

Trustworthy enough? Evaluation of an AI decision support system for healthcare professionals

Kristýna Sirka Kacafirková, Sara Polak, Myriam Sillevs Smitt, Shirley A. Elprama, An Jacobs

imec-SMIT, Vrije Universiteit Brussel

Abstract

Exploring end-users' needs and involving them in designing and evaluating an AI system should be a priority. Understanding the system is essential to assess whether to trust it or not. This paper discusses a use case of a decision support system integrated into a platform for healthcare professionals (medical call operