The Contribution of XAI for the Safe Development and Certification of AI: An Expert-Based Analysis

Benjamin Fresz[*†‡] **Vincent Philipp Göbels**[§] **Safa Omri**[§] **Danilo Brajovic**[*†]

Andreas Aichele[*] **Janika Kutz**[§] **Jens Neuhüttler**[§] **Marco F. Huber**[*†]

Abstract

Developing and certifying safe—or so-called trustworthy—AI has become an increasingly salient issue, especially in light of recently introduced regulation such as the EU AI Act. In this context, the black-box nature of machine learning models limits the use of conventional avenues of approach towards certifying complex technical systems. As a potential solution, methods to give insights into this black-box—devised in the field of eXplainable AI (XAI)—could be used. In this study, the potential and shortcomings of such methods for the purpose of safe AI development and certification are discussed in 15 qualitative interviews with experts out of the areas of (X)AI and certification. The interview results are summarized as a set of recommendations for policy makers and XAI researchers and developers. Overall, XAI methods are found to be a helpful asset for safe AI development, as they can show biases and failures of machine learning models, but since certification relies on comprehensive and correct information about technical systems, their impact is expected to be limited.

1 Introduction

In the rapidly evolving domain of machine learning (ML), the integration of ML systems into safety-critical applications presents unique challenges, primarily due to the ML-inherent opacity. Often characterized as "black-box" systems, such models are based on learning patterns instead of being explicitly programmed, thus complicating transparency and reliability Castelvecchi [2016]. This opacity not only challenges their integration into environments where safety is paramount but also impedes established system certification processes, which shape the available technology and how people interact with it. However, the scope of safety concerns related to AI systems extends beyond physical harms, encompassing broader sociotechnical risks such as algorithmic bias, discrimination, and societal inequalities. Therefore, the design of new assessment techniques for AI systems is of societal importance.

The field of eXplainable Artificial Intelligence (XAI) seeks to address the black-box problem by improving the transparency of ML models [Rai, 2020]. XAI aims to make the decision-making processes of AI systems comprehensible to human stakeholders, thereby increasing the trustworthiness of AI systems and facilitating their integration into regulated domains [Martinie, 2021, Brajovic et al., 2023]. As of now, the role of XAI for the assessment of AI and in other legally relevant use cases is part of an ongoing discussion, as showcased in legal cases around the world, e.g., in the case ACCC

[*]Fraunhofer Institute for Manufacturing Engineering and Automation IPA, Stuttgart, Germany
[†]Institute of Industrial Manufacturing and Management, University of Stuttgart, Germany
[‡]Corresponding author: `benjamin.fresz@ipa.fraunhofer.de`
[§]Fraunhofer Institute for Industrial Engineering IAO, Heilbronn, Germany

Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

vs. Trivago in Australia [Fraser et al., 2022] or multiple cases before the Court of Justice of the European Union [CJEU, 07.12.2023, 27.02.2025]. Additionally, new legislation such as Article 86 of the AI Act requires explanations for users. Although some publications discuss the legal requirements for XAI, they remain on a theoretical level [Fresz et al., 2024, Bringas Colmenarejo et al., 2025], and the utility of XAI in improving the safety and assessment (and thus certification) of AI systems has not been thoroughly investigated empirically. This paper addresses this gap by examining the potential and limitations of XAI with respect to the certification of AI systems. To the best of the authors' knowledge, this paper offers the first in-depth exploration of how XAI can be utilized in the certification and safeguarding of AI systems, evaluating the practical utility of XAI tools through the experiences of practitioners. Specifically, this paper addresses three primary questions:

1. What are the positive and negative experiences of practitioners using XAI in the field?
2. What is and could be the function of XAI in the development of safe AI?
3. Is it feasible to incorporate XAI into existing and future certification frameworks for AI systems?

To answer these questions, qualitative interviews with 15 experts, who operate at the intersection of AI development and certification, are conducted. While some of the criticisms levied towards XAI methods themselves in these interviews are not new, this paper adds to the aforementioned discussion of XAI for certification and legally relevant uses.

Note that XAI as a research field is criticized for its lack of clear foundations and even for a lack of scientific rigor [Weber et al., 2024]. One of those lacking foundations are vague and differing definitions of the terms "explainability" (as in XAI) and "interpretability" (as in interpretable ML, iML), often complicated by different disciplines having differing definitions [Miller, 2019, Weber et al., 2024]. Additionally, in the social sciences, explanations entail more than what is offered by most current XAI methods, e.g., aspects such as context- and user-dependence and interactivity of explanations [Miller, 2019, Liao et al., 2020, Rohlfing et al., 2021, Weber et al., 2024]. In this paper, the focus is on "technical" or "algorithmic" explainability [Weber et al., 2024], as the underlying XAI methods will limit what is possible with explainability in certification. Appendix A provides relevant background knowledge of XAI and related techniques. Within the "technical" explainability, a broad view on explainability methods is taken to get an overview of the entire field. Where applicable, not only the broad term XAI but more specific descriptions—based on the definitions in Appendix A—are used.

The paper is organized as follows: After an introduction into the safe development of technical systems, the related works are presented, with a focus on the context of certification of AI systems. In Section 3, the methodology and participant profiles are introduced, followed by the presentation of the interview results in Section 4. Further pathways for XAI and limitations of the used approach are discussed in Section 5. After summarizing the previous results in the form of recommendations for XAI developers and researchers, and policy makers in Section 6, the paper closes with a summary in Section 7.

2 Related Works

The following sections present an overview of certification processes for safe technical systems and the current challenges of AI certification. Although the certification of AI systems is not standardized as of now, multiple scientific publications provide potential avenues of approach. Some of these are discussed in Sections 2.2 and 2.3, to outline the issues and potential solutions for safe AI development and certification.

2.1 Safe Development of Technical Systems

For non-AI products, safe development and certification processes are well established, as they are subject to numerous legal and standardization requirements. For example, the Machinery Directive 2006/42/EC of the European Union regulates the provisions for placing machinery on the market in the European Economic Area. A key point of this directive is the minimum requirements for safety and health protection. Specific requirements are derived from references to corresponding harmonized standards. For technical systems with AI functionalities, which are the focus of this paper, the area of electrical, electronic and programmable electronic systems is most likely to apply. If a system in this area is developed with a safety function, IEC 61508 describes a process model, methods to be used, and various required activities and work products. The basic procedure is to identify potential situations that pose a risk to life and limb. The relevance or dangerousness of situations is determined by means of a risk assessment. In the case of particularly dangerous situations, further methods must be used to avoid systematic errors. In the case of random faults, a quantitative assessment of the components with a maximum permissible probability of failure is required. Companies and/or products are certified to confirm compliance with these requirements. An independent body checks whether the requirements specified in the standard(s) have been met and provides a certificate of the system's conformity to the standard(s). Specific standards, such as ISO 21448 (Road vehicles - Safety of the intended functionality), already recommend analyzing the interpretability of ML software to increase its trustworthiness.

2.2 AI Certification Challenges

The certification of AI systems presents complex challenges, as highlighted by various publications [Falcini and Lami, 2017, Stoica et al., 2017, Vanderlinde et al., 2022, Mahilraj et al., 2023, Anisetti et al., 2023, Winter et al., 2021, Levene and Wooldridge, 2023]. Certifying AI systems is particularly difficult due to factors that diverge from traditional software certification. The discussion around AI assessment (or auditing) and the corresponding certification is summarized below.

Falcini and Lami [2017] emphasize the need for new certification schemes, in this example in the automotive industry, while Stoica et al. [2017] stress the importance of AI systems capable of making safe decisions in unpredictable environments. The growing body of work on AI auditing, as noted by Birhane et al. [2024], suggests that audits can serve as a meaningful accountability mechanism for AI systems [Levene and Wooldridge, 2023, Anisetti et al., 2023]. Recommendations from Costanza-Chock et al. [2022] advocate independent algorithmic audits to ensure compliance with defined standards.

Further studies highlight the need for new approaches to tackle AI certification challenges. Vanderlinde et al. [2022] focus on potential solutions, while Mahilraj et al. [2023] discuss issues of robustness, transparency, reliability, and safety. One proposed assessment scheme for low-risk applications is outlined by Winter et al. [2021].

Two significant challenges in AI certification are *data dependency* and *dynamic behavior*. The quality of training data directly influences AI performance [Landgrebe, 2022]. Ensuring datasets are relevant, representative and unbiased is complex and differs from traditional software validation.The *dynamic behavior* of AI systems that learn post-deployment, introduces unpredictability. This contrasts with traditional software, where behaviors can be tested and certified against fixed specifications. To address this, Bakirtzis et al. [2023] propose a dynamic certification approach involving iterative testing and revision of use-context pairs. Establishing flexible and robust certification processes that monitor AI system changes over time remains a key challenge [Stodt et al., 2023].

2.3 XAI and its Role in Certification

Regarding the previously described challenges of safe AI development, several publications propose XAI as a potential solution. Some of them and the ongoing discourse are presented in the following.

Gyevnar et al. [2023] coin the term "transparency gap" as the fundamental discrepancy between XAI's narrow focus on algorithmic explanation—treating transparency as an end in itself—and the broader legal perspective, like that in the AI Act, which views transparency as a means to achieve accountability, human rights, and other societal values. To bridge this gap the authors call for clearly defining and scoping transparency to tailor explanations according to risk and stakeholder needs. They also advise clarifying the legal status of XAI so that XAI tools are integrated with the underlying AI systems, supported by unified conformity assessments and standardized documentation practices. For that, Brajovic et al. [2023] describe a framework for the documentation of AI (as precursor to certification), based on model cards [Mitchell et al., 2019] and data cards [Pushkarna et al., 2022], including XAI as a potentially necessary part of development. A similar notion is provided by Martinie [2021], who views XAI as key to make AI in critical interactive systems transparent to users and certification stakeholders. An application for these use cases is shown by Saraf et al. [2020], as they develop a proof of concept tool to generate local explanations for a trajectory anomaly detection model to demonstrate how XAI can help towards user acceptance and certification. To be able to assess transparency in the safe development and certification, robust measures for XAI need to be developed and integrated into AI assessment [Stodt et al., 2023].

Several studies test XAI methods for use cases other than certification, but possibly with implications for AI certification as well. Fostering user trust in an AI application with the help of XAI is a currently discussed topic, as there exist other influences on human trust in AI models, e.g., model performance [Papenmeier et al., 2019] or task complexity [Vered et al., 2023]. Ideally, XAI could allow users to "calibrate" their trust, i.e., to accept correct decisions and reject false ones [Turner et al., 2020, Vered et al., 2023, Zhang et al., 2020, Ma et al., 2023]. Some works criticize explanations fostering trust when they are incorrect or not informative, thus increasing—instead of decreasing—automation bias [Kim et al., 2022, Schemmer et al., 2023, Ehsan et al., 2024, Eiband et al., 2019]. Such problems are even more relevant in the context of certification, as malicious actors might want to cheat the certification procedure by providing intentionally misleading explanations and thus hiding biases within an AI system [Zhou and Joachims, 2023].

Further criticism of XAI approaches as certification aid has emerged, with Landgrebe [2022] noting a shift from the initial goal of providing objective understanding of ML models. They suggest "certified AI" as an alternative, emphasizing specification, realization, and tests, incorporating ontology and formal logic. Additionally, Henriksen et al. [2021] argued in 2021 that XAI-generated explanations fail to meet their intended role in policy documents, tying back to the discussion about legal requirements for XAI as touched on in Section 1.

A popular use case for XAI is model debugging, i.e., detecting biases [Achtibat et al., 2023, Adebayo et al., 2020, 2022, Colin et al., 2022, Fel et al., 2023, Lin et al., 2021]. In user studies, several influences on the performance of the participants to detect biases have been reported, e.g., the choice of explanation method [Achtibat et al., 2023], whether the type of bias is known beforehand [Adebayo et al., 2022], or the specific type of bias in the data [Adebayo et al., 2020]. While several surveys on the topic exist, they mostly report unclear or mixed results about the ability of study participants to spot biases in ML models via XAI methods [Schemmer et al., 2022, Müller, 2024, Fok and Weld, 2024, Kandul et al., 2023, Rong et al., 2023]. Due to these mixed results, it is particularly interesting to see whether practitioners and certification experts expect the potential benefits of XAI methods to be realized in practice.

While the utility of XAI in enhancing transparency is often recognized, there is a notable gap in empirical research concerning its integration into the certification processes of AI systems. Most existing studies focus on theoretical frameworks or specific use case scenarios, with less emphasis on systematic, empirical evaluations of XAI's role in the broader certification processes [Landgrebe, 2022, Brajovic et al., 2023]. While Gyevnar et al. [2023] identify a transparency gap, real-world insights on how XAI methods support legal transparency, conformity assessment, and human oversight are crucial in further assessing the size of the gap. This study aims to examine the firsthand experiences of XAI in practice and evaluate its potential and limitations in the context of AI certification.

3 Methodology

To survey the potential of XAI in general and in certification processes in particular, 15 interviews were conducted. Participants were recruited from the fields of industry, academia and consulting, with the proportions shown in Figure 1. Potential participants were taken into account based on publicly funded research projects and involvement in standardization bodies and AI-focused networks. Additionally, snowball sampling was used, with potential participants being able to recommend further experts. All interviews were conducted on a voluntary basis without reimbursement. In the following, the methodology for the interviews is described, including the participant selection and profiles and the interview. The coding process based on Mayring [2019] can be found in Appendix B.

3.1 Participant Profiles

To limit culture-specific influences, all participants either originated from Germany, Austria, or Switzerland or live there permanently. Most interviews were conducted in German (10), and some in English (5). Inclusion criteria required knowledge about both AI certification and XAI, established through current projects or published works. Participants had to have experience working with AI,

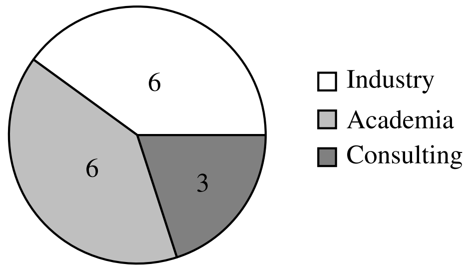

Figure 1: Working fields of interview participants (n=15). Participants reporting multiple fields were classified by their self-reported main field.

experimenting with XAI, and be actively involved in certification processes. Most participants had more than four years of experience in XAI (minimum two years) and two years in AI certification (minimum 1.5 years). Due to the specific expertise requirements, purposive sampling [Guest et al., 2014] was used. An anonymized list of participants and their expertise is found in Table 2 in Appendix D. The backgrounds of the participants span the fields of psychology, neuro-science, computer science, finance and engineering to ensure a broad perspective on XAI.

3.2 Interview Process

The interviews for this study were conducted from January to March of 2024 via Microsoft Teams and mostly lasted between 30 and 60 minutes. Interviews were conducted by two researchers: one led the conversation while the other followed up on relevant points and remarks from the interviewee. All interviews were recorded and transcribed with the consent of the participants. After the agreement of participants to participate in the study, the general outline of the interviews was presented, including the following structure (full list of interview questions in Appendix C):

- **Participant profile:** Participants were asked about their current position, their previous work and their current task relating to AI.
- **Use of (X)AI:** Participants were invited to explain their current aim and use of XAI, since challenges and the state of the art might differ between the aim and field of use.
- **XAI in certification:** After participants spoke about their experience with XAI in general, they were asked about specific challenges and requirements for XAI in the field of certification of AI.
- **Look into the future:** Since most of the previous questions focused on challenges in the field of XAI, participants were invited to share their thoughts and hopes regarding the future development of XAI.

4 Results

Among the 15 expert interviews conducted, certain findings were consistently agreed upon by 13 participants, which are presented below. For these key findings, data saturation can be argued, as suggested by Hennink and Kaiser [2022], who note that saturation often occurs within 9 to 17 interviews for homogeneous populations and narrow objectives.

In addition to the standard questions, the semi-structured format of the interviews allowed for the exploration of other remarks and engaging discussions. These insights, while valuable, as they offer unique viewpoints not commonly found in existing literature, do not reach data saturation and are thus presented separately in Appendix E.

4.1 Use of (X)AI

4.1.1 Aim of XAI Use

The participants of the study unanimously thought of transparency or explainability as an important topic in safe AI development. This could also be explained by selection bias, as all participants are working on related topics (see Section 3.1). While they considered transparency and explainability as important, a common thread that emerged is that the integration of appropriate methods into standard development processes is still lacking in most cases. This shortcoming is partly due to the perceived lack of sufficient value-addition from explainability to warrant the necessary funding, especially when AI projects are externally commissioned. When applied, the objectives of explainability methods are multifaceted and often abstract, encompassing aspects such as enhancing the public perception of projects, adhering to regulatory or customer requirements, detecting errors in ML systems, and facilitating internal communication about the capabilities and operation of ML methods across different departments (e.g., compliance and ethics checks).

4.1.2 Choice of XAI Method

With the aims of XAI use varying significantly, a clear framework for assessing the performance of XAI methods was not discernible from the interviews, also due to the wide range of XAI methods employed. These span from neuro-symbolic ML systems, graph- and concept-based explanations to white-box models, and feature importance methods like SHAP and LIME. The multitude of available methods and the ambiguity in evaluating the respective objectives make it challenging for practitioners to identify the most suitable method for a particular application. Consequently, methods that are easy to implement and provide accessible information are often chosen. Due to its open-source nature and ease of use, SHAP is commonly used, although the interviewees are aware of the criticisms levied at this method, e.g., by Slack et al. [2020], Kumar et al. [2020], and thus sceptical of its performance and reliability.

4.1.3 Experiences with XAI

Despite the challenges described before, the interviewees reported successful applications of XAI procedures, particularly in identifying errors in existing ML systems, conducting plausibility checks on models during development, and enhancing data understanding. However, it was also noted that current XAI methods are not well-suited for all use cases, with projects often failing due to common reasons. These included the incomprehensibility of generated explanations to the target audience, lack of time for experts to interact with the explanations, and difficulty in verifying found correlations due to insufficient AI or domain expertise. Examples for explanations being incomprehensible to the target audience included knowledge graph explanations being too complex for lay users and saliency map explanations being not discriminative enough for physicians. Further complicating the deployment of XAI is the unstable nature of some methods, leading to non-reproducible results, and the lack of comprehensive research on methods for specific data types like time series.

The use of XAI methods to foster (or "calibrate") trust among end-users, often highlighted in scientific literature, was viewed critically in many interviews. The complexity of XAI methods effectively shifts the problem of an untrustworthy black-box (the ML system) to another black-box (the XAI method), the trustworthiness of which is also questioned due to the controversial nature of existing XAI methods. This issue is exemplified by the disagreement problem [Krishna et al., 2022], where different XAI methods provide different explanations for a single decision of an ML model, making it unclear what the "true" explanation is. Note that the black-box nature of the XAI method stems more from the lack of in-depth expertise about XAI than from a general incomprehensibility such as for the ML model. Because of this, a potential solution mentioned by P2 to the trust issue created by the double black-box is the provisioning of training on AI and XAI for users, resulting in "calibrated" trust.

4.2 XAI in Certification

Additional to XAI in development, interviewees were asked about their expectations and perceived challenges of XAI in AI certification. Central to this discussion is the challenge of measuring "appropriate" transparency and explainability in XAI methods (as demanded by the AI Act), a task that varies significantly depending on the specific purpose and function of the AI system in question.

Overall, two main groups of opinions about the use of XAI in certification can be distinguished: From the perspective of some experts, the influence of XAI on the certification of AI systems is seen to be minor. This viewpoint stems from the existence of other regulatory measures such as thresholds for certain performance metrics for AI systems or the belief that XAI methods, particularly in complex applications, are not and cannot be sufficiently robust or comprehensive. In such applications, explanations generated by XAI methods themselves become too intricate, thus detracting from their utility. This viewpoint is underpinned by the interviewees almost unanimously agreeing that explainability is not (and will not be) truly measurable, or will at least require user studies to do so.

In contrast, other experts—especially ones who successfully used XAI in the past—maintain that XAI has demonstrated its potential in improving ML models by identifying problems early in the development process. As the main goal of safeguarding and certifying AI is to prevent potentially harmful defects, it is argued that XAI—even without quantitative performance metrics—helps towards that goal and should thus be part of AI certification. Interviewees with this viewpoint often additionally pointed out that XAI could only be one of many tools for AI certification (additional to testing, formal verification, etc.), providing only a small piece of evidence for certification. Irrespective of their expectations for the future use of XAI in certification processes, participants emphasized that human inspectors remain integral to the certification process.

4.3 Expectations for XAI

Looking towards the future, the expectations and hopes associated with the development of XAI are diverse. While the ideal of achieving complete, globally applicable explanations is largely seen as unattainable, some optimism persists around the evolution of new XAI approaches. These include concept-based, mechanistic, and neuro-symbolic methods, which are hoped to enable a new form of explanations elucidating the fundamental operation of ML models.

The interviewees also highlighted the necessity of user-centric and industry-focused approaches in order to fully realize the potential of XAI. While XAI methods are seen as offering the capacity to detect errors in ML systems, and thus should ideally be integrated into the development processes, other methods are expected to be of higher importance for the certification landscape. These include AI examination by alternate AI systems, formal verification of specific properties, and uncertainty quantification of AI decisions. A major challenge for the explainability of AI systems is seen in the difference of new AI paradigms, as future Large Language Models (LLMs) might provide multi-modal inputs and outputs and explanations for e.g. time series or image data need to be fundamentally different than ones for other data types.

A recurring theme across discussions was the call for clear, definitive requirements for AI certification, such as specific metrics. Without such clear guidance, the interviewees felt that companies would lack the available resources and information to ensure that their AI systems comply with the relevant transparency requirements, such as those in the AI Act. This call for clear requirements is complemented by the advocacy for the use of simple, intrinsically interpretable AI solutions, wherever possible. The use of AI in high-risk applications was considered inappropriate in general by one interviewee, while others argued for the use of white-box models where possible, emphasizing the need for caution and discretion in AI deployment.

5 Discussion

The conducted expert interviews illuminate potential pathways for the advancement of XAI, which will be explored in this section. Additionally, constraints and limitations inherent in the study's design are addressed in Section 5.5.

5.1 Integration of Diverse Expert Perspectives

This research integrates insights from experts with dual expertise in XAI and certification. The diversity in expertise and backgrounds enriched the analysis, providing a well-rounded understanding of both the potential and limitations of XAI in AI certification processes. While the initial expectations anticipated these insights, the nuanced opinions offered by participants exceeded the predictions, underscoring the complex interplay between XAI capabilities and certification standards. Due to the required expertise, only a limited number of participants could be interviewed. While more participants might have provided additional insights, the main opinions of using XAI as an incomplete debugging tool or not at all in certification converged. With the sample size of 15 interviews and based on Hennink and Kaiser [2022], this can be used to argue for data saturation for research question 3 "Is it feasible to incorporate XAI into existing and future certification frameworks for AI systems?".

5.2 XAI in Certification: From Debugging Tool to Certification Aid

XAI's role in certification could be pivotal but is constrained by several factors. As of now, there is a critical gap between the theoretical advantages of XAI and its practical utility in ensuring compliance with stringent certification protocols. As a debugging tool, XAI already provides valuable insights into AI behavior, identifying biases and failure points. However, transitioning from debugging to a certification context requires XAI to offer more definitive guarantees (or at least information in the form of confidence estimates) about the correctness of explanations and AI systems' behaviors and outcomes, a transition that is currently underdeveloped.

