Evaluation of Quality Requirements for Explanations in AI-based Healthcare Systems

Zubaria Inayat

University of Twente, 7500 AE, Enschede, The Netherlands.

Abstract

In the field of explainable artificial intelligence (XAI), methods are being developed to explain AI results. These methods form the range of implementation choices available to XAI designers when dealing with the explainability requirements to a system. While in the discipline of Requirements Engineering, explainability has been conceptualized and operationalized as a non-functional requirement, there was so far little focus specifically on the quality aspects of the explanations themselves. Yet, quality requirements issues pertaining to the explanations of AI systems lead to issues such as lack of transparency, trust, and user confidence. The present PhD research makes a step towards closing this gap. The research aims to formulate a solution for determining the quality of explanations in AI systems, particularly in the healthcare domain. We believe that this research will benefit healthcare professionals in maintaining confidence and trust in AI-based healthcare systems.

Keywords

Requirements for Explanations, Quality Requirements, Explainable Artificial Intelligence, Healthcare, Artificial Intelligence in Medicine, Empirical Research Method.

1. Introduction

Artificial Intelligence (AI) has huge potential for bringing innovation in many domains [1]. Particularly, in the domain of healthcare, it is used for suggesting ways to prevent mistakes, assisting in disease diagnosis and treatment, and aiding in health records management for healthcare professionals [1] [2]. For example, in partnership with healthcare networks, Google leverages AI technologies to improve medical imaging and genomic analysis as well as algorithm-based screening for diabetic retinopathy; Google also builds predictive models from big data to warn clinicians of patients' high-risk conditions, such as sepsis, heart failure and blindness. Generally, explainable AI (XAI) [3] is thought to make the life of medical experts easier by supplying them with explanations supposed to help them understand the results rather than just believing the algorithmic processing. Largely, explanations are proven useful in almost all application domains, to make results of AI-based systems transparent and trustworthy. However, as explainability has only recently become a focus of intense research efforts in the Requirements Engineering (RE) community, some explainability aspects are not fully explored. In particular, the quality requirements concerning the explanations in AI-based systems and the assessment of explanations' quality have been

In: A. Ferrari, B. Penzenstadler, I. Hadar, S. Oyedeji, S. Abualhaija, A. Vogelsang, G. Deshpande, A. Rachmann, J. Gulden, A. Wohlgemuth, A. Hess, S. Fricker, R. Guizzardi, J. Horkoff, A. Perini, A. Susi, O. Karras, A. Moreira, F. Dalpiaz, P. Spoletini, D. Amyot. Joint Proceedings of REFSQ-2023 Workshops, Doctoral Symposium, Posters & Tools Track, and Journal Early Feedback Track. Co-located with REFSQ 2023. Barcelona, Catalunya, Spain, April 17, 2023.

EMAIL: z.i.ms.zubaria@utwente.nl

ORCID: 0000-0002-8515-2761

under-researched. The conceptualizations of explainability and its attributes are considered as a non-functional requirement (NFR) [13] and while those gain attention in the RE field, very little has been done until now to characterize the quality of the explanations themselves and to design metrics that help adequately assess the quality of explanations. This knowledge gap motivated us to initiate a PhD research project on the quality of explanations as our topic of interest. Our overall aim is to propose a framework for evaluation of quality requirements for explanations of AI-based systems in the healthcare domain.

2. Problem Statement and Relevance

As our interest is in healthcare, for this research we will use the term 'AI in medicine' (AIM) as defined by Payrovnaziri et al. [8] to mean "AI specialized to medical applications". Nowadays, all the healthcare organizations are moving towards digital record maintenance of their patients. In turn, the availability of the electronic health records has made the application of AIM much easier and smoother. However, in practice, it turns out that most of the currently employed AI-based assistive tools have limited scope of use and do not provide the required transparency for clinicians to establish an appropriate diagnosis, administer treatments, perform management of patient-related data, or maintenance of electronic health records [2]. Scholars (e.g. [1]) attribute the limited scope of AIM usage to the challenges to embed them into the actual clinical workflows in healthcare organizations and to the poor integration with existing healthcare systems. This in itself is a requirements misalignment problem. In fact, AIM systems have built-in assumptions (i) about the clinical workflow which they are part of and are supposed to support, and (ii) about the information needs of each of the medical experts that may happen to use them. These assumptions seem not to be realistic in many cases [1]. Furthermore, scholars (e.g. [11]) trace back the observation of insufficient transparency – and trust, to the lack of adequate explanations. While AIM systems augment healthcare practitioners' efforts to care for patients, many AI algorithms might well be hard to interpret, sometimes even by a qualified physician. As per the review of Carvalho et al. on state-of-the art of XAI [11], explanations are context-dependent, and nearly impossible to generalize; plus, there is still no method available to explainability engineers to evaluate explanations' quality. The present PhD project tackles the requirements misalignment problem from quality evaluation perspective and is set out to design and evaluate a possible solution to it.

This research has immediate relevance to both RE and healthcare. First, it responds to the call of RE researchers (e.g. [4][5]) for exploring new processes in support of complex systems' adaptation in various contexts. The quality requirements for explanations and their evaluation against context-relevant and stakeholders-relevant benchmarks have evaded so far the attention of RE researchers. In this sense, our project contributes to narrowing an existing gap of research in the field. Second, the intention of this PhD project is to add to the body of empirical RE knowledge on healthcare systems by providing an evaluative viewpoint into one particular type of requirements (quality requirements for explanations) in one specific context (healthcare) and from the perspective of the clinical workflow of medical practitioners working in that context. If RE practitioners working in the healthcare domain, have a framework of metrics models that could be used to assess the extent to which the quality requirements for explanations are satisfied or satisficed from the healthcare user perspective, then these RE practitioners could possibly design AI-based systems with more predictable transparency, and ultimately, trust and user confidence.

3. Related Work

The meaning of XAI has been explicated by several scholars (e.g. [12]). Many definitions along with the attributes of explainability had been proposed to elucidate the purpose of explainability in many

domains. In the RE community, consensus exists that explainability is an emerging NFR [13][21]. As such, RE scholars proposed ways to operationalize it [14] and possibly to quantify some of its aspects. While these works concern explainability as such, research on quality requirements for explanations and the quality assessment of explanations turn out to be scarce. To the best of our knowledge, there are only four publications [6][11][15][16] that at least partly treated our topic of interest. The study of Sarp et al. [6] focuses on improving the understanding of the results generated from complex AI models for the classification of human wounds. These authors propose additional explanations be given for lay users that are concerned with the AI-model-generated results. Furthermore, the research of Langer et al. [15] investigates how user satisfaction with AIM could be increased by improved understanding of the explanations generated by the AI models. Based on empirical data, the authors propose a model that stresses the quality metric of understanding (the explanations) for good performance by the concerned groups of users. Next, the literature review of Carvalho et al. [11] reports the state-of-the-art research on machine learning interpretability with a strong focus on the societal impact and on the methods and metrics proposed to assess the quality of the explanations. These authors found that while many proposals were put forward, there was very little empirical evaluation of those proposals and no comparative evaluation at all on what metric might be useful in what context. Finally, the work of Mittelstadt et al. [16] treats the topic of explanations and the context of XAI, from the perspective of philosophy of science. These authors examine research on explanations in philosophy, cognitive science, and social sciences in order to compare "the different schools of thought on what makes an explanation". Mittelstadt et al. argue that if XAI systems are to give good quality explanations, and explainability specialists should investigate how the quality of explanations is conceptualized in other disciplines where explanations play a central role.

Furthermore, as per the SLR on explainability methods of Vilone and Longo [6][17], when considering human understanding and performance as a measure of evaluating explanations, empirical researchers group the metrics into two categories: objective metrics (using automated approaches for evaluation) and human-centric metrics (using feedback and judgments from the users to evaluate). Furthermore, researchers referenced in this SLR concluded that there are no standards, no frameworks, and no consensus among the scholars to guide the quality evaluation of explanations. This conclusion agrees with the understanding of Bohlender and Köhl [18] from a RE perspective. The conclusion also agrees with the observation of Amparore et al. [19] about the lack of consensus on definitions of ''explanation quality" in the XAI literature. The quality of explanations is context-dependent because the understanding of explanations is tightly linked to the experience, the skill level and the background of the concerned user [11]. Moreover, even if methods of quality assessment are proposed, these do not completely address the challenges of quality evaluation of the AIM explanations in the healthcare domain [17].

