Explainable AI: A key driver for AI adoption, a mistaken concept, or a practically irrelevant feature?

Julia Dvorak[1], Tobias Kopp[2], Steffen Kinkel[2] and Gisela Lanza[1]

[1] Institute of Production Science (wbk), Karlsruhe Institute of Technology
julia.dvorak@kit.edu
[2] Institute for Learning and Innovation in Networks (ILIN), Karlsruhe University of Applied Sciences
tobias.kopp@h-ka.de

Abstract. Explainable artificial intelligence (xAI) has become a popular subject of research amongst AI scholars in the last years. Some scholars consider xAI a significant driver of AI adoption in practice. However, at date, only a few studies investigated the conditions under which xAI solutions provide benefits in practice. Additionally, there is still a lot of controversy and inconsistency about related terminology revealing large conceptual differences between the understanding of explanations from a theoretical social science viewpoint and from a technological viewpoint. In this article, we strive to contribute to a more realistic picture of the potential and practical application scenarios of xAI. Thereby, we clarify the question whether xAI is a key driver for AI adoption, a mistaken concept from a theoretical point of view or perhaps a practically irrelevant feature and bridge the gap between different disciplines.

Keywords: explainable AI; artificial intelligence; trust; manufacturing; real-life applications.

1 Introduction

Along with the growing interest in the research community, the number of publications in the field of explainable artificial intelligence (xAI) has increased sharply in the last approximately five years [1]. xAI is attributed the potential to significantly drive AI adoption in practice [1], which is supposedly hindered by the black box nature of many AI models [2]. Thereby, xAI refers to the capacity of an AI system to explain either the model itself to the developer to achieve intrinsic interpretability or a specific AI outcome to a user as part of some kind of post-hoc reverse engineering process [3–5]. In that sense, xAI aims at explaining "the way in which an algorithm works in order to understand how and why it has delivered particular outcomes" [4].

The assumption that xAI can drive AI adoption in practice leads to the question of its practical benefits. In recent years, several researchers have strived to identify and collect possible motivations to implement xAI solutions. Explanations can be used to evaluate the AI application, to justify its reliance, to control its outcomes, to discover and to learn from it, i.e., to serve educational purposes [1, 5, 6]. In newer publications, the motivations for xAI applications are enriched with the purpose to manage AI applications [6]. Apart from social and practical reasons, some of these purposes address legal issues such as the demand for transparency [7], which has gained importance since it has been defined as a key criterion for trustworthy AI by the EU [8], or the right to "obtain meaningful information about the logic involved" [9] as granted by the European's General Data Protection Regulation (GDPR) [1, 2].

Despite the multitude of possible motivations to use xAI, the actionability of its output remains unclear, i.e., there is a lack of empirical studies investigating whether recipients of explanations are able to derive beneficial practical implications [10]. Additionally, there is still a lot of controversy and inconsistency about the terminology. This leads to a variety of co-existing definitions and large conceptual overlaps with related terms such as interpretability or transparency, which also triggers diverse expectations towards xAI by relevant user groups [7, 11]. Analogous to the term intelligence, explanation refers to a concept that has its roots in human and social sciences but is also present in everyday language. Consequently, these concepts are loaded with associations and expectations. Simply transferring them to technical contexts can easily result in exaggerated expectations.

Figure 1 visualizes the elements relevant to a xAI system. An AI application is applied to a specific data set and generates an output. The output of the machine learning algorithm and its quality thereby depend on the quality of the respective input data. Since the output of an AI algorithm is often opaque, a xAI can be used to gain further insights into the mechanisms that have led to this particular output or the model of the AI itself. A

human-machine interface makes these insights accessible to the user in a specific context. The human feedback may in turn be used as an input for the xAI.

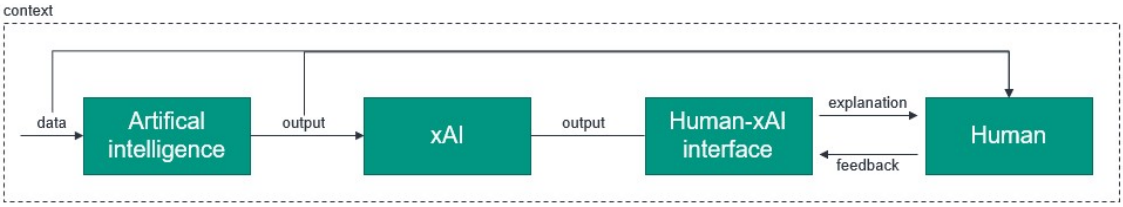

Figure 1 AI, xAI and Human embedded in a specific context following [12]

In this article, we aim to critically reflect on the opportunities and the range of practical use cases of xAI from two perspectives to contribute to a more realistic understanding and more adequate expectations regarding xAI. First, we highlight some conceptual limitations of xAI from a theoretical viewpoint mainly driven by arguments from scholars of philosophy and the social sciences. Thereby, we want to answer the question whether and in what sense xAI might be a mistaken and sometimes misunderstood concept. Second, we will adopt a practical viewpoint to analyse whether and in what application contexts xAI might be either a beneficial or a practically irrelevant feature. With this in mind, we will narrow and specify the potential application domains of xAI from two thrusts, while also highlighting relevant potential use cases in appropriate contexts. Finally, we will bring theoretical and practical considerations together in order to derive practical implications as well as future research avenues in the final chapter.

2 Theoretical viewpoint: Explainable AI as a mistaken concept?

Human beings are familiar with explanations since they represent an essential part of everyday communication. We regularly ask why a certain event (the explanandum) happened and hope to be provided with an answer (the explanans) that satisfies our information needs regarding reasons and causes. After having received an explanation, we might be asked whether we feel that we have understood a certain phenomenon. However, in many cases, the answer to this question is gradual in a sense that we have understood something "a bit more", but our information needs are not entirely satisfied. At the contrary, the provided explanation let further questions arise, so that explaining becomes a complex, iterative, and interactive process [13–15].

Due to our familiarity with explanations, we can quite easily develop a vision of what is supposedly meant by xAI or by AI-generated explanations. However, this also imposes the danger that we subliminally expect AI explanations to be of a similar kind than explanations we know from our everyday communication amongst human beings. Unfortunately, despite being a common term in everyday language and a concept of folk psychology, there is no consensus on the criteria a certain piece of information has to meet in order to be conceived as an explanation, let alone on the criteria for a *good* explanation [1]. Accordingly, it is challenging to develop objective evaluation metrics for xAI approaches, an issue only addressed by a comparably low number of studies so far [16]. Hence, in the first place, xAI algorithms provide certain information, but do they qualify to be regarded as explanation in a narrower sense? In the remainder of this chapter, we provide four theses about the limits of xAI, which contradict common assumptions about the similar nature and explanatory power of xAI-generated and human explanations.

2.1 xAI does not provide explanations that resemble human explanations

Whereas scholars with technical backgrounds tend to consider explainability as a feature of the AI system, social scientists consider explanations as audience-dependent involving cognitive as well as social processes [15, 17]. The explainer exchanges some information regarding the explanandum with the explainee. However, when someone provides information about the conditions under which a certain outcome arises, this does not imply that the outcome is sufficiently explained. In contrast, the information recipient ultimately decides whether the information had explanatory power and supported him/her in interpreting the outcomes or in the understanding of the model itself. Put in other words, the definition of an explanation is relative to the explainee and its individual expectations, prior knowledge on the explanandum, information needs and so on [17]. Hence, the output of an xAI represents a necessary but not a sufficient piece of information which serves as an input for an explanatory cognitive process on the user side [17]. Ideally, this process should result in a feeling that

someone has understood certain phenomena and its formation conditions. Hence, whereas explainability by definition focusses on whether some outcome is explainable per se, the focus should be shifted to the user side and its understanding of a phenomenon after having processed an explanans [18].

This argumentation draws on an interactional and processual understanding of explanations, which corresponds with the way humans explain things to each other as part of social interaction and conversation [14, 17]. Processing explanations is not an entirely rational and optimized process, which can be best supported by a large amount of exact data. Quite the contrary, those processes usually suffer from cognitive biases and are shaped by social expectations, which might prevent someone from revealing one's true reasons [14]. Furthermore, people prefer pragmatic, functional explanations, which can guide future behaviour, instead of mechanistic, data-driven explanations [18]. From a social sciences perspective, most explanations are contrastive, selective, social and do not refer to probabilities [14]. Accordingly , people usually do not seek for a complete list of causes when they ask for an explanation, but rather ask why a certain event happened instead of another plausible event [14]. Seemingly, most existing xAI approaches fail to take these criteria sufficiently into account. However, some approaches like counterfactual explanations or approaches to equip explanation engines such as SHAP (Shapley Additive exPlanations) with interactive features [19] can be considered as a step towards explanation processes that more closely resemble interactive human-to-human explanation processes. In particular, counterfactual explanations, which have a long tradition of research in analytic philosophy [20], resemble human explanations and thus might be positively evaluated in user studies, although they provide its user with only a limited set of information about possible causes [21]. Hence, whilst there are approaches to provide more human-like explanations, the conceptual differences seem considerably large at the moment. However, this does not mean that human-like explanations are the best possible explanations from a normative viewpoint, but they are at least the type of explanations which we are familiar with and which we might intuitively expect from xAI applications. Consequently, the conceptual differences at least underpin the relevance of expectation management.