5.3 XAI as a Requirement

This paper considered XAI as a tool for certification. In practice, XAI could also become a requirement for AI systems, potentially through Article 14 "Human oversight" and Article 86 "Right to explanation of individual decision-making" of the AI Act. For these, standards should specify clear requirements for the implementation of XAI. Based on current literature, that seems rather difficult, thereby potentially decreasing the practical use of these articles [Nnawuchi and George, 2024].

5.4 Societal and Ethical Considerations

The discourse around XAI goes beyond the technical boundaries and touches on the broader societal and ethical implications. Current certification frameworks primarily address technical compliance, but the integration of XAI requires a broader consideration of ethical standards and societal impacts. This requires a paradigm shift in certification, from purely technical evaluations to more holistic assessments that consider the societal implications of AI technologies.

5.5 Limitations

While this study provides valuable insights into the role of XAI in the certification and safe development of AI systems, several limitations need to be acknowledged:

The primary limitation pertains to sample characteristics. While the qualitative methodology yielded in-depth insights from 15 experts, all participants were recruited from the DACH-region (Germany, Austria, Switzerland) through professional networks. This geographic focus ensured consistency within a shared regulatory context but potentially restricts the generalizability of findings to other regions with differing certification frameworks. Furthermore, the purposive sampling approach, while essential for targeting specialized expertise, may have introduced selection bias.

A second limitation involves stakeholder representation. The participant pool primarily comprised technical experts actively engaged in AI certification processes. Although their expertise provided valuable technical insights, this composition underrepresents critical perspectives from end-users impacted by certification decisions, policymakers and regulators shaping the frameworks, and organizations implementing AI certification without deep technical proficiency. Consequently, the findings may lean toward an implementation-focused perspective, leaving broader policy and user-centric considerations less explored.

Finally, the qualitative nature of this study, while facilitating an in-depth exploration of expert perspectives, introduces inherent subjectivity. Findings rely on individual experiences and interpretations, which may not comprehensively represent broader certification contexts or objective measures of XAI's effectiveness. Additionally, opinions on XAI are shaped over a longer period of time, potentially misrepresenting the most up-to-date developments in such a highly dynamic field. While this subjectivity enriches the understanding of practitioner viewpoints, it underscores the need for complementary approaches to validate the insights.

6 Recommendations for XAI Researchers and Developers

In the following, the previous results and discussion points are condensed into recommendations on how to shape the future development and use of XAI. Since these recommendations differ based on one's role in relation to XAI, they are separated into recommendations for XAI developers and researchers, and recommendations for policy makers.

6.1 Recommendations for XAI Developers and Researchers

The conducted interviews point towards common challenges in the development of XAI systems. While most of these challenges are known in the scientific literature, they still pose problems for the real-world use of XAI, mostly pertaining to how and when XAI is used in the development process. Thus, the following recommendations can be given to XAI developers and researchers:

1. Make sure you understand the explanation requirements of your use case.
2. Integrate explainability and explanation requirements as early as possible into the development process.
3. When designing and using XAI methods, state clearly what their purpose is and what the expected benefits of the methods are (e.g., based on [Sokol and Flach, 2020]).
4. When designing and using XAI methods and evaluation metrics (e.g, from [Pawelczyk et al., 2021, Agarwal et al., 2024, Monke et al., 2025], be aware of their assumptions.
5. Be aware of the inherent interdisciplinarity of XAI.
6. Get involved into standardization processes.

6.2 Recommendations for Policy Makers

As touched on in Section 1, there is an ongoing discussion about the legal necessity for and benefits of XAI. This paper aims to provide some input for such a discussion, with the following points condensing the interview results:

1. Be aware of the Work in Progress status of XAI, especially if XAI is to be used in legislation.
2. Make sure new norms leave room for potential future changes of XAI.
3. If XAI should be used: Provide clear explanation requirements, in contrast to current legislation such as Article 86 of the AI Act.
4. Realize that XAI has its use, while not being a comprehensive solution to AI certification or transparency.
5. Make the certification of AI an interdisciplinary effort.

7 Summary

As the field of AI continues to evolve, the adaptability of certification processes—and the role of XAI within these—will be paramount. XAI is often touted as a potential solution to the black-box nature and thus, certification, of AI. This notion was examined empirically in this paper via qualitative interviews with 15 experts both in XAI and AI certification. In these interviews, the current state of XAI was often viewed skeptically due to the known problems of such methods and the overall difficulty of using more complex explanation techniques. Despite that, the interviewees often came up with examples where they used XAI successfully. In these, it showed that XAI is able to highlight errors in ML applications, while it does not seem well suited to provide simple and understandable explanations to end users or domain experts. Based on the shortcomings of current XAI methods, the interviewees largely expect XAI to be at most a helpful asset in AI certification, but no comprehensive answer for the associated difficulties. The interviewees also highlighted further avenues for XAI research, especially into data types for which XAI methods are less common like time series and natural language and new explanation types like concept-based and multi-modal explanations.

Looking ahead, the integration of XAI into certification processes poses significant challenges and opportunities. The evolving regulatory landscape, particularly with frameworks like the EU AI Act, will likely include explainability as a core component. However, the absence of standardized measures for assessing the sufficiency of explainability complicates this integration. To be able to integrate XAI into certification processes, practitioners need clear guidance on which XAI method should be used when and future research must focus on developing robust, quantifiable metrics for (X)AI that align with certification standards and contribute effectively to the safety and reliability of AI systems.

Acknowledgements

We thank all reviewers of the previous versions of this paper, who helped to improve it with their valuable feedback.

This paper is funded in parts by the German Federal Ministry for Economic Affairs and Climate Action under grant no. 19A21040B (project "veoPipe") and by the Fraunhofer Gesellschaft under grant no. PREPARE 40-02702 (project "ML4Safety").

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 15662535. doi: 10.1016/j.inffus.2021.05.008.

Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023. doi: 10.1038/S42256-023-00711-8. URL `http://arxiv.org/pdf/2206.03208`.

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Julius Adebayo, Michael Muelly, Hal Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation, 2022. URL `https://arxiv.org/abs/2212.04629`.

Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. Evaluating explainability for graph neural networks. *Scientific data*, 10(1):144, 2023. doi: 10.1038/s41597-023-01974-x.

Chirag Agarwal, Dan Ley, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations, 2024. URL `https://arxiv.org/abs/2206.11104`.

Marco Anisetti, Claudio A. Ardagna, Nicola Bena, and Ernesto Damiani. Rethinking certification for trustworthy machine-learning-based applications. *IEEE Internet Computing*, 27(6):22–28, oct 2023. ISSN 1089-7801. doi: 10.1109/MIC.2023.3322327. URL `https://doi.org/10.1109/MIC.2023.3322327`.

Georgios Bakirtzis, Steven Carr, David Danks, and Ufuk Topcu. Dynamic certification for autonomous systems. *Communications of the ACM*, 66(9):64–72, August 2023. ISSN 0001-0782. doi: 10.1145/3574133. URL `https://doi.org/10.1145/3574133`.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024.

Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. AI auditing: The Broken Bus on the Road to AI Accountability . In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 612–643, Los Alamitos, CA, USA, April 2024. IEEE Computer Society. doi: 10.1109/SaTML59370.2024.00037. URL `https://doi.ieeecomputersociety.org/10.1109/SaTML59370.2024.00037`.

Danilo Brajovic, Niclas Renner, Vincent Philipp Goebels, Philipp Wagner, Benjamin Fresz, Martin Biller, Mara Klaeb, Janika Kutz, Jens Neuhuettler, and Marco F. Huber. Model reporting for certifiable ai: A proposal from merging eu regulation into ai development, 2023.

