Can Requirements Engineering Support Explainable Artificial Intelligence? Towards a User-Centric Approach for Explainability Requirements

Umm-E- Habiba
Institute of Software Engineering
University of Stuttgart
Stuttgart, Germany
Email: umm-e-habiba@iste.uni-stuttgart.de

Justus Bogner
Institute of Software Engineering
University of Stuttgart
Stuttgart, Germany
justus.bogner@iste.uni-stuttgart.de

Stefan Wagner
Institute of Software Engineering
University of Stuttgart
Stuttgart, Germany
stefan.wagner@iste.uni-stuttgart.de

***Abstract*—With the recent proliferation of artificial intelligence systems, there has been a surge in the demand for explainability of these systems. Explanations help to reduce system opacity, support transparency, and increase stakeholder trust. In this position paper, we discuss synergies between requirements engineering (RE) and Explainable AI (XAI). We highlight challenges in the field of XAI, and propose a framework and research directions on how RE practices can help to mitigate these challenges.**

Keywords: *Requirements Engineering, Explainability, XAI*

I. Introduction

In the last decade, Artificial Intelligence (AI) systems, especially based on Machine Learning (ML), have achieved remarkable feats in many areas that were previously computationally impossible [1]. AI is now prevalent in our professional and personal lives. Progress has also been made in the fields of medicine, automated vehicles, bioinformatics, and recommender systems. Despite these advancements, trust issues and disillusionment are emerging in several areas. One of the most common problems is the lack of transparency in these systems, as the black-box nature of many ML models collides with the human desire to understand the system and its output [2]. This is especially concerning when these systems are used in contexts where they have a major impact on human lives, such as medical diagnostic systems or financial and legal matters.

Software systems' growing autonomy and complexity make it harder for software engineers and domain experts to comprehend them [3], especially if they are composed of ML components [4]. This leads to a need for explainable systems. Explanations help to understand system decisions, which increases confidence in a system and improves trustworthiness. Additionally, it also justifies actions taken, increases usability, aids in uncovering causes of mistakes, and reduces the potential of human error because humans cannot make informed decisions without access to the system rationale for its recommendation. An appropriate explanation for unexpected or inaccurate system behavior might also help to determine problems, for instance, the misinterpretation of a requirement or a mistake in the system design.

Moreover, a lack of explainability not only causes ethical, social, and legal issues [5]. It also exacerbates distrust, reduces user acceptability [6], and stifles the adoption of innovative technology. In this regard, the Ethics Guidelines for Trustworthy AI [7] have been proposed by the High-Level Expert Group on Artificial Intelligence (AI HLEG). They highlighted transparency as a key requirement for trustworthy AI, which is further comprised of traceability, explainability, and communication. Depending on the context of relevant stakeholders, these guidelines emphasize a proper explanation of the decision-making process of AI systems. Suitable development techniques are required to provide a certain level of explainability. The development process should incorporate explainability to make it explicit. However, the concrete scope for *explainability* is still unclear, and many researchers are actively exploring different courses of action to support explainability.

Approaches such as LIME [8] and Shapley values [9] are based on mathematical methods, while the Requirements Engineering (RE) community is also exploring different approaches [10], [11] to support explainability, although it has already been identified as a key non-functional requirement to support transparency and reduce opacity. Current research is either inclined towards model explainability, which particularly focuses on ML engineers, or user experience, which addresses user needs for explainability and how users react to certain explainability techniques. A problem arises when it comes to the interaction between ML engineers and multiple stakeholders, as they have different expectations and views of explainability. However, there is a lack of studies that investigate the gap between the understanding of ML engineers and end-users regarding explainability. We believe RE practices can help to address this challenge.

Position and Contribution: In this paper, we provide an overview of the state-of-the-art in Explainable AI (XAI). Our objective is to highlight current challenges in XAI from the perspective of RE and provide future directions on what could be options to mitigate these challenges.

II. Explainability vs. Interpretability

Terminology related to explainability in ML is not consistently used [12], with terms such as "explainable AI" [13], "interpretability" [14], and "comprehensibility" [15]. They are all aimed at making the system more transparent, i.e., explaining the system's inner working and making it more understandable.