4. Research Questions and Method

The overall goal of this PhD research is to propose a solution for quality evaluation of the explanations generated by the complex XAI systems working in the healthcare domain. To achieve this goal, we plan to answer the following research question (RQs):

RQ1: What are the methods or metrics of quality evaluation of explanations in AI systems in healthcare, according to published literature? The exploration of methods and metrics of quality evaluation will help to deeply understand the possible range of properties characterizing the quality of explanations in the healthcare domain. Knowledge on these properties and the ways in which each property could be quantified or qualitatively assessed, is needed in order to be able to operationalize the quality requirements for explanations in AI-based healthcare systems. In turn, equipping requirements engineers with possible operationalizations will help them specify quality requirements for explanations in verifiable fashion [9].

RQ2: What are the potential issues of quality evaluation for AI systems in the healthcare domain? While for XAI designers in the healthcare domain the systems' outputs might be clear, there is no standard available to evaluate the worth, accuracy, and appropriateness of the explanations for healthcare professionals. Therefore, we aim to identify the possible challenges in the evaluation of the explanation's quality from the perspective of medical experts. The knowledge of these challenges is instrumental to formulate goals and design criteria for our solution design [20].

RQ3: How to evaluate the quality of explanations generated for XAI systems in healthcare? RQ3 leads to the design of the new solution that addresses the issues identified with RQ2 and is expected to improve upon the existing solutions in literature. We envision our proposed solution to be a framework for evaluating he quality of explanations in AI systems.

RQ4: What is the usefulness of the proposed framework? As our proposed artifact is intended for use in practice, RQ4 helps us understand how useful our framework is to healthcare domain experts. We expect this solution to help the experts to build trust in the results produced by the XAI systems. RQ4 also aims to explore whether there will be a need for assistance in the decision-making process for healthcare practitioners by using our proposed method.

To answer our four RQs, we plan to use mixed method research methodology [20] and divide our work into three phases as shown in Figure 1:

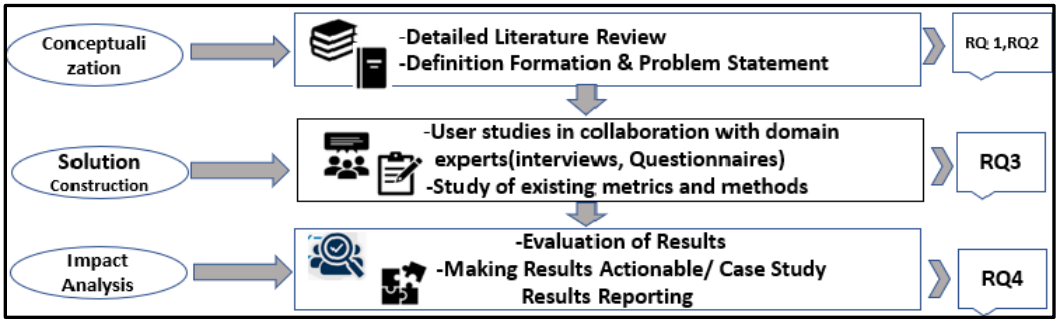

Figure 1 Research Design

Phase 1 addresses RQ1 and RQ2 by exploring the methods and metrics available in existing publications. We aim to use a systematic literature review (SLR) to find (i) the existing methods and metrics of quality evaluation (RQ1) and (ii) the issues (RQ2). ***Phase 2*** is focused on solution construction, i.e., the framework for the quality evaluation of the explanations in the healthcare domain (RQ3). This phase will be informed by the SLR from phase 1 and by qualitative interviews with domain experts. First, our framework, we will draw on previously published proposals that shed light on evaluating quality requirements for explanations of XAI. As these proposals are from other domains (and not healthcare), this phase also includes unearthing the proposals' tacit assumptions about the context of XAI use and evaluating the extent to which these assumptions might be realistic [20] to the healthcare context. Second, to assure the relevance of our proposal to RE practice, we will do qualitative interviews with two types of experts: explainability specialists in healthcare (responsible for the quality requirements for explanations, transparency and ethicality requirements), and medical experts (the users of the supplied explanations). Lastly, ***phase 3*** is about the validation of the proposed approach. For this, we plan to use two different strategies among those suggested by Wieringa [20]. One employs empirical evaluation techniques, namely through a selection of appropriate real-world cases for the application of the proposed framework, the selection and analysis of concomitant data, and reporting of the findings to RE researchers. We plan the cases to be selected based on specific criteria: (1) the clinical tasks for which XAI is used is diagnosis; we