Alejandra Bringas Colmenarejo, Laura State, and Giovanni Comandé. How should an explanation be? a mapping of technical and legal desiderata of explanations for machine learning models. *International Review of Law, Computers & Technology*, pages 1–32, 2025. ISSN 1360-0869. doi: \url{10.1080/13600869.2025.2497633}.

Davide Castelvecchi. Can we open the black box of ai? *Nature*, 538:20–23, 2016. URL `https://api.semanticscholar.org/CorpusID:4465871`.

CJEU. Schufa holding (scoring), 07.12.2023. Judgement, C-634/21, ECLI EU:C:2023:957.

CJEU. Dun & bradstreet austria, 27.02.2025. Judgement, C-203/22, ECLI EU:C:2025:117.

Julien Colin, Thomas FEL, Remi Cadene, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 2832–2845. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/13113e938f2957891c0c5e8df811dd01-Paper-Conference.pdf`.

Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1571–1583, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533213. URL `https://doi.org/10.1145/3531146.3533213`.

Richard A. DeMillo, Richard J. Lipton, and Frederick G. Sayward. Hints on test data selection: Help for the practicing programmer. *IEEE Computer*, 11(4):34–41, 1978.

Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. The who in xai: How ai background shapes perceptions of ai explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642474. URL `https://doi.org/10.1145/3613904.3642474`.

Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich 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*, CHI EA '19, page 1–6, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359719. doi: 10.1145/3290607.3312787. URL `https://doi.org/10.1145/3290607.3312787`.

Richard Evans, Matko Bošnjak, Lars Buesing, Kevin Ellis, David Pfau, Pushmeet Kohli, and Marek Sergot. Making sense of raw input. *Artificial Intelligence*, 299:103521, 2021. ISSN 00043702. doi: 10.1016/j.artint.2021.103521. URL `https://www.sciencedirect.com/science/article/pii/S0004370221000722`.

Fabio Falcini and Giuseppe Lami. Challenges in certification of autonomous driving systems. In *2017 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 286–293, 2017. doi: 10.1109/ISSREW.2017.45.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadénc, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2711–2721, 2023. doi: 10.1109/CVPR52729.2023.00266.

Raymond Fok and Daniel S. Weld. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. *AI Mag.*, 45(3):317–332, July 2024. ISSN 0738-4602. doi: 10.1002/aaai.12182. URL `https://doi.org/10.1002/aaai.12182`.

Henry Fraser, Rhyle Simcock, and Aaron J. Snoswell. Ai opacity and explainability in tort litigation. In Timo Speith, editor, *A review of taxonomies of explainable artificial intelligence (XAI) methods*, pages 185–196, New York and Saarbrücken, 2022. ACM and Saarländische Universitäts- und Landesbibliothek. ISBN 9781450393522. doi: 10.1145/3531146.3533084.

Benjamin Fresz, Elena Dubovitskaya, Danilo Brajovic, Marco F. Huber, and Christian Horz. How should ai decisions be explained? requirements for explanations from the perspective of european law. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):438–450, Oct. 2024. doi: 10.1609/aies.v7i1.31648. URL `https://ojs.aaai.org/index.php/AIES/article/view/31648`.

Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic ai: the 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, 2023. ISSN 0269-2821. doi: 10.1007/s10462-023-10448-w.

Greg Guest, Elizabeth E. Tolley, and Christina M. Wong. Qualitative research methods. *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society*, pages 1947–1952, 2014.

Balint Gyevnar, Nick Ferguson, and Burkhard Schafer. *Bridging the Transparency Gap: What Can Explainable AI Learn from the AI Act?* IOS Press, September 2023. ISBN 9781643684376. doi: 10.3233/faia230367. URL `http://dx.doi.org/10.3233/FAIA230367`.

Monique Hennink and Bonnie N Kaiser. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social science & medicine*, 292:114523, 2022. doi: https://doi.org/10.1016/j.socscimed.2021.114523.

Anne Henriksen, Simon Enni, and Anja Bechmann. Situated accountability: Ethical principles, certification standards, and explanation methods in applied ai. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 574–585, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462564. URL `https://doi.org/10.1145/3461702.3462564`.

Lukas-Valentin Herm. Impact of explainable ai on cognitive load: Insights from an empirical study, 2023.

Daniel Kahneman. *Thinking, fast and slow*. Penguin psychology. Penguin Books, London, 2012. ISBN 9780141033570.

Serhiy Kandul, Vincent Micheli, Juliane Beck, Markus Kneer, Thomas Burri, Francois Fleuret, and Markus Christen. Explainable ai: A review of the empirical literature. *SSRN Electronic Journal*, 01 2023. doi: 10.2139/ssrn.4325219.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/kim18d.html`.

Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. Hive: Evaluating the human interpretability of visual explanations. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 280–298, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19775-8.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective, 2022.

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle A. Friedler. Problems with shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Matthew Landers and Afsaneh Doryab. Deep reinforcement learning verification: A survey. *ACM Computing Surveys*, 55(14s):1–31, 2023. ISSN 0360-0300. doi: 10.1145/3596444.

Jobst Landgrebe. Certifiable ai. *Applied Sciences*, 12(3):1050, 2022. ISSN 2076-3417. doi: 10.3390/app12031050. URL `https://www.mdpi.com/2076-3417/12/3/1050/pdf?version=1642673131`.

Martin L. Levene and Jeffrey Wooldridge. Certification of machine learning applications in the context of trustworthy ai with reference to the standardisation of ai systems, March 2023. URL `http://eprintspublications.npl.co.uk/9683/`.

Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376590. URL https://doi.org/10.1145/3313831.3376590.

Yi-Shan Lin, Wen-Chuan Lee, and Z. Berkay Celik. What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1027–1035, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467213. URL https://doi.org/10.1145/3447548.3467213.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017. URL http://arxiv.org/abs/1705.07874.

Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581058. URL https://doi.org/10.1145/3544548.3581058.

Jenifer Mahilraj, M. Satheesh Pandian, Muthuraman Subbiah, Sanjna Kalyan, Vadivel R, and Nirmala S. Evaluation of the robustness, transparency, reliability and safety of ai systems. *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 1: 2526–2535, 2023. doi: 10.1109/ICACCS57279.2023.10113057.

Célia Martinie. Challenges for operationalizing XAI in Critical Interactive Systems. In *ACM CHI Workshop on Operationalizing Human-Centered Perspectives in Explainable AI (HCXAI @ CHI 2021)*, Online Virtual Conference (originally Yokohama), Japan, May 2021. ACM. URL https://hal.science/hal-03221502. ACM Conference on Human Factors in Computing Systems (CHI).

Philipp Mayring. Qualitative inhaltsanalyse – abgrenzungen, spielarten, weiterentwicklungen: Forum qualitative sozialforschung / forum: Qualitative social research, vol 20, no 3 (2019): Qualitative content analysis i. 2019. doi: 10.17169/FQS-20.3.3343.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2018.07.007. URL https://www.sciencedirect.com/science/article/pii/S0004370218305988.

Tim Miller. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 333–342, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594001. URL https://doi.org/10.1145/3593013.3594001.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Helena Monke, Benjamin Sae-Chew, Benjamin Fresz, and Marco F. Huber. From confusion to clarity: Protoscore - a framework for evaluating prototype-based xai. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 2215–2231, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732151. URL https://doi.org/10.1145/3715275.3732151.

Romy Müller. How explainable ai affects human performance: A systematic review of the behavioural consequences of saliency maps. *International Journal of Human–Computer Interaction*, 0(0):1–32, 2024. doi: 10.1080/10447318.2024.2381929. URL https://doi.org/10.1080/10447318.2024.2381929.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.*, 55 (13s), July 2023. ISSN 0360-0300. doi: 10.1145/3583558. URL `https://doi.org/10.1145/3583558`.