In particular, explainability and interpretability appear frequently in the literature, and many researchers use them interchangeably [16] as no specific mathematical definitions for explainability or interpretability exist, nor have they been quantified by any metric [17]. While they refer to overlapping concepts, there is some work that attempts not only to define these two concepts but also related notions such as comprehensibility [4], [14], [18]. Miller [19] defines interpretability as "the degree to which a human can understand the cause of a decision". Another definition comes from Doshi-Velez and Kim [14], who define interpretability as "the ability to explain or communicate in intelligible words to a human". These definitions refer to interpretability as a property of how input and output are associated and how easily end-users can identify cause-and-effect relations.

In contrast, "explainability" deals with the internal logic of the system, i.e., how the ML system has been trained and how a particular output has been generated [20]. Doshi-Velez and Kim [14] define explanations as "the currency in which we exchange beliefs". Rudin [21] differentiates between interpretability and explainability via the concepts "inherently interpretable" and "post-hoc explanations". In [14] and [18], the authors describe interpretability as a broader concept and argue that explainability is essential to support interpretability.

Table I summarizes how different authors strived to define explainability and interpretability. However, we can see from this literature, there are no agreed definitions of explainability or interpretability. The AI research community is still working towards a unified position. To motivate research in this area, researchers are trying to understand and elicit different requirements, e.g., why these explanations and interpretations are needed, what their context is, and who the stakeholders in these explanations are.

To address end-user needs, research must consider their perspective of explanations. While the human-computer interaction (HCI) community is actively exploring user-centric XAI [23]–[26], a more extensive contribution is still required from the RE community to provide a systematic approach to enable software engineers to effectively incorporate explainability requirements in the development process. In the following section, we propose an alliance between RE and XAI.

III. Synergies between RE and XAI

The interactions between the two fields RE and XAI are still evolving. A few studies discuss various interactions of RE and XAI. However, the major focus is concrete requirements for XAI. Hall et al. [9] proposed a 5-step methodology to understand explainable AI requirements. Their goal was to understand explanation requirements from different perspectives. Kohl et al. [12] proposed a process for eliciting, specifying, and verifying explainability as a non-functional requirement. Chazette et al. [27] also concentrated on a user perspective of explanations. Their work is more focused on software transparency and user opinion about embedded explanations in software. Another work [28] from the same authors explores the interaction between explainability and other non-functional requirements. They also describe potential trade-offs of incorporating explainability into the system. Similarly, Liao et al. [29] interviewed 20 user experience and design practitioners to identify gaps in current practices. They provided insights into the design space of XAI, and a question bank that can be used to create user-centered XAI. Eiband et al. [30] improved existing user interface guidelines by proposing a stage-based participatory process. Their approach is aimed more towards design management and user interface design.

Most of the presented work is focused on user-centered design and interactions of explainability with other non-functional requirement. There are studies centered on different stakeholder perspectives, but several challenges remain. We will outline them in the following section.

IV. Challenges

Current research in XAI is inclined towards model explainability and that these explanations are more focused on ML engineers rather than end-users of the system. Most existing explanation techniques and tools are only comprehensible by technical stakeholders with an ML background, like data scientists and ML engineers. Approaches that pay attention to the end-user are more rare. Although there are studies that attempted to work in the direction of RE and try to ask stakeholders about XAI requirements, many challenges remain.

a) Absence of a mediator role: Recently, several open-source toolkits [31] emerged for XAI. They provide explanations which are mostly comprehensible by ML engineers. Yet, these approaches are difficult to adapt for end-users, e.g., due to the gap between user needs and available technical capabilities to improve the effectiveness of explainability in AI products [29]. Thus, the role of mediators arises, whose job it is to fill the gap between stakeholders. They have the experience to identify user needs and to convey them to the developers, which supports a better translation of end-user needs into technical aspects. In this sense, they serve as a bridge between users and ML engineers, keeping in mind the demands and constraints on both sides. Thus, explainable solutions for AI systems also need to provide explanations for the mediators, whose job is to bridge the gaps among stakeholders.

b) No coherent definition of explainability: Usually, demands for explainability and how it can be accomplished are ambiguous [12]. Several authors presented different facets of explainability, with no agreed on definition, which can, e.g., impact communication in AI projects.

c) Lack of stakeholder-centric approaches: Current solutions for model interpretability do not describe or specify