2.2 xAI does not necessarily lead to correct expectations and higher trust in an AI system

Increasingly complex organisational styles of leadership as well as seemingly non-deterministic and unpredictably acting modern technology have given rise to the relevance of trust in human-technology interaction [22]. The described opaque, intangible, and complex nature of AI systems, their novelty as well as the manifold expectations and associations make them difficult to comprehend. Users often struggle to anticipate outcomes and to adequately assess the ability of AI systems, which in turn affects the evaluation of its trustworthiness [23]. Consequently, calibrating users' trust in AI and thereby influencing users' willingness to rely on an AI system is an important practical issue.

Basically, trust is about expectations regarding the unknown course of the future [24]. Hence, trust becomes relevant in situations in which persons perceive a lack of knowledge about the future while possessing some relevant information about the past from which they derive their expectations [24]. By enhancing users' knowledge about the working principles of AI systems, xAI applications strive to support users in making predictions about the future output of AI and thereby affect their trust in AI systems. Accordingly, fostering users' trust in an AI system is often considered as a major goal of xAI [7, 25] and measuring real-time trust as an opportunity to assess the quality of generated explanations [26].

However, research in related contexts of human-machine interaction have highlighted the significant relevance of a large variety of contextual as well as individual factors on trust evolution [27–29]. Preliminary results from the few available user studies in the human-AI interaction context seem to confirm this finding by indicating a complex and far from straightforward relationship between the characteristics of the xAI system and perceived trust [18]. For instance, [30] found that visual explanations increased perceived observability of an AI's working principles, but did not lead to an increase in trust. Additionally, meaningless placebic explanations without any explanatory power resulted in comparable levels of perceived trust than real explanations [31], just to name two examples. Consequently, much more theoretical and empirical work is required in order to analyze the relationship between perceived interpretability of AI output and human-AI trust under consideration of relevant contextual factors. Meanwhile, the widespread assumption that xAI implementations necessarily foster users' trust in any situation remains doubtable. An interdisciplinary perspective incorporating existing literature from sociology, psychology, philosophy of technology, and human-robot interaction seems promising to close this research gap [32].

2.3 xAI does not address the need for meaningful narratives

Explanations are usually provided in response to a why-question. Many reasons can drive someone to ask such a question [14]. For instance, if an AI application in manufacturing suggests replacing a certain part of a machine, one could ask "Why should I do this?". Or if an AI system refuses to grant creditworthiness, the applicant might ask "Why am I refused?".

Whilst the relation between why-questions and explanations seems intuitive and straightforward, a closer look reveals different notions and meanings of this question. As philosopher Daniel Dennett states, why-questions sometimes ask for a retrospective process narrative precisely answering the question "How did you come to this decision?" [33]. In case of the application for evaluating creditworthiness, the system could present a series of causes, mathematical calculations, and correlations to describe the process of generating its final decision. However, this kind of explanation is unlikely to satisfy the applicant and make him/her accept the decision. Indeed, automated decision-making systems often lack acceptance, because they fail to satisfy humans' need to understand decisions [2]. Here, it is crucial to have a closer look at the possible notions of the application's why-question. The latter may not only ask for a causal series of events, but also refer to intentions and meaning. Hence, they precisely ask the question "What did this happen for?" [33] and thereby request "justifications of the system's actions or recommendations *(why)"* as opposed to "rule-oriented explanations of *how* the system reasoned" [6]. This is in line with humans' tendency to build stories around their reasons for certain decisions to provide a deeper meaning and to foster trust. They "explain themselves by referring to post-factum coherent stories" [2] including additional environmental information. Oftentimes, humans also expect such stories as explanations. One's request for an explanation is not necessarily driven by the need for exact mathematical calculations, but by humans' condemnation to make-sense even in a "world of data, correlations, and probabilities" [34]. This process of sense-making requires convincing narratives instead of pure information. If a good friend betrayed you, you most likely would not be willing to forgive him/her, if he/she would only explicate some factors and correlations that made it quite probable that he/she would ever betray you. In this context, [34] concludes that humans have a narrative responsibility which is actually challenged by the rise of AI applications. The latter can provide a series of possible causes but fail to generate stories and narratives and to answer the important *what for*-question. In that sense, not every need for an explanation in a practical scenario can be addressed by xAI, because the ones asking for an explanation might indeed seek for a narrative.