Uchenna Nnawuchi and Carlisle George. A grand entrance without a blueprint: A critical analysis of the right to explanation in article 86 of the european union artificial intelligence act. *Journal of AI Law and Regulation*, 1(4), 2024. doi: \url{10.21552/aire/2024/4/6}.

Frederik Pahde, Maximilian Dreyer, Wojciech Samek, and Sebastian Lapuschkin. Reveal to revise: An explainable ai life cycle for iterative bias correction of deep models. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 596–606, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43895-0.

Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. How model accuracy and explanation fidelity influence user trust, 2019.

Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms, 2021. URL `https://arxiv.org/abs/2108.00783`.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1776–1826, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533231. URL `https://doi.org/10.1145/3531146.3533231`.

Arun Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48:137–141, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL `https://doi.org/10.1145/2939672.2939778`.

Katharina J. Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel-Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. Explanation as a social practice: Toward a conceptual framework for the social design of ai systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3):717–728, 2021. doi: 10.1109/TCDS.2020.3044366.

Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(4):2104–2122, nov 2023. ISSN 0162-8828. doi: 10.1109/TPAMI.2023.3331846. URL `https://doi.org/10.1109/TPAMI.2023.3331846`.

Aditya P. Saraf, Kennis Chan, Martin Popish, Jeff Browder, and John Schade. Explainable artificial intelligence for aviation safety applications. In *AIAA AVIATION 2020 FORUM*, Reston, Virginia, 2020. American Institute of Aeronautics and Astronautics. ISBN 978-1-62410-598-2. doi: 10.2514/6.2020-2881.

Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 617–626, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534128. URL `https://doi.org/10.1145/3514094.3534128`.

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 410–422, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701061. doi: 10.1145/3581641.3584066. URL `https://doi.org/10.1145/3581641.3584066`.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 180–186, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL `https://doi.org/10.1145/3375627.3375830`.

Kacper Sokol and Peter Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 56–67. ACM, January 2020. doi: 10.1145/3351095.3372870. URL `http://dx.doi.org/10.1145/3351095.3372870`.

Timo Speith. A review of taxonomies of explainable artificial intelligence (xai) methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, New York, NY, USA, 2022. ACM. ISBN 9781450393522. doi: 10.1145/3531146.3534639.

Diomidis H Stamatis. *Failure mode and effect analysis*. Quality Press, 2003.

Robert R. Sterling. A test of the uniformity hypothesis. *Abacus*, 5(1):37–47, 1969. ISSN 0001-3072. doi: 10.1111/j.1467-6281.1969.tb00159.x.

Jan Stodt, Christoph Reich, and Nathan Clarke. A novel metric for xai evaluation incorporating pixel analysis and distance measurement. In Anna Esposito, editor, *2023 IEEE 35th International Conference on Tools with Artificial Intelligence*, pages 1–9, Piscataway, NJ, 2023. IEEE. ISBN 979-8-3503-4273-4. doi: 10.1109/ICTAI59109.2023.00009.

Ion Stoica, Dawn Song, Raluca Ada Popa, David A. Patterson, Michael W. Mahoney, Randy H. Katz, Anthony D. Joseph, Michael Jordan, Joseph M. Hellerstein, Joseph Gonzalez, Ken Goldberg, Ali Ghodsi, David E. Culler, and Pieter Abbeel. A berkeley view of systems challenges for ai, Oct 2017. URL `http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.html`.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 6021–6029, Apr. 2020. doi: 10.1609/aaai.v34i04.6064. URL `https://ojs.aaai.org/index.php/AAAI/article/view/6064`.

Amy Turner, Meenakshi Kaushik, Mu-Ti Huang, and Srikar Varanasi. Calibrating trust in ai-assisted decision making, 2020. URL `https://www.ischool.berkeley.edu/projects/2020/calibrating-trust-ai-assisted-decision-making`.

Jake Vanderlinde, Kevin Robinson, and Benjamin Mashford. The challenges for artificial intelligence and systems engineering. *Australian Journal of Multi-Disciplinary Engineering*, 18(1):47–53, 2022. doi: 10.1080/14488388.2022.2044607.

Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. The effects of explanations on automation bias. *Artificial Intelligence*, 322:103952, 2023. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2023.103952. URL `https://www.sciencedirect.com/science/article/pii/S000437022300098X`.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31: 841–887, 04 2018. doi: 10.2139/ssrn.3063289.

Rosina O Weber, Adam J Johs, Prateek Goel, and João Marques Silva. Xai is in trouble. *AI Magazine*, 45(3):300–316, 2024. doi: https://doi.org/10.1002/aaai.12184. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12184.

Philip Matthias Winter, Sebastian Eder, Johannes Weissenböck, Christoph Schwald, Thomas Doms, Tom Vogt, Sepp Hochreiter, and Bernhard Nessler. Trusted artificial intelligence: Towards certification of machine learning applications, 2021. URL https://arxiv.org/abs/2103.16910.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 295–305, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372852. URL https://doi.org/10.1145/3351095.3372852.

Joyce Zhou and Thorsten Joachims. How to explain and justify almost any decision: Potential pitfalls for accountability in ai decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 12–21, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593972. URL https://doi.org/10.1145/3593013.3593972.

A XAI Taxonomy

To contextualize the presented publications and the interview findings in this paper, this section provides a short overview over the current state of the art of XAI.

In general, different taxonomies for XAI exist, grouping XAI methods mostly by their general function, their type of result or via their underlying concepts [Speith, 2022]. One of the most common distinction between methods is provided by their "scope", e.g., whether XAI methods are intended to explain a single decision (*local explanation*), the functioning of an entire ML model (*global explanation*), or the dataset of an ML task (*data explanation*).

Some of the earliest—and most common—XAI methods are so-called feature-importance methods, either model-agnostic ones like LIME [Ribeiro et al., 2016] or SHAP [Lundberg and Lee, 2017], or model-specific ones, mainly for computer vision tasks with neural networks, like Integrated Gradients [Sundararajan et al., 2017]. They present their explanations as the "importance" of features towards a decision, e.g., via highlighting input values of tabular data or as saliency maps ("heatmaps") that highlight (super-)pixels of images. Because such explanations can provide ambiguous and thus, difficult to understand, information by only highlighting an area of an image without further information whether the form or the texture of some object is used, concept-based explanations were proposed [Kim et al., 2018]. They try to decompose single decisions (or the general logic of an ML model) into human-understandable concepts, e.g., via concepts such as "striped" or "square-shaped". In combination with such concepts, data-based explanations can be used, which show data instances similar to the input in question or relevant for the concept(s) used [Achtibat et al., 2023]. Another common explanation type are so-called "counterfactual" explanations. They provide information of the type "If feature X would have value y, the outcome would be different", e.g., if a user wants to know why they were not granted a loan. By design, such explanations only provide a limited amount of information to prevent reverse-engineering of the model, but equip users with enough information to be able to challenge a decision or adapt accordingly to receive a different outcome [Wachter et al., 2018].

A further way to explain ML decisions is called "mechanistic interpretability", a bottom-up approach which tries to decompose models into fine-granular explanations by taking their exact computations into account [Bereska and Gavves, 2024]. Corresponding explanations often entail specific neural circuits that are linked to specific behaviors or concepts (comparable to parts of [Achtibat et al., 2023]). Similarly concerned with exact computational behavior of ML models is the field of formal verification, where methods aim at formal or statistical guarantees for specific properties such as robustness against specific input perturbations [Landers and Doryab, 2023]. While theoretically sound (and especially thought to be relevant for ML safety [Landers and Doryab, 2023]), such approaches often struggle with the computational complexity of neural networks for real-world applications.

Often not viewed as part of explainability itself, the closely related field of uncertainty quantification tries to provide ML decisions with uncertainty estimates, enabling users to spot potentially unsafe model decisions [Abdar et al., 2021]. Some authors also call for uncertainty (or "confidence") estimates of explanations themselves to show whether a generated explanation should be trusted [Nauta et al., 2023, Fresz et al., 2024].

