Peeking Outside the Black-Box: AI Explainability Requirements beyond Interpretability

Jakob Droste[1,*,†], Hannah Deters[1,†], Ronja Fuchs[1] and Kurt Schneider[1]

[1]*Leibniz University Hannover, Software Engineering Group, Hannover, Germany*

Abstract

With the rise of artificial intelligence (AI) in society, more people are coming into contact with complex and opaque software systems in their daily lives. These black-box systems are typically hard to understand and therefore not trustworthy for end-users. Research in eXplainable Artificial Intelligence (XAI) has shown that explanations have the potential to address this opacity, by making systems more transparent and understandable. However, the line between interpretability and explainability is blurry at best. While there are many definitions of explainability in XAI, most do not look beyond the justification of outputs, i.e., to provide interpretability. Meanwhile, contemporary research outside of XAI has adapted wider definitions of explainability, and examined system aspects other than algorithms and their outputs. In this position paper, we argue that requirements engineers for AI systems need to consider explainability requirements beyond interpretability. To this end, we present a hypothetical scenario in the medical sector, which demonstrates a variety of different explainability requirements that are typically not considered by XAI researchers. This contribution aims to start a discussion in the XAI community and motivate AI engineers to take a look outside the black-box when eliciting explainability requirements.

Keywords

Explainable Artificial Intelligence, Interpretability, Requirements Engineering

1. Introduction

In modern society, AI permeates an increasing range of professional and personal spaces. Intelligent software systems are used by companies and private end-users alike. These systems are often opaque in nature and are referred to as black-boxes [1], as their inner workings are not visible and understandable to the average observer. As they are commonly used in safety-critical environments, such as autonomous driving [2] or the medical sector [3, 4], it is necessary for these systems to be transparent and trustworthy for their users. The research area of XAI aims to address this issue, by providing explanations that help end-users understand the AI systems they are working with, and by providing transparent reasoning for the AI systems'

In: D. Mendez, A. Moreira, J. Horkoff, T. Weyer, M. Daneva, M. Unterkalmsteiner, S. Bühne, J. Hehn, B. Penzenstadler, N. Condori-Fernández, O. Dieste, R. Guizzardi, K. M. Habibullah, A. Perini, A. Susi, S. Abualhaija, C. Arora, D. Dell'Anna, A. Ferrari, S. Ghanavati, F. Dalpiaz, J. Steghöfer, A. Rachmann, J. Gulden, A. Müller, M. Beck, D. Birkmeier, A. Herrmann, P. Mennig, K. Schneider. Joint Proceedings of REFSQ-2024 Workshops, Doctoral Symposium, Posters & Tools Track, and Education and Training Track. Co-located with REFSQ 2024. Winterthur, Switzerland, April 8, 2024.

[*] Corresponding author.

[†] These authors contributed equally.

jakob.droste@inf.uni-hannover.de (J. Droste); hannah.deters@inf.uni-hannover.de (H. Deters); ronja.fuchs@stud.uni-hannover.de (R. Fuchs); kurt.schneider@inf.uni-hannover.de (K. Schneider)

0000-0001-8746-6329 (J. Droste); 0000-0001-9077-7486 (H. Deters); 0000-0002-7456-8323 (K. Schneider)

behavior and decisions. This is also referred to as providing interpretability [5]. In this paper, we argue that explainability as a non-functional requirement (NFR) goes beyond just providing interpretability. To this end, we identify three types of needs for explanations that are typically not covered by XAI research and engineering: privacy information, system interaction and domain-related information. To demonstrate the applicability of these explainability needs to AI systems, we discuss a use case of image recognition and classification algorithms in the medical sector. The goal of this contribution is to raise awareness for AI explainability requirements outside of providing interpretability, and to motivate discourse among XAI researchers and engineers about explainability as an NFR.

2. Interpretability in Explainable Artificial Intelligence

Gilpin et al. [5] refer to explainability as a means to achieve interpretability and completeness, with interpretability referring to system internals and completeness referring to describing the internal operations of a system in an accurate manner. When the term explainability is used in XAI papers, authors typically refer to explanations of the model or to the reasoning for its outputs [6, 7, 8]. In other words, explainability is used as a means to provide interpretability.