TABLE I
INTERPRETABILITY VS EXPLAINABILITY IN AI LITERATURE

Papers	Interpretability	Explainability
[4]	Interpretability addresses "how does the model work?"	Explainability addresses "what else can the model tell me?"
[14]	Interpretability is "the ability to explain or to present something in understandable terms to a human"	Explanations are "the currency in which we exchange beliefs"
[18]	Mostly concerned with the intuitions behind the output of a model "to describe the internals of a system in a way that is understandable to humans"	Mostly concerned with the internal logic, i.e., "models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions"
[19]	"The degree to which an observer can understand the cause of a decision"	"Explicitly explaining decisions to people"
[22]	"An interpretation is the mapping of an abstract concept into a domain that humans can make sense of"	"An explanation is the collection of features of the interpretable domain that have contributed to a given example to produce a decision"

their intended users. Therefore, most of these solutions unwittingly are more understandable to the people who build them, i.e., ML engineers. Alternately, the interpretable model built for end-user has issues such as simple visuals with general explanations often not useful for people in practice [32].

d) No common vocabulary for all ML stakeholders: Many terms have recently been introduced in explainable AI. However, these terms appear interchangeably, which leads to confusion in this rapidly expanding field. Furthermore, certification authorities, researchers, end-user, domain experts, and ML engineers all use explainable AI with different expectations and understanding. This leads to more problems, as each stakeholder desires separate objectives to be satisfied by the term explainable AI [33].

To address these challenges, we propose a general framework to provide guidance on how these issues can be minimized and how explainability can be better supported by RE practices. For each step, we are working on more detailed methods and guidelines to support practitioners.

V. PROPOSED FRAMEWORK

In this section, we i) draw a boundary between interpretability and explainability and ii) propose a framework to mitigate the challenges stated above.

As we have seen, these two terms have been used interchangeably, even though some authors draw a boundary between them. As a synthesis of these definitions, we differentiate between the two terms to make it explicit that our framework supports explainability:

1) *Interpretability*: looking *at* the model, i.e., the stakeholder can understand the cause-and-effect relationship between input and output.
2) *Explainability*: looking *inside* the model, i.e., the stakeholder can understand the inner workings of the model and how the model has produced a particular output.

Activities in our framework will be performed by the mediating role of the requirements engineer. This framework explicitly aims to identify stakeholders with the requirements of explainability. Furthermore, it will help to establish a shared understating of explainability among multiple stakeholders. The framework comprises the following steps (Figure 1):

1) **Identifying stakeholders:** In this step we aim to address the question "explainable for whom?" We will identify relevant stakeholders for the system, i.e., end-users, ML engineers, and domain experts. We will also determine the stakeholders for explanations.
2) **Identifying requirements:** We will elicit and document the concrete requirements for explainability, together with rationales like why explainability is required and what the impact of each explanation will be on the system. We will also describe what types of explanations will be provided. This will answer the question "why is explainability needed?"
3) **Establishing a common vocabulary:** Using knowledge of end-users, domain experts, and ML engineers, we will establish a common vocabulary. We will introduce common terms and create a shared system understanding, which supports discussing explainability requirements.
4) **Negotiating and validating requirements:** This is an iterative process. The identified explainability requirements will be validated in this step. This will answer questions like "what if explainability is not possible?" or "how much uncertainty is tolerable?" If there is any trade-off between the requirements or some are not feasible, they will be negotiated.
5) **Classifying requirements:** In this step, requirements will be classified with respect to stakeholders. Who are the stakeholders for particular explanations, and what do they want to achieve with these explanations? The main question to answer is "how much explainability is enough for which stakeholder?" This classification will also help to address each stakeholder requirements.

VI. CONCLUSION AND FUTURE WORK

In this position paper, we discussed differences between interpretability and explainability, and pointed out existing challenges in the XAI space. Taking a RE perspective, we propose a work-in-progress framework to address the challenges, with emphasis on a user-centered approach to explainability requirements and highlight the potential of RE to support the explainability of AI systems.