chose this criterion, as currently XAI is deemed relatively most efficient in identifying the diagnosis of different types of diseases [10], (2) the clinical workflow supported by the system should be well understood by at least one non-technical stakeholder, (3) at least one domain expert is available to use to collaborate in the analysis of quality requirements for explanation and the verification of the extent to which these requirements are met in the AI system. Second, we also plan to gather data through questionnaires and expert interviews. These will be conducted both before (as indicated in phase 2) and after the construction of the proposed framework to analyze its usefulness and correctness. As explanations of XAI in healthcare cannot be generalized, we will maintain prolonged contact with the domain experts to get critical feedback for the issues under consideration over longer periods. We also plan to conduct focus groups with the domain experts to discuss the impact of our findings.

5. Proposed Solution

This section explains briefly our proposed conceptual framework. It will consist of (1) the description of the candidate metrics and their properties for quality evaluation of AIM explanations and (2) the guidelines for flexible implementation of the metrics and contextual guidance considering the ethical limitations of healthcare. This framework will help the explainability experts, particularly, and RE experts in general, to design better explainability models in future. The process of designing the framework is divided in three phases as explained below: Phase 1.**Conceptualisation:** In this phase the differentiation is established between the core concepts and terms that are often mixed up or used alternatively. This is important to set the baseline of our framework and to understand the true concept of the relevant terminology. Phase 2.**Solution Construction:** This phase of our proposed approach is based on the construction of our framework for quality evaluation. This will be grounded on the principles stated in the standards ISO 8402 and ISO 9402 for quality evaluation. E.g. in our framework, the criterion "fitness for use" stated in ISO 8402 and its related metric such as understandability will be operationalized and supplemented with evaluation guidelines, to help evaluate the quality of explanations of AIM systems. Phase 3.**Impact analysis**: This phase will determine the impact of the framework on the domain experts. It will determine the usefulness of the guidelines for quality evaluation.

5.1. Progress to date

The first year of this PhD project included a SLR on what is known about quality of explanations produced by AI systems, according to published empirical studies, and what quality metrics from other domains might be applicable to healthcare. At the time of writing this doctoral paper, the SLR is in the process of being finalized. One of our conclusions based on the SLR findings is that the quality evaluation of explanations cannot be considered complete until the quality metrics satisfy all related quality properties. For example, understandability is one of the quality properties that is often used as an alternative for explainability in a broader context. Understandability has multiple sub-properties or indicators such as (i) explanatory power, (ii) accountability for interdependent factors, and (iii) a sense of inference, among others; all indicators for those sub-properties must satisfy their respective acceptance criteria, in order to claim (sufficient) understandability. Currently, we are working on the mapping of indicators (metrics) and their properties which will be used as a basis for our proposed solution approach. Once this is done, we will be moving forward toward the formulation of the solution for the quality evaluation of explanations.

5.2. Novelty

The adoption of XAI systems is much slower in healthcare than in other domains because of a lack of trust in the results. The current proposals for generally-applicable metrics for the evaluation of the explanations do not seem to work specifically for critical domains such as healthcare. Next to this, our SLR indicates that literature so far does not provide any method to determine the quality of the explanations in healthcare. In light of this, the novelty of this PhD research is twofold: (1) to the best of our knowledge, this work is the first that treats the trust and quality of explanations in AI-based healthcare systems as a requirements misalignment problem. Drawing on prior work (e.g. [19]) we admit there is relationship between the requirements for trust and the quality requirements for explanations. The nature of this relationship is, however, unclear and as long as is unclear, it will be hard to come up with effective ways to align these requirements. This PhD project lays out a new foundation to solve the requirements misalignment problem by bringing new knowledge about the quality requirements for explanations in AI-based healthcare systems in terms of quality properties important to consider and operationalize when specifying and evaluating quality requirements for explanations; (2) the key deliverable of this PhD work will be a framework to evaluate the quality of the explanations in healthcare, along with guidelines for evaluation specialists on which metrics to consider as candidates for inclusion in the evaluation process based on suitability to context. To the best of our knowledge, this research project is the first attempt to design and evaluate such a framework.