2.4 xAI does not sufficiently take into account diverging addresses and their needs

The arguments brought forth so far have stressed an interactional understanding of explanation processes strongly involving the explainee's cognitive processes. This hints at intersubjective differences in these processes. However, many articles about xAI do neither specify their target audience nor name the motives for which their xAI can be usefully applied. Thereby, they overlook that the requirements and information needs of different target groups for explanations substantially differ [2], especially if they have varying levels of expertise with respect to AI technology and domain-specific knowledge [2, 6]. This can have various implications in terms of how they understand and emotionally evaluate the explanations. Whereas novices might require very detailed and easy-to-comprehend explanations, AI-experts might feel offended by easy explanations which disregard their experience and knowledge [2]. Additionally, laypersons might tolerate a less accurate but understandable explanation whereas experts might prefer more accurate but very technical and complex explanations [1]. It has to be taken into account that the real-life end users of xAI systems can also encompass non-technical users [10].

Whereas most publications omit to address this issue, some scholars have tried to raise attention for variety of possible user groups and their diverging demands in recent years [35–37]. For instance, [37] proposed to differentiate stakeholders into regulators, users, developers, and persons affected by AI outputs. Still, the knowledge about the influence of user characteristics on users' perception of explanations and the attitudinal, affective, and behavioral implications is limited [5, 11, 38]. However, these are key questions in terms of xAI adoption in practice as highlighted by a recent meta-analysis on the future of AI at work [39]. In that sense, apart from the explanandum (what to explain?) and the explanans (how to explain?), the explainee (to whom to explain?) should be considered as a third essential building block of an explanation [2]. This would also stress the value-added by personalized explanations [6]. However, since some of the most popular xAI approaches like LIME (Local Interpretable Model Agnostic Explanation) or SHAP are model- and application-agnostic, they lack methods of personalization. Hence, this advantage in terms of wide applicability of these techniques backfires in terms of an appropriate consideration of application-specific user needs [40].

Surprisingly, although companies represent the key application domain for AI and xAI solutions, an organizational and workplace-oriented perspective on xAI is underrepresented in related literature and in public discourse [11].

3 Practical viewpoint: Explainable AI as a practically irrelevant feature?

Until this point, we discussed xAI with a strong focus on the nature of explanations and the differences between human and xAI-generated explanations. The forthcoming section focusses on the practical usefulness of xAI in an organizational context, without considering technical details of these applications. Highlighting the relation between AI, xAI and the user, several criteria to assess existing use cases regarding their potential of an xAI application will be discussed. The identification of use cases goes thereby in line with a set of design recommendations for an xAI. We use the domain of production as an example. However, we believe that the criteria identified are rather associated with the nature of xAI than with a specific application domain and it is the goal to develop criteria that can identify xAI use cases on a domain-independent level. .

3.1 Applications of AI and xAI in company

Use cases for AI are found across a broad range of domains as for example production, medical, mobility, but also education or scientific work. In a study conducted by Fraunhofer IAO potential fields of application for AI are identified, including autonomous robots and transport, but also cognitive assistants and smart devices [41]. Production as it is considered in the course of this publication is not a “primary” domain such as for example autonomous driving, but there remains a broad range of AI use cases. In [42], a variety of such use cases are worked out abstractly, including maintenance, quality management and control, automation technology, and product and process development. In particular in small and medium enterprises, the assumed potential of AI applications is far from being identified or even realized, mainly due to a lack of competence, obstacles for the actual implementation and data problems [43].

xAI use cases in industry include cases on predictive maintenance, business management, anomaly detection and modeling [16]. However, it remains open what the potential for xAI actually is across all levels from shop floor to management level. For practical applications, the motivation to use an xAI is essential to justify its application, as it has been discussed beforehand in the introduction [44]. Ranging from shop floor to management level, the availability of high-quality data is an essential basis to make decisions. Confidence in how the systems work must thus be created, but at the same time it must also be ensured that the database is of sufficient quality [45]. In addition to strategic decisions, other possible applications for xAI are, for example, the training of employees, product development and the sharpening of process understanding [46].

3.2 Derivation of criteria for use case analysis and xAI application

So far, we have highlighted some typical application domains, but the question how to identify future use cases to assess the potential of xAI applications remains open. In the course of this section, several criteria will be presented that could serve practitioners as a first indication for an xAI potential. Thereby, neither exhaustiveness nor completeness is claimed.

The first aspect to mention is the need for an explanation [5], although from the authors' point of view this aspect can also be well justified via the objectives of an xAI. As a second criterion the criticality of for example a decision is to mention [5, 31]. The criticality can therefore refer to the fault tolerance of a use case, the criticality in time [41] and the criticality of data itself, resulting in criticality of the use case itself and the resulting implications [47]. Along with time criticality comes the decision if an explanation must be provided in real time or if the post-hoc provision of an explanation is sufficient for evaluation and knowledge enhancement purposes [48]. This question is particularly relevant in the context of manufacturing work, which is often associated with time pressure. Furthermore, the time window for a reaction to a decision and the explanation provided by an AI or xAI must be considered. For decisions of the strategic management in companies which are associated with a high impact, time criticality is not an important aspect, however, since those may be based on simulations, explanations are of crucial importance [46]. Also from a practical point of view the prior knowledge of stakeholders of a practical use case or system must be subject of consideration [49].