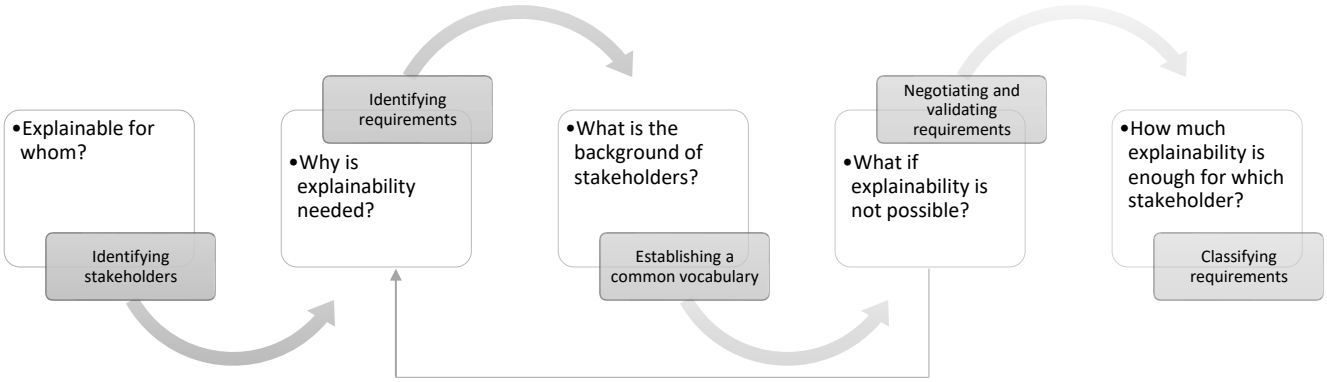

Fig. 1. Proposed Framework

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, 2015.

[2] L. Yang, H. Wang, and L. A. Deleris, "What does it mean to explain? a user-centered study on ai explainability," in *International Conference on Human-Computer Interaction*. Springer, 2021.

[3] DARPA, "Broad agency announcement, explainable artificial intelligence (xai)," *DARPA-BAA-16-53*, pp. 7–8, 2016.

[4] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[5] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi *et al.*, "AI4People–an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, 2018.

[6] M. Lahijanian and M. Kwiatkowska, "Social trust: a major challenge for the future of autonomous systems," in *2016 AAAI Fall Symposium Series*, 2016.

[7] "Ethics guidelines for trustworthy AI," https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html\#Transparency, accessed: 2022-03-30.

[8] M. T. Ribeiro, S. Singh, and C. 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*, 2016, pp. 1135–1144.

[9] M. Hall, D. Harborne, R. Tomsett, V. Galetic, S. Quintana-Amate, A. Nottle, and A. Preece, "A systematic method to understand requirements for explainable AI (XAI) systems," in *Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019)*, vol. 11, 2019.

[10] C. T. Wolf, "Explainability scenarios: towards scenario-based xai design," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 252–257.

[11] D. Cirqueira, D. Nedbal, M. Helfert, and M. Bezbradica, "Scenario-based requirements elicitation for user-centric explainable ai," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2020, pp. 321–341.

[12] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender, "Explainability as a non-functional requirement," in *2019 IEEE 27th International Requirements Engineering Conference (RE)*. IEEE, 2019.

[13] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[14] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[15] A. A. Freitas, "Comprehensible classification models: a position paper," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 1, pp. 1–10, 2014.

[16] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, 2019.

[17] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE access*, vol. 6, 2018.

[18] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[19] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[20] M. Turek, "Explainable artificial intelligence (XAI)," *Defense Advanced Research Projects Agency*, 2018.

[21] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[22] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, 2018.

[23] J. J. Ferreira and M. Monteiro, "Designer-user communication for XAI: An epistemological approach to discuss XAI design," *arXiv preprint arXiv:2105.07804*, 2021.

[24] M. Naiseh, D. Al-Thani, N. Jiang, and R. Ali, "Explainable recommendation: when design meets trust calibration," *World Wide Web*, vol. 24, no. 5, pp. 1857–1884, 2021.

[25] S. Najafian, A. Delic, M. Tkalcic, and N. Tintarev, "Factors influencing privacy concern for explanations of group recommendation," in *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 2021, pp. 14–23.

[26] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable AI," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–15.

[27] L. Chazette, O. Karras, and K. Schneider, "Do end-users want explanations? analyzing the role of explainability as an emerging aspect of non-functional requirements," in *2019 IEEE 27th International Requirements Engineering Conference (RE)*. IEEE, 2019.

[28] L. Chazette and K. Schneider, "Explainability as a non-functional requirement: challenges and recommendations," *Requirements Engineering*, vol. 25, no. 4, 2020.

[29] Q. V. Liao, D. Gruen, and S. 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*, 2020.

[30] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann, "Bringing transparency design into practice," in *23rd International Conference on Intelligent User Interfaces*, 2018.

[31] R. Marcinkevičs and J. E. Vogt, "Interpretability and explainability: A machine learning zoo mini-tour," *arXiv:2012.01805*, 2020.

[32] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan, "Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[33] 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*, 2020.