6. Plan for Completion

Once the SLR is over, we would continue with a qualitative interview-based study to collect the perceptions of practitioners regarding the quality properties of explanations and the issues surrounding their evaluation. After this, both the findings from the SLR and from the interview-based study will be used to construct the solution for the evaluation of the quality of explanations in healthcare. The next major step then will be the empirical validation [20] and impact analysis of our framework. To this end, the first and foremost approach in our case will be the use of expert opinions to validate the usefulness and the utility of the framework from the perspective of professionals working in the field. We will also develop and deploy a task ontology for our proposed framework to figure out the relationship between the ontological information and its impact on the user (i.e. the impact on trust). This approach will not only validate our framework but will also answer our RQ4 which is intended to check the impact of our proposed solution. The main challenge that we foresee during this research is the willingness of the healthcare experts to participate in our empirical studies as research [16] indicates that there is lack of multidisciplinary efforts in this line of work. This risk is partly mitigated due to the healthcare research history of the department in which the PhD student works. To reach out to relevant experts, we would use the partnering healthcare organizations in the Netherlands who supported previous research collaborations with the involvement of the promoter and the daily supervisor of the PhD student.

7. Conclusion

This PhD research intends to contribute to the systematic management and improvement of quality of explanations of AIM systems. If these systems could consistently provide clear, unambiguous, and transparent results, this will encourage their implementation and usage and will also help overcome user resistance to AIM due to lack of trust. Using a mixed method research process, we will design and empirically evaluate a framework consisting of candidate metrics and guidelines for their selection and evaluation of explanation quality in AIM contexts. We hope this framework will at least partly alleviate the

practical and research-related challenges concerning the quality of explanations in AIM and the current misalignment of quality requirements concerning trust and those concerning explanations.

References:

[1] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 6(2), pp. 94-98, 2019

[2] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of AI technologies in medicine," *Nat. Med.*, vol. 25, no. 1, pp. 30–36, 2019

[3] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for AI in healthcare: a multidisciplinary perspective," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–10, 2020

[4] Cleland-Huang, J."Disruptive change in requirements engineering research", RE 2018, pp. 1–2.

[5] Gregory, S. What Does the Future Hold for Requirements Engineers? IEEE Softw. 39(4), pp. 18-21 (2022)

[6] S. Sarp, M. Kuzlu, E. Wilson, U. Cali, and O. Guler, "The enlightening role of explainable AI in chronic wound classification," *Electron.*, vol. 10, no. 12, 2021

[7] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing Design Practices for Explainable AI User Experiences," *Conf. Hum. Factors Comput. Syst. - Proc.*, 2020

[8] S. N. Payrovnaziri *et al.*, "Explainable AI models using real-world electronic health record data: A systematic scoping review," *J. Am. Med. Informatics Assoc.*, vol. 27, no. 7, pp. 1173–1185, 2020

[9] Lauessen, S. Requirements Specification Styles and Techniques, Wiley, 2000.

[10] Kumar Y, Koul A, Singla R, Ijaz MF. AI in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput. 2022

[11] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics (Switzerland)*, vol. 8, no. 8. 2019

[12] F. Hussain, R. Hussain, and E. Hossain, "Explainable Artificial Intelligence (XAI): An Engineering Perspective," pp. 1–11, 2021, [Online]. Available: http://arxiv.org/abs/2101.03613

[13] L. Chazette, W. Brunotte, and T. Speith, *Explainable software systems: from requirements analysis to system evaluation*, vol. 27, no. 4. Springer London, 2022

[14] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, 2018

[15] M. Langer *et al.*, "What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," *Artif. Intell.*, vol. 296, no. February, 2021

[16] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," *FAT* 2019 - Proc. 2019 Conf. Fairness, Accountability, Transpar.*, pp. 279–288, 2019

[17] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, vol. 76, no. April, pp. 89–106, 2021

[18] D. Bohlender and M. A. Köhl, "Towards a Characterization of Explainable Systems?," *arXiv*, pp. 1–11, 2019.

[19] E. Amparore, A. Perotti, and P. Bajardi, "To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods," *PeerJ Comput. Sci.*, vol. 7, pp. 1–26, 2021

[20] R. Wieringa, *Design science methodology*, Springer 2014

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