3.3 xAI design for practical application

In line with the previous aspects, it should be noted that the user of the xAI plays a central role and that the design of an xAI must be determined to its user group [50]. In their study, [50] contrast different explanations to different user groups and conclude that also especially the application context of an xAI is of great importance. This is also confirmed in other publications (e.g. [44]), particularly more recent publications. In addition to the type of information that is provided to the user, it is also recommended to consider the level of detail of information that should be provided to the user to ensure a communication appropriate to the target group [50]. It is emphasized that the user should be included in the early stages of an xAI development to ensure a fulfillment of expectations [12]. In current developments, this aspect has not been paid significant attention to [6]. If one considers further the relation between AI-xAI-human as shown in Figure 1, the user interface and the way of information visualization is an additional critical aspect [46]. From an entrepreneurial perspective, further challenges can be identified apart from the actual application. In [46], it is emphasized that its economic evaluation of xAI is difficult. If we look in particular at small and medium enterprises, we already find barriers to the adoption of AI approaches [51]. For example, in smaller companies it may be the case that human intervention is required along the information chain [52] and that an insufficient amount of data is provided for an AI application [53]. This may be the case due to several reasons, for example, if a number of different systems is used and data availability to even train an AI is not guaranteed [52].

Moreover, cost may be a critical criterion for the decision if an AI or xAI is even considered within the scope of potential solutions. Usually, before a product is fully developed, a number of prototypes can be created. However, in the case of an AI, this is not possible without major effort, including financial effort [51] and it is assumed that this applies analogously to xAI which also leads to financial risk at this point. Based on the previous chapters, it can also be stated that explanations that appeal to the user and are thus truly perceived as explanations offer the greatest added value. [49] underpins this approach by focusing on current demand and thus user requirements. Nevertheless, according to [47], these so-called application-grounded explanations, including real humans and real tasks, are also the explanations that are associated with the highest costs, which again represents a barrier for companies. However, it is conspicuous that even for the adoption of AI cost is not a primary factor, but major reasons range around competencies, data, or infrastructure [43].

4 Bridging the gap between theory and practice

In this section, we try to bring together the human-centric and the application-specific perspective on explanations and xAI. Alike for any other application within a company or production, a tool or technology must fulfill a function. The first statement refers to the human perception of explanations that are provided by an xAI, but that do not resemble human explanations. Thinking in terms of a practical use case, this aspect is to be considered as controversial, since sometimes human beings provide an inadequate amount of information than it is actually needed in a specific use case. This can on the one hand cause problems in time critical environments, but on the other hand contribute to trust building. From a practitioner's point of view, trust in a system is an essential aspect. Trust is hampered by the fact that AI system's behavior is not to be anticipated, but the question arises to which degree this must be the case in practical use cases. Explanations do not seem relevant in any cases, since we oftentimes accept decisions based on human reasoning, which is also a black box, and humans' explanations, which might not detail causes but only represent ex-post rationalizations [2]. xAI is often assigned the quality to foster users' trust in a system by unfolding the AI's working principles and decision mechanisms. However, if the underlying problem was so deterministic and obvious, would there be a need for an AI at all? AI is working with data, whilst human beings often share a broader view on a phenomenon as a whole including observations and implicit knowledge [54]. Thinking of use cases on the shop floor and in higher management, different context and settings are found with respect to time, consequences of decisions, and so on. The actual need for trust might be context-specific and there might even be cases, in which trust in an explanation is not an essential requirement (e.g. in cases without any human-AI interaction). However, in many cases, trust might be an essential success factor, but xAI applications might fail to modulate trust appropriately due to other overshadowing contextual factors.

Human beings do not only expect the transmission of an information, but rather information embedded in a story that provides a full picture with potentially more information that are required from a merely rational stance. Nevertheless, human beings do not act merely rationally and thereby desire and expect explanations in the form of narratives. Such explanatory stories can be shaped by subjective impressions. Thinking about production and for example predictive maintenance, the information itself may be of greater interest, but its

visualization format may be of lower interest. In any case, the provided information and its visual representation should contribute to an adequate expectation management addressing the fact that human's expectations differ. However, since algorithms tend to work the same, a different set of algorithms may be needed for different people and different use cases.

This leads to the last aspect of consideration, the use specificity. From a practical perspective, there is a difference in talking to a developer or talking to a worker on the shop floor. Both will have different requirements, whilst an engineer might want to understand how the implemented algorithm works, the higher management wants to know why a specific decision should be taken based on a simulation. The claim for user specific explanations is addressed from a theoretical and also practical viewpoint alike. Information must be provided with respect to the context, the stakeholder, and the prior knowledge of a person.