Further approaches entail adapting the model structure to inject previous knowledge about the data structure and to assist in explanation generation. As indicated by their name, graph-based neural networks adapt the network structure to allow the interpretation as a graph, potentially improving data approximation and explainability [Agarwal et al., 2023]. Neuro-symbolic approaches combine deep-learning approaches, e.g., for perception tasks [Evans et al., 2021], with classical reasoning, to not rely on post-hoc explanations but to understand local decisions and the global logic of the resulting model [Garcez and Lamb, 2023].

B Coding

For the coding of the interviews, an inductive qualitative analysis based on Mayring [2019] was used. Initially, two coders independently reviewed three interview transcripts to become familiar with the material and identify preliminary categories emerging from the data. Following this, an intermediate coding workshop was conducted in which coders compared their initial codes, discussed discrepancies, and collaboratively refined the coding scheme before proceeding with the coding. During further coding, any potential ambiguities were marked for discussion. In the final coding workshop, these ambiguities were thoroughly examined and consensus was reached on the categorization, ensuring reliability and validity in the coding process. This systematic approach allowed for the inductive development of categories grounded in the interview data. The finalized codes encompassed themes such as the use of AI models—with image processing and large language models (LLMs) being predominant, mentioned in 12 and 10 use cases, respectively, the importance of transparency for internal and external stakeholder groups, and evaluations of XAI methods. Challenges in the use of XAI for certification were identified, including usability concerns (13 mentions), dependence on assumptions (8 mentions), the necessity for use-case-specific requirements (10 mentions), and the identification of 15 relevant—often overlapping or not clearly defined—XAI attributes (e.g., robustness, faithfulness, sensitivity). Additionally, the codes captured the potentials of XAI, its applicability for certification, and aspirations for future developments, highlighting the need for human involvement and further methodological advancements.

C Interview Guide

In the following, the interview guide for the interviews is provided.

C.1 Interviewee Profile

1. What is your role in the company/organization?
2. What is your background, jobwise and course of study?
3. What's the relation between AI and your company/organization? Do you use it, test it, ...?

C.2 (X)AI Use

1. What AI applications are used in your company/organization? Are these self-developed or purchased?
2. To what extent do you consider transparency and explainability requirements in the development and deployment of AI applications?
3. Do you specifically use XAI methods to enhance transparency and explainability?
4. (a) If yes:
    - Which methods are used?
    - What is the goal of using them?
    - Do the current XAI methods assist in achieving this goal?
    - How is the achievement of the goal evaluated? Are there specific attributes that are particularly emphasized?

   (b) If no: Why not? What do you do instead?
5. Are there any specific use cases or examples in your company/organization where the use of XAI has been particularly challenging or successful?

C.3 XAI in certification

This part of the interview was started with a short explanation of the EU AI Act requiring "sufficient" explainability of AI systems, followed by these questions:

1. Do you think explainability/transparency is currently measurable enough to be assessed in a certification process?
2. What are the open questions regarding the measurability of appropriate explainability?

3. Do you think XAI methods should be part of AI certification?
4. What should XAI methods fulfill to be helpful in the certification process?
5. In your opinion, does XAI allow fulfilling of transparency requirements (of the AI Act or other regulations)?

C.4 Outlook

1. Which new trends or technologies in XAI do you see as particularly promising?
2. How do you envision the future of XAI, especially in terms of ethical and regulatory aspects?

D Interview results

In Table 1 the main findings across all 15 interviews are summarized. In Table 2, all conducted interviews are summarized briefly, to provide a better overview of the statements given. Note that the information given is kept general to keep the interviewees non-identifiable.

Table 1: Summary of the main statements by the interviewees about the current state and future development of XAI.

	High Potential	Low Potential
XAI in general	• Communication between domain/AI experts • Clear guidance on when to use which XAI method	• Explaining to lay users • Explaining in situations where the underlying processes are too complex or not well understood
XAI in Certification/safe AI	• Plausibility check of ML model by developers • Discovery of Bias/Errors • Improved Data Understanding	• Assurances about AI safety
Future of XAI	• Increased focus on user needs • New explanation types (concept-based, mechanistic, multi-modal) • Uncertainty quantification of (X)AI	• Comprehensive measurement of transparency/explainability
Future of AI Certification	• XAI as an additional asset of certification processes • Formal verification of safety-relevant AI properties • New AI approaches (e.g., neuro-symbolic)	• XAI as a comprehensive answer to AI certification

Table 2: Summary of the conducted interviews. For the (X)AI-expertise, the following conventions were used: 0 = no expertise, 1 = working expertise with AI, 2 = working expertise with AI and experimenting with XAI, 3 = extended XAI knowledge (without XAI being the focal point of the own work), 4 = active research on XAI. Similar conventions were used for the certification expertise.

Identifier	**Certification**	**(X)AI**	**Noteworthy failed/successful projects with XAI**	**Opinion on XAI State of the Art**	**Opinion on XAI in certification**	**Expected impact on certification**
P1	4	4	Project failed due to explanations being too complex for users. Successful projects with explanations plus contextualisation.	Not yet where it should be.	Helpful asset, since errors can be found. Difficult to evaluate XAI with user studies due to individual differences in users.	Medium
P2	4	3		Not yet where it should be.	Helpful asset.	Medium
P3	4	3	Project showed new clusters in data, which were deemed sensible by domain experts.	"True" Explanations will not be possible (see Section E).	Helpful asset for error detection, improvements of data knowledge.	Low/ Medium
P4	4	3		No hope for the development of global explanations, overall not yet where it should be.	white-box models should be used and AI should not learn during deployment. XAI not really helpful for certification.	Low
P5	4	3		Not yet where it should be.	Helpful asset. Assumptions in XAI methods should be documented (see Section E).	Medium
P6	4	2		Not yet where it should be.	XAI only truly relevant when guarantees for (X)AI can be given.	Low
P7	3	4	Project, where XAI showed errors in ML application for image data.	Not yet where it should be.	Helpful asset.	Medium
P8	3	4		Not yet where it should be.	Helpful Asset.	Medium
P9	3	4		Not yet where it should be.	Helpful Asset.	Medium
P10	3	3	Failed to produce useful explanations for time series, successful for image data.	Not yet where it should be.	Hopes for formal verification/robustness analysis and performance metrics for AI certification.	Low

Continued on next page

Table 2 continued from previous page

Identifier	**Certification**	**(X)AI**	**Noteworthy failed/successful projects with XAI**	**Opinion on XAI State of the Art**	**Opinion on XAI in certification**	**Expected impact on certification**
P11	3	3	SHAP is used for internal communication (given enough experience). XAI fails due to too complex models/pipelines. Counterfactuals fail due to too many immutable/sensitive attributes.	Not yet where it should be.	Helpful asset.	Medium
P12	3	3	Tested XAI methods could not detect frequency domain features for time series.	Not yet where it should be.	XAI would need clear guidelines, then it would be a helpful asset.	Low/ Medium
P13	3	2		Not yet where it should be.	Helpful asset.	Low/ Medium
P14	2	4	XAI showed bias in text application.	Good due to the theoretical guarantees of methods (especially IntGrad).	Human interpretation and access to model and data important to make XAI helpful asset.	Medium/ High
P15	2	3	Explanations failed due to being too complex and fundamental connections in data not known.	Not yet where it should be.	Helpful asset (especially for fairness).	Medium

E General Remarks about AI Certification and XAI

Additional to the more universal statements on XAI and XAI for certification, some interviewees also voiced concerns and opinions on specific topics. Since these remarks are not commonly found throughout literature and believed by the authors to add interesting viewpoints to the discussion aimed at by this paper, they are presented in this section. To clarify the distinction between the statement made and additional information provided to contextualize the statements during the writing of this paper, the initial statement is given in italic. Note that these statements are not direct quotes. Most of them were translated from German and edited for brevity and readability, as the direct quotes were spoken language and embedded in the context of the corresponding interview.