Interpretability is seen as a means to improve trust and transparency by offering human-understandable justifications for the decisions made by an AI system by offering explanations [9]. Moreover, explaining model behavior is thought to improve model bias understanding and fairness [9]. Fan et al. [6] conducted a comprehensive survey on interpretability of neural networks and compiled a definition of interpretability:

> **Definition:** Interpretability refers to the extent of human's ability to understand and reason a model. [6]

Notably, this definition of interpretability does not go beyond explaining a system's model and reasoning. XAI research usually divides all AI explanation methods into local and global methods [1, 7, 9]. Local methods refer to individual data instances and global methods to the model as a whole. Once again, explaining the model itself is the focus of XAI.

3. Explainability Requirements Beyond Interpretability

In recent years, explainability has also been researched independently of AI. Mucha et al. [10] investigated how interfaces for explanations should be designed. Other studies looked into the personalization of explanations [11, 12]. Chazette et al. [13] investigated the influence of explainability on related NFRs, such as understandability and usability. Through on a comprehensive systematic literature review (SLR), they examined different definitions of explainability and derived a general definition of explainable systems:

> **Definition:** A system S is explainable with respect to an aspect X of S relative to an addressee A in context C if and only if there is an entity E (the explainer) who, by giving a corpus of information I (the explanation of X), enables A to understand X of S in C. [13]

Following this definition, explanation can be used to make different kinds of system aspects more understandable. While Chazette et al. [13] do not provide a comprehensive overview of all explainable system aspects, they name some examples that they identified via their SLR. System aspects such as the system's inner logic and its reasoning process are aspects that correspond to conventional XAI explanations, i.e., they support interpretability.

However, Chazette et al. [13] also consider a system's knowledge about its user to be an explainable system aspect. Explanations that allow users to understand how their personal information is stored and processed differ from typical XAI explanations. In this context, Brunotte et al. [14] introduce the concept of **privacy explanations** as part of explainability. The need for privacy explanations is also underlined by the works of Hamon at al. [15] and Jobin et al. [16], that highlight the importance of data protection and privacy in AI systems.

Chazette et al. [17] also state that explanations can guide users, and that they can make systems more easily operable. This reveals another type of explanatory need, namely the need to explain end-users' interactions with the system. Deters et al. [18] investigated the need for explanations in a modified social media app and identified that the most common need for explanations was user guidance. This includes explaining both the navigation and the operation of a system. **Interaction explanations** are required when users have a goal in mind and know what they want to do, but do not know how to achieve this goal within the system.

Lastly, we consider explanations of domain-specific elements to be part of a system's explainability. For example, explanations of specific terminology, that lay users do not know, could potentially increase the usability of a given system [19, 20]. Chazette et al. [17] stated that domain aspects have an influence on how explanations should be designed. We argue that **domain-related explanations** of specific terms and system elements can support end-users and can enable the effective use of the system.

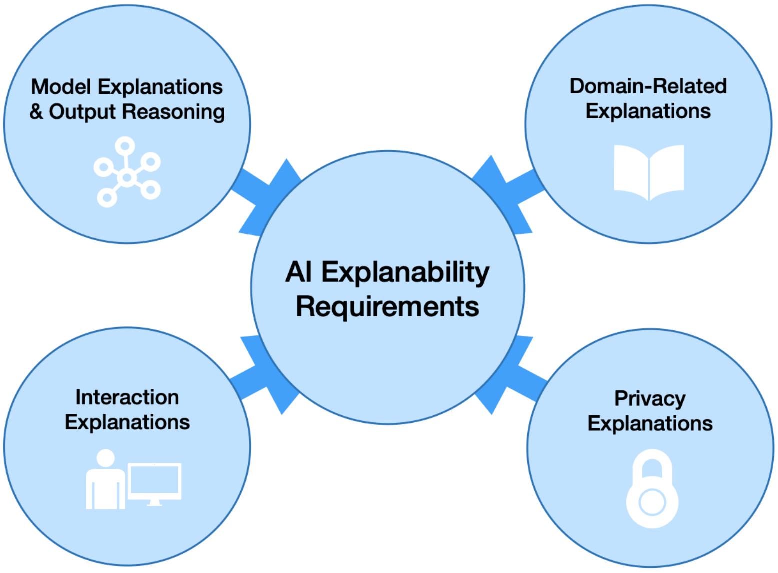
Model Explanations & Output Reasoning
Domain-Related Explanations
AI Explanability Requirements
Interaction Explanations
Privacy Explanations

Figure 1: Four types of AI explainability requirements.

