

Mccarthy, J., Minsky, M., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12–14. https://doi.org/10.1609/aimag.v27i4.1904

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., & 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

Mikalef, P., Popovic, A., Lundström, J. E., & Conboy, K. (2020). special issue call for papers: Dark side of analytics and AI. *The European Journal of Information Systems*. https://www.journalconferencejob.com/ejis-dark-side-of-analytics-and-ai

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

Narla, A., Kuprel, B., Sarin, K., Novoa, R., & Ko, J. (2018). Automated Classification of Skin Lesions: From Pixels to Practice. *Journal of Investigative Dermatology*, *138*(10), 2108–2110. https://doi.org/10.1016/j.jid.2018.06.175

Papamichail, K. N., & French, S. (2005). Design and evaluation of an intelligent decision support system for nuclear emergencies. *Decision Support Systems*, *41*(1), 84–111. https://doi.org/10.1016/j.dss.2004.04.014

Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing Bias in Artificial Intelligence in Health Care. *JAMA*, *322*(24), 2377–2378. https://doi.org/10.1001/jama.2019.18058

Radaelli, L., de Montioye, Y.-A., Singh, V. K., & Pentland, A. P. (2015). Unique in the shopping mall: On the reidentifiability of credit and card metadata. *Science*, *347*(6221), 536–539. https://doi.org/10.1126/science.1256297

Rahman, M., Yu, X., & Srinivasan, B. (1999). A neural networks based approach for fast mining characteristic rules. In *Foo N,(eds) advanced topics in artificial intelligence. AI 199. Lecture notes in computer science, vol 1747* (pp. 36–47). Springer. https://doi.org/10.1007/3-540-46695-9_4

Rai, A., Constantinides, P., & Sarker, S. (2019). Editor's comments: Next-generation digital platforms: Toward human-AI hybrids. *MIS Quarterly*, *43*(1), 3–4. https://dl.acm.org/doi/10.5555/3370135.3370136

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçütürk, U. Güçlü, & M. Gerven (Eds.), *Explainable and interpretable models in computer vision and machine learning. The Springer series on challenges in machine learning* (pp. pp. 19–36). Cham, Schweiz.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

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

Schneider, J., & Handali, J. (2019). Personalized explanation in machine learning: A conceptualization. *27th European Conference on Information Systems (ECIS 2019)*, 1–17. https://arxiv.org/pdf/1901.00770.pdf

Schneider, J., Handali, J., Vlachos, M., & Meske, C. (2020). Deceptive AI Explanations: Creation and Detection. *arxiv*, *2001*, 07641. https://arxiv.org/pdf/2001.07641.pdf

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Shirer, M., & Daquila, M. (2019). *Worldwide spending on artificial intelligence systems will be nearly $98 Billion in 2023, According to New IDC Spending Guide*. International Data Corporation (IDC).

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, *23*(5), 828–841. https://doi.org/10.1109/TEVC.2019.2890858

Sutton, S. G., Arnold, V., & Holt, M. (2018). How much automation is too much? Keeping the human relevant in knowledge work. *Journal of Emerging Technologies in Accounting*, *15*(2), 15–25. https://doi.org/10.2308/jeta-52311

Swartout, W. R., & Smoliar, S. W. (1987). On making expert systems more like experts. *Expert Systems*, *4*(3), 196–208. https://doi.org/10.1111/j.1468-0394.1987.tb00143.x

Union, E. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da* (Off. J. Eur. Union L119; pp. 1–88).

van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. *Proceedings of the 16th Conference on Innovative Applications of Artificial Intelligence*, 900–907. https://www.aaai.org/Papers/IAAI/2004/IAAI04-019.pdf

Vidgen, R., Shaw, S., & Grant, D. B. (2017). Management challenges in creating value from business analytics. *European Journal of Operational Research*, *261*(2), 626–639. https://doi.org/10.1016/j.ejor.2017.02.023

Wang, H., Li, C., Gu, B., & Min, W. (2019). Does AI-based credit scoring improve financial inclusion? Evidence from online payday lending. *In proceedings of the 40th international conference on information systems, Paper ID 3418*, Munich, Germany, pp. 1–9.

Watson, H. (2017). Preparing for the cognitive generation of decision support. *MIS Quarterly Executive*, *16*(2), 153–169. https://www.semanticscholar.org/paper/Preparing-for-the-Cognitive-Generation-of-Decision-Watson/766825192ccec1419564c9882a857339cc4e9a44

Wood, S., & Schulman, K. (2019). The doctor-of-the-future is in: patient responses to disruptive health-care innovations. *Journal of the Association for Consumer Research*, *4*(3), 231–243. https://doi.org/10.1086/704106

Xiao, L., Shen, X.-L., Cheng, X., Mou, J., & Zarifis, A. (2020). Call for Papers - The Dark Sides of AI. *Electronic Markets*. https://www.springer.com/journal/12525/updates/17695144

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

Ye, L. R., & Johnson, P. E. (1995). The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly*, *19*(2), 157–172. https://doi.org/10.2307/249686

Zolbanin, H. M., Delen, D., Crosby, D., & Wright, D. (2019). A predictive analytics-based decision support system for drug courts. *Information Systems Frontiers*, *22*, 1–20. https://doi.org/10.1007/s10796-019-09934-w

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75–89. https://doi.org/10.1057/jit.2015.5