E.1 Incorrect Evidence?

P2 + P6: *Certification so far examines whether evidence is in line with the requirements of standards and norms. There is no process in place to check whether this evidence is correct.*

Expert discussions, particularly with P2 and P6, highlighted that AI certification introduces new challenges compared to traditional product certification, even outside the technical challenges described before. Traditionally, certification verifies whether evidence provided by manufacturers conforms to standards and norms, assuming this evidence is accurate and reflects the actual processes or product. However, with evidence generated by XAI, this assumption may not hold. Malicious actors might generate arbitrary explanations for their AI systems using established methods [Slack et al., 2020, Zhou and Joachims, 2023], and incorrect evidence might be provided unintentionally without malicious intent. Therefore, responsibilities must be clarified: Do manufacturers guarantee the correctness of the evidence—which is hardly feasible with the current state of XAI—or must certifiers consider the generation process of the evidence, requiring in-depth knowledge of AI and XAI?

P5: *Until now, the "Uniformity Hypothesis" and the "Competent Programmer Hypothesis" were helpful pieces of building and certifying safety-critial systems.*

Similar to the point above, some previous assumptions might not hold for AI certification. Usually, the here mentioned "Uniformity Hypothesis" [Sterling, 1969] has been applied, as it describes that specific data points can be selected for tests, whose findings generalize across an equivalence class of similar data. Additionally, the "Competent Programmer Hypothesis" [DeMillo et al., 1978] postulates that safety-relevant software does not produce completely unpredictable errors because it was created by a competent programmer who can avoid errors that appear random (e.g., by buffer overflows or pointers, as commented by P5). Note that the "Competent Programmer Hypothesis" does not warrant blind trust to competent programmers but is supported by programming guidelines, static code analysis and further measures to help avoid seemingly random errors during execution. However, both assumptions are violated by the black-box nature of AI: for the generalizability of tests to particular data, equivalence classes are difficult to find and the decision-making process of an ML system has not been explicitly programmed, while the range of techniques to check ML systems for errors (such as static code analysis for classical code) is quite small, although P5 noted that XAI can be used here. Due to the complexity of ML models, resulting errors may appear random. Nevertheless, P5 noted that a good step towards safer AI is to document the assumptions made in AI and XAI, which is often not done for assumptions such as the "Uniformity Hypothesis" and the "Competent Programmer Hypothesis" in classical software development. Regarding the criticism faced by some assumptions in XAI, P7 explicitly pointed out that scientific progress typically comes from challenging existing ideas. In the field of XAI, this leads to a complexity that is difficult for practitioners to penetrate. Initially, XAI was proposed for the evaluation or testing of AI, but now there are also metrics for the testing of XAI, and even metrics for evaluating those metrics [Tomsett et al., 2020]. Consequently, a goal of applied research could now be to provide explicit recommendations on how to select XAI methods for specific use cases.

E.2 Fundamental Changes in Certification

P1: *If AI is to be certified, there needs to be a discussion about a shift from value-based to utilitarism-based certification.*

Due to the uncertainties in testing AI systems, an expert speculated that the culture of certification has to fundamentally change to accommodate AI: Existing certification is principally guided by societal values and norms. For example, for the norm “safety” of a technical system, a threshold can be defined, which can then be adhered to based on a detailed analysis of an overall system, for example through methods such as Failure Mode and Effect Analysis [Stamatis, 2003]. If this value is not maintained, countermeasures must accordingly be taken from the development side. A particularly well-known example of the traceability of ethical values in technical systems can be found in the field of autonomous driving, the so-called “trolley problem”. In this scenario, an immediate choice must be made before an accident as to which involved individuals are subjected to a higher risk of severe injuries or potentially death. For AI, however, such thresholds and ethical decisions are currently not sufficiently determinable. Therefore, the expert suspected that the use of AI might need to be assessed more from utilitarian viewpoints, meaning “If the use of AI is expected to result in fewer injuries or fatalities in traffic, then its use is sensible.”

E.3 Responsibility for Explainability Requirements

P1 + P2: *Explainability is more of a societal than technical topic, as such the standardization bodies are not well equipped to deal with it.*

To be able to certify a technical system, standards are used. Tasked with creating such standards are organizations such as DIN (for Germany), CEN/CENELEC (European Committee for Electrotechnical Standardization), or ISO (International Organization for Standardization). As these organizations commonly create technical standards, P1 and P2 argued that issues such as explainability and fundamental considerations like the compliance with and negotiation of ethical values (see above) should not be technical discussions, but rather socio-political debates. It is also particularly noteworthy that while technical evaluation methods for XAI do exist, they were not considered to be effective by the majority of study participants, which suggests that purely technical standardization is unlikely to resolve the open questions surrounding the assessment and certification of AI. P2 additionally noted that participants of standardization committees might lack the time to be well informed about topics as current as XAI (due to other obligations), thus resulting in standards that might not represent the current state of science.

E.4 Fundamental Doubts about XAI

P3 (with a background in neuroscience): *For AI, the requirements are stricter than ever possible for humans. At best, XAI might provide justifications, while the only possible explanation for an AI system is its complete calculation from input to output.*

Around the topic of AI certification, there exists a discussion of whether AI should be subject to stricter requirements than humans doing the same task (as also touched on by Fresz et al. [2024]). P3 extended this by linking explanations to the description of thought processes provided by Kahneman [2012], dividing thought processes in system 1 thinking (fast, low effort, ‘intuitive’) and system 2 thinking (slow, high effort, deliberate). P3 argued that humans may justify their behavior upon request after the fact (system 2), but such justifications are not identical to the actual motives, especially for decisions that are often made intuitively (system 1). They suggested that the same applies to XAI: XAI could produce a justification for ML behavior that is understandable to humans (system 2), but the true explanation could only be found within the computational chain of the ML system and, although fundamentally ‘transparent’ (i.e., visible), not entirely understandable to humans due to the potentially huge number of calculations made by the ML system. It could be argued here that the ideal conception of XAI enables the computation chain to be summarized in such a way that a correct explanation is produced (e.g., via concepts), which provides users with insights into the ML behavior.

E.5 New Paradigms for XAI

P8: *For the field of XAI, I am particularly optimistic about the feedback of XAI information into ML training.*

P8 identified the combination of the explanation process with the associated model improvement as particularly promising in the field of XAI. So far, XAI has mostly been viewed unidirectionally—even if errors and biases can be identified in existing models, there is no simple way yet to intervene in the model or training data to correct existing problems. A new paradigm (similar to the one proposed by Pahde et al. [2023]) could offer the possibility to interact with explanations, correct them, and integrate these corrections back into the model training process. Thus, insights gained from XAI could be efficiently used for the error correction of ML models.

E.6 Differences between Research and Practice

P7 + P8 + P9: *In research, cognitive load and interaction time with explanations are often not explicitly considered.*

Multiple participants criticized that in XAI research, the explicit experience and aims of domain experts are not considered enough. They noted that XAI research seems to operate under the assumption that complete explanations should be generated in all circumstances. In contrast, domain experts, such as physicians, typically only require explanations in specific instances.

Furthermore, users are more likely to interact with and have a positive experience with explanations that serve to reduce the cognitive load associated with the task at hand. The majority of users, in their daily routines, lack the time and cognitive resources to engage with overly complex explanations. This effectively undermines the core objective of XAI, which is to make AI more accessible. To make explanations easier to understand, P1 mentioned that explanations need to be contextualized to fulfill their potential, which could potentially increase or decrease the cognitive load based on the specific implementation. While interaction time and cognitive load are not commonly evaluated in XAI literature, there is some existing research that explores the idea of reducing the cognitive load of explanations [Herm, 2023]. In another approach, users cannot see the result of an ML prediction without first interacting with an explanation and making their own prediction [Miller, 2023]. Thus, user interaction with explanations is enforced, thereby potentially improving task performance by increasing the cognitive load of the task at hand.