Explainable Artificial Intelligence:
Objectives, Stakeholders and Future Research Opportunities

Christian Meske¹, Enrico Bunde¹, Johannes Schneider², Martin Gersch¹

¹Freie Universität Berlin and Einstein Center Digital Future
Department of Information Systems
Garystr. 21, 14195 Berlin, Germany
Corresponding author: Christian Meske, christian.meske@fu-berlin.de

²University of Liechtenstein
Institute of Information Systems
Fürst-Franz-Josef-Strasse, 9490 Vaduz, Liechtenstein

Recommended citation:
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
1 Introduction

Artificial Intelligence (AI), a research area initiated in the 1950ies (McCarth 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 usable 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 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.

2 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 towards 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 towards wrongful output. Caliskan, Bryson, and Narayanan (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), 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.

3 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 artefacts’ 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, Tanner, & Josephson, 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 publications1 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.

4 Explainable Artificial Intelligence

4.1 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 II, 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 II, 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

---

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.
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 artefact (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).

4.2 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 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 affects 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.

![Figure 2: Generalized objectives of Explainable Artificial Intelligence](image)

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 artefact’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.

![Figure 3: Stakeholder groups of Explainable Artificial Intelligence](image)

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.

4.3 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 explanee (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).

*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., 2018). *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 explanees (Binns, 2018; Kusner et al., 2017). Figure 4 summarizes the quality criteria for personalized explanations.

![Figure 4: Quality criteria for personalized explanations](image)

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., 2018).
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 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 the following table.

<table>
<thead>
<tr>
<th>Research stream</th>
<th>Research question</th>
<th>Research contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral Science</td>
<td>How do AI explanations influence the users’ and mangers’ cognitive perception of the AI?</td>
<td>Knowledge about how explainability may be an important variable in existing theories about human perception of the world and IS artefacts (e.g., affordance theory, mental model theory, sensemaking, UTAUT, and others).</td>
</tr>
<tr>
<td></td>
<td>How do explanations influence employees’ compliance behavior and work practices?</td>
<td>Knowledge on how AI explanations support IT governance.</td>
</tr>
<tr>
<td></td>
<td>How do explanations help to detect bias in managerial decision making?</td>
<td>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).</td>
</tr>
<tr>
<td>Question</td>
<td>Knowledge</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Under which circumstances do explanations support or inhibit individual’s trust towards the AI?</td>
<td>Knowledge on how different levels of expertise and personality traits like risk aversion elicit different reactions to AI explanations.</td>
<td></td>
</tr>
<tr>
<td>How can explanations fulfill task-related needs of the different XAI stakeholders?</td>
<td>Knowledge on when and how explanations should be presented to users in order to increase task performance.</td>
<td></td>
</tr>
<tr>
<td>What are adequate metrics to evaluate AI explanations?</td>
<td>Knowledge on the dimensions that are relevant for explanations to be effective; differentiate “good” from “bad” explanations.</td>
<td></td>
</tr>
<tr>
<td>How do explanations influence (de)skilling of employees?</td>
<td>Knowledge on how explanations help to maintain or increase user qualification and self-efficacy regarding AI usage.</td>
<td></td>
</tr>
<tr>
<td><strong>Design Science</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>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)?</td>
<td>Bring together knowledge and methodical expertise of different disciplines in order to accelerate and improve XAI research across research communities.</td>
<td></td>
</tr>
<tr>
<td>Which features in explanations support the evaluation of an AI’s ethicality and morality?</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>How can the transdisciplinary design of AI explainability across different stakeholders look like?</td>
<td>Conceptualization of a standardized design process for fair, accountable and transparent AI, that take the needs of different stakeholders into account.</td>
<td></td>
</tr>
<tr>
<td>What are design principles on how to build explainable AI systems that allow for a flexible adaptation of explanations by users (based on technical possibilities)?</td>
<td>Knowledge of technical possibilities to allow for a flexible adaptation of explanations by users (based on technical possibilities).</td>
<td></td>
</tr>
</tbody>
</table>
stakeholder- and domain-specific personalization? 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.

Table 1: Summary of potential research opportunities and contributions

6 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.
References


International Data Corporation (IDC).