Interdisciplinarity, therefore, fosters shedding light on relevant issues from different perspectives. Thereby, a holistic picture of current limitations, obstacles and challenges regarding xAI and its use in practice is created. Nonetheless, the importance of each aspect is determined by the context in which an AI and an xAI is applied. We therefore highly stress the importance of research focusing more on the actual adoption of xAI and system design, but also the transfer of xAI to real use cases and not solely on the algorithmic development. In addition, proper expectation management towards xAI is crucial. This may then again transfer xAI from a theoretical construct into a practically relevant feature.

5 Conclusion and implications for research and practice

In this contribution we approached xAI from a twofold perspective trying to answer the question if xAI is either a key driver for AI adoption, a mistaken concept, or practically irrelevant feature. First, the term explanation itself has been challenged and contrasted with human explanations. Second, the term trust within the scope of AI and xAI has been considered as a subject of interest, which is not necessarily influenced by the provision of AI-generated explanations. Third, the human need for storytelling instead of factual mathematical explanations was highlighted. This goes along with the fourth aspect pointing out the that current xAI implementations tend to provide general explanations, which are neither adapted to a specific situation nor a specific person. From a practical perspective, insights were provided into criteria that may promote or hamper the application of xAI. The widespread use of AI is a prerequisite for xAI adoption, but to use AI, data must be provided which is seen as a major obstacle by small and medium enterprises. Criticality of time and decision, control, speed of response and acceptance are identified as first criteria to assess AI-use cases for their potential of xAI. For sure, this list of criteria should be extended in further work.

In order to answer the research question, theoretical as well as practical issues and challenges were discussed. Thereby, we analyzed some important issues from different perspectives. Based on the theoretical-philosophical and the practical view the following implications for further research and practice can be derived:

- **Context specificity and user centeredness**. We highly emphasize the consideration of the context of an xAI application and the specific characteristics and needs of users. We therefore suggest not to neglect any of the relations indicated in Figure 1.
- **Empirical context-specific user studies.** The demand for a user-centered design of xAI approaches in practice requires a precise understanding how users perceive different explanation types in their typical usage contexts. Much more empirical findings are needed to be able to adequately design xAI approaches.
- **Expectation management**. Since the terms intelligence and explanation are common in everyday language and loaded with associations, practitioners will most likely expect something that explains in similar way than humans when they are told about xAI approaches. This can easily lead to false expectations and misleading ability attributions, which can negatively influence appropriate trust calibration. Mechanisms for expectation management and trust calibration should be developed and empirically tested to tackle this issue.
- **Promotion of AI for small and medium-sized enterprises (SMEs)**. xAI should not be considered as a universal solution to foster AI adoption in any company and usage context. Instead, it can be helpful for example when the AI adoption is hindered by transparency requirements. However, other typical obstacles such as missing technical competencies or a lack of high-quality data, especially in SMEs, must first be overcome to enable the use of xAI.

- **Criteria for xAI identification and heat map**. From a practical perspective, there is a need to develop and provide a guide for practitioners that empowers them to recognize beneficial xAI use cases based on a set of context- and use case-specific criteria.
- **Empirical analysis of overall potential of xAI in practice**. Apart from technical progress and single-case studies, empirical studies should also evaluate the potential of xAI in practice from a macro perspective. Thereby, they should address the research question, which practical use cases indeed fulfill the basic criteria for xAI to be feasible and advantageous. Such a study is foreseen within the context of the research project KARL.

Acknowledgements

This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within the "The Future of Value Creation – Research on Production, Services and Work" program (funding number 02L19C250) and managed by the Project Management Agency Karlsruhe (PTKA). The authors are responsible for the content of this publication.

References

[1] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable Artificial Intelligence," *WIREs Data Mining Knowl Discov*, vol. 11, no. 1, 2021, doi: 10.1002/widm.1391.

[2] N. Burkart and M. F. Huber, "A Survey on the Explainability of Supervised Machine Learning," *jair*, vol. 70, pp. 245–317, 2021, doi: 10.1613/jair.1.12228.

[3] A. B. Arrieta *et al.,* "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Oct. 2019. Accessed: Nov. 12 2021.

[4] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," *International Journal of Human-Computer Studies*, vol. 146, 2021, doi: 10.1016/j.ijhcs.2020.102551.

[5] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[6] C. Meske, E. Bunde, J. Schneider, and M. Gersch, "Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities," *Information Systems Management*, vol. 39, no. 1, pp. 53–63, 2022, doi: 10.1080/10580530.2020.1849465.

[7] A. Brennen, "What Do People Really Want When They Say They Want "Explainable AI?" We Asked 60 Stakeholders," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA, 2020, pp. 1–7.

[8] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," 2019. Accessed: Jun. 7 2022.