Figure 1 shows the four types of AI explainability requirements identified in this paper. Note that we do not claim this to be a complete taxonomy and that there might still be more types of explainability requirements that warrant further investigation. Research into different types of explainability requirements is already ongoing for AI-independent software systems [14, 18]. However, there is a lack of research into explainability requirements specific to AI systems that do not relate to interpretability. If requirements engineers of AI systems continue to reduce explainability requirements to the explanation of algorithms and system outputs, important end-user needs might be omitted.

4. Exemplary Scenario

To demonstrate the applicability of different explainability requirements to AI systems, we present a hypothetical scenario from the medical sector. In clinical diagnostics, image recognition algorithms are used to support both medical professionals and their patients. In dermatology in particular, image recognition and classification is used for the identification of malignant tissue in suspicious skin areas [4]. While there are several particular technologies to enable this, modern systems in this domain usually rely on machine learning and related AI algorithms [4].

In this scenario, we consider two important stakeholder groups that interact with image recognition systems: medical professionals and lay users. Medical professional like dermatologists are trained to examine suspicious skin areas and identify malignant tissue such as melanomas. To support their decision making and avoid fatal errors in diagnosis, they can use image recognition systems. Lay users, such as patients, may use the same image recognition technology to get a preliminary diagnosis before they decide whether they want to visit a doctor.

Different kinds of explanations do not have equal significance across these stakeholder groups. Furthermore, providing inappropriate explanations can potentially cause frustration and confusion in end-users [17, 21]. To provide appropriate explanations for every stakeholder group, the diverse explainability needs of each group must be considered [12]. For example, medical professionals need explanations on why the system made a certain diagnosis instead of another. If professionals are not able to understand and verify the diagnoses made by the system, they might ultimately not trust the diagnoses, even if they are correct. Conversely, as trained professionals, they need less explanations concerning the operation of the system. Furthermore, they do not need detailed explanations related to the medical terminology used by the system, and they do not need in-depth privacy explanations, as long as they know that their patients' medical data is not misused.

Lay users might prefer explanations on how to correctly input the data or on what certain medical terms mean. If they are unable to correctly navigate the system or provide the necessary inputs, they might be unable to use the system or they might accidentally make incorrect inputs that can lead to incorrect diagnoses by the system (e.g. not inputting an appropriate photo of the suspicious skin area). In contrast to medical professionals, lay users do not understand most of the medical terminology used by the system. In order to be able to understand a diagnosis, they need explanations for the medical terminology that is used. As the system uses their personal medical data, lay users might also want privacy explanations that detail exactly how their data is stored and processed.

5. Conclusion and Research Perspective

Within the field of XAI, explainability is typically reduced to explaining the inner workings of black-box AI systems, or to justifying their outputs, i.e., to provide interpretability. By providing insights on how an AI system works and how it reaches its results, XAI engineers aim to increase the transparency and understandability of the system. In contrast to this, recent research in explainability requirements outside of AI has shown that there are explainability requirements that go beyond interpretability. Examples of this are privacy explanations, aiming to foster trust, as well as interaction and domain explanations, aiming to increase the usability of the system.

In this position paper, we discussed the applicability of these different types of explanations to AI systems. In particular, we provided the example of image recognition in the medical sector to show how different kinds of explainability requirements can apply to an existing AI use case. We hope that this demonstration provides the necessary push for XAI researchers and engineers to take a look outside the black-box, and to start considering explainability requirements beyond interpretability when designing explainable systems.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No.: 470146331, project softXplain (2022-2025).

References

[1] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[2] S. Atakishiyev, M. Salameh, H. Yao, R. Goebel, Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions, arXiv preprint arXiv:2112.11561 (2021).

[3] F. Klauschen, J. Dippel, P. Keyl, P. Jurmeister, M. Bockmayr, A. Mock, O. Buchstab, M. Alber, L. Ruff, G. Montavon, et al., Erklärbare künstliche intelligenz in der pathologie, Die Pathologie (2024) 1–7.

[4] Z. Li, K. C. Koban, T. L. Schenck, R. E. Giunta, Q. Li, Y. Sun, Artificial intelligence in dermatology image analysis: current developments and future trends, Journal of Clinical Medicine 11 (2022) 6826.

[5] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, 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.

[6] F.-L. Fan, J. Xiong, M. Li, G. Wang, On interpretability of artificial neural networks: A survey, IEEE Transactions on Radiation and Plasma Medical Sciences 5 (2021) 741–760.

[7] Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, IEEE Transactions on Emerging Topics in Computational Intelligence 5 (2021) 726–742.

[8] G. Guizzardi, N. Guarino, Semantics, ontology and explanation, arXiv preprint arXiv:2304.11124 (2023).

[9] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (XAI): A survey, CoRR abs/2006.11371 (2020). URL: https://arxiv.org/abs/2006.11371. arXiv:2006.11371.

[10] H. Mucha, S. Robert, R. Breitschwerdt, M. Fellmann, Interfaces for explanations in human-ai interaction: Proposing a design evaluation approach, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, CHI EA '21, Association for Computing Machinery, New York, NY, USA, 2021. doi:10.1145/3411763.3451759.

[11] H. Ramos, M. Fonseca, L. Ponciano, Modeling and evaluating personas with software explainability requirements, in: Human-Computer Interaction: 7th Iberoamerican Workshop, HCI-COLLAB 2021, Sao Paulo, Brazil, September 8–10, 2021, Proceedings 7, Springer, 2021, pp. 136–149.

[12] J. Droste, H. Deters, J. Puglisi, J. Klünder, Designing end-user personas for explainability requirements using mixed methods research, in: 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), IEEE, 2023, pp. 129–135.

[13] L. Chazette, W. Brunotte, T. Speith, Exploring explainability: A definition, a model, and a knowledge catalogue, in: 2021 IEEE 29th International Requirements Engineering Conference (RE), 2021, pp. 197–208. doi:10.1109/RE51729.2021.00025.

[14] W. Brunotte, A. Specht, L. Chazette, K. Schneider, Privacy explanations – a means to end-user trust, Journal of Systems and Software 195 (2023) 111545. doi:https://doi.org/10.1016/j.jss.2022.111545.

[15] R. Hamon, H. Junklewitz, G. Malgieri, P. D. Hert, L. Beslay, I. Sanchez, Impossible explanations? beyond explainable ai in the gdpr from a covid-19 use case scenario, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 549–559.

[16] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature machine intelligence 1 (2019) 389–399.

[17] L. Chazette, K. Schneider, Explainability as a non-functional requirement: challenges and recommendations, Requirements Engineering 25 (2020) 493–514.

[18] H. Deters, J. Droste, M. Fechner, J. Klünder, Explanations on demand - a technique for eliciting the actual need for explanations, in: 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), 2023, pp. 345–351. doi:10.1109/REW57809.2023.00065.

[19] M. N. K. Boulos, A. C. Brewer, C. Karimkhani, D. B. Buller, R. P. Dellavalle, Mobile medical and health apps: state of the art, concerns, regulatory control and certification, Online journal of public health informatics 5 (2014) 229.

[20] Y. Hong, K. Ehlers, R. Gillis, T. Patrick, J. Zhang, A usability study of patient-friendly terminology in an emr system, in: MEDINFO 2010, IOS Press, 2010, pp. 136–140.

[21] M. Grüning, T. Wolf, M. Trenz, A stressful explanation: The dual effect of explainable artificial intelligence in personal health management, in: Proceedings of the 57th Hawaii International Conference on System Sciences, 2024.