[9] Parliament and Council of the European Union, "General data protection regulation.," 2016. [Online]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_de

[10] S. Verma, A. Lahiri, J. P. Dickerson, and S.-I. Lee, "Pitfalls of Explainable ML: An Industry Perspective," Jun. 2021. Accessed: Jun. 7 2022. [Online]. Available: https://arxiv.org/pdf/2106.07758.pdf

[11] E. Hafermalz and M. Huysman, "Please Explain: Key Questions for Explainable AI research from an Organizational perspective," *Morals & Machines*, vol. 1, no. 2, pp. 10–23, 2021, doi: 10.5771/2747-5174-2021-2-10.

[12] T. A. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. van den Bosch, "Human-centered XAI: Developing design patterns for explanations of clinical decision support systems," *International Journal of Human-Computer Studies*, vol. 154, p. 102684, 2021, doi: 10.1016/j.ijhcs.2021.102684.

[13] A. Vultureanu-Albisi and C. Badica, "Recommender Systems: An Explainable AI Perspective," pp. 1–6, 2021, doi: 10.1109/INISTA52262.2021.9548125.

[14] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019, doi: 10.1016/j.artint.2018.07.007.

[15] P. Madumal, “Explainable Agency in Intelligent Agents,” in *AAMAS 2019*, 2019. Accessed: Feb. 12 2021.
[16] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, “A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks,” *Applied Sciences*, vol. 12, no. 3, p. 1353, 2022, doi: 10.3390/app12031353.
[17] K. O’Hara, “Explainable AI and the philosophy and practice of explanation,” *Computer Law & Security Review*, vol. 39, p. 105474, 2020, doi: 10.1016/j.clsr.2020.105474.
[18] A. Páez, “The Pragmatic Turn in Explainable Artificial Intelligence (XAI),” *Minds & Machines*, vol. 29, no. 3, pp. 441–459, 2019, doi: 10.1007/s11023-019-09502-w.
[19] M. Chromik, “Making SHAP Rap: Bridging Local and Global Insights Through Interaction and Narratives,” in *Springer eBook Collection*, vol. 12933, *Human-Computer Interaction – INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30 – September 3, 2021, Proceedings, Part II*, C. Ardito et al., Eds., 1st ed., Cham: Springer International Publishing; Imprint Springer, 2021, pp. 641–651.
[20] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR,” *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.
[21] G. Warren, M. T. Keane, and R. M. J. Byrne, “Features of Explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI,” 2022.
[22] J. D. Lee and K. A. See, “Trust in automation: designing for appropriate reliance,” *Hum Factors*, vol. 46, no. 1, pp. 50–80, 2004, doi: 10.1518/hfes.46.1.50_30392.
[23] R. C. Mayer, J. H. Davis, and F. D. Schoorman, “An Integrative Model of Organizational Trust,” *The Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995.
[24] P. Sumpf, *System Trust: Researching the Architecture of Trust in Systems*: Springer Fachmedien Wiesbaden, 2019.
[25] S. Laato, M. Tiainen, A. Najmul Islam, and M. Mäntymäki, “How to explain AI systems to end users: a systematic literature review and research agenda,” *INTR*, vol. 32, no. 7, pp. 1–31, 2022, doi: 10.1108/INTR-08-2021-0600.
[26] R. Setchi, M. B. Dehkordi, and J. S. Khan, “Explainable Robotics in Human-Robot Interactions,” *Procedia Computer Science*, vol. 176, pp. 3057–3066, 2020, doi: 10.1016/j.procs.2020.09.198.
[27] S. Marsh and M. R. Dibben, “The role of trust in information science and technology,” *Ann. Rev. Info. Sci. Tech.*, vol. 37, no. 1, pp. 465–498, 2003, doi: 10.1002/aris.1440370111.
[28] K. A. Hoff and M. Bashir, “Trust in automation: integrating empirical evidence on factors that influence trust,” *Human factors*, vol. 57, no. 3, pp. 407–434, 2015, doi: 10.1177/0018720814547570.
[29] P. A. Hancock, T. T. Kessler, A. D. Kaplan, J. C. Brill, and J. L. Szalma, “Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses,” *Human factors*, 1-34, 2020, doi: 10.1177/0018720820922080.
[30] T. Schrills and T. Franke, “Color for Characters - Effects of Visual Explanations of AI on Trust and Observability,” in *Springer eBook Collection*, vol. 12217, *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings*, H. Degen and L. Reinerman-Jones, Eds., 1st ed., Cham: Springer International Publishing; Imprint Springer, 2020, pp. 121–135.
[31] M. Eiband, D. Buschek, A. Kremer, and H. Hussmann, “The Impact of Placebic Explanations on Trust in Intelligent Systems,” in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk, 2019, pp. 1–6.
[32] T. Kopp, “Facets of trust and distrust in collaborative robots at the workplace: Towards a multidimensional and relational conceptualisation,” *International Journal of Social Robotics*, under review.
[33] D. C. Dennett, *From bacteria to Bach and back: The evolution of minds*. New York, London: W.W. Norton & Company, 2017.
[34] M. Coeckelbergh, “Narrative responsibility and artificial intelligence,” *AI & Soc*, 2021, doi: 10.1007/s00146-021-01375-x.
[35] A. Kirsch, “Explain to whom? Putting the User in the Center of Explainable AI,” *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 colocated with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), 2017, Bari, Italy.*, 2018.

[36] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, “Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems,” *in Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden*, pp. 8–14, 2018.
[37] M. Langer *et al.,* “What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research,” *Artificial Intelligence*, vol. 296, p. 103473, 2021, doi: 10.1016/j.artint.2021.103473.
[38] L.-V. Herm, J. Wanner, F. Seubert, and C. Janiesch, “I Don’t Get It, but It Seems Valid! The Connection Between Explainability and Comprehensibility in (X)AI Research,” in *European Conference on Information Systems (2021)*, 2021.
[39] M. Langer and R. N. Landers, “The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers,” *Computers in Human Behavior*, vol. 123, p. 106878, 2021, doi: 10.1016/j.chb.2021.106878.
[40] C. Panigutti, A. Beretta, F. Giannotti, and D. Pedreschi, “Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems,” in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA, 2022, pp. 1–9.
[41] D. Hecker, I. Döbel, U. Petersen, A. Rauschert, V. Schmitz, and A. Voss, “Zukunftsmarkt Künstliche Intelligenz: Potenziale und Anwendungen,” 2019.
[42] B. Hatiboglu, S. Schuler, A. Bildstein, and M. Hämmerle, “Einsatzfelder von künstlicher Intelligenz im Produktionsumfeld,” Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA, Stuttgart; Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO, Stuttgart, 2019.
[43] P. Ulrich, V. Frank, and M. Kratt, “Adoption of artificial intelligence technologies in German SMEs — Results from an empirical study,” in *Corporate governance: A search for emerging trends in the pandemic times*, 2021, pp. 76–84.
[44] B. Wickramanayake, C. Ouyang, C. Moreira, and Y. Xu, “Generating Purpose-Driven Explanations: The Case of Process Predictive Model Inspection,” in *Lecture Notes in Business Information Processing, Intelligent Information Systems*, J. de Weerdt and A. Polyvyanyy, Eds., Cham: Springer International Publishing, 2022, pp. 120–129.
[45] N. Tanwar and Y. Hasija, “Explainable AI; Are we there yet?,” *2022 IEEE Delhi Section Conference (DELCON)*, 2022.
[46] C. J. Turner and W. Garn, “Next generation DES simulation: A research agenda for human centric manufacturing systems,” *Journal of Industrial Information Integration*, vol. 28, p. 100354, 2022, doi: 10.1016/j.jii.2022.100354.
[47] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” *arXiv: Machine Learning*, 2017, doi: 10.48550/arXiv.1702.08608.
[48] C. H. Park, “Anomaly Pattern Detection on Data Streams,” in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Shanghai, 2018, pp. 689–692.
[49] T. Weber, H. Hußmann, and M. Eiband, “Quantifying the Demand for Explainability,” in *Springer eBook Collection*, vol. 12933, *Human-Computer Interaction – INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30 – September 3, 2021, Proceedings, Part II*, C. Ardito et al., Eds., 1st ed., Cham: Springer International Publishing; Imprint Springer, 2021, pp. 652–661.
[50] X. Wang and M. Yin, “Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons,” *ACM Trans. Interact. Intell. Syst.*, 2022, doi: 10.1145/3519266.
[51] G. Dove, K. Halskov, J. Forlizzi, and J. Zimmerman, “UX Design Innovation: Challenges for working with machine learning as a design material,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA, 2017, pp. 278–288.
[52] B. Pokorni, M. Braun, and C. Knecht, “Menschenzentrierte KI-Anwendungen in der Produktion: Praxiserfahrungen und Leitfaden zu betrieblichen Einführungsstrategien,” Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO, 2021.
[53] G. Siaterlis *et al.,* “An IIoT approach for edge intelligence in production environments using machine learning and knowledge graphs,” *Procedia CIRP*, vol. 106, pp. 282–287, 2022, doi: 10.1016/j.procir.2022.02.192.
[54] X. Li, Y. Koren, and B. I. Epureanu, “Complementary learning-team machines to enlighten and exploit human expertise,” *CIRP Annals*, vol. 71, no. 1, pp. 417–420, 2022, doi: 10.1016/j.cirp.2022.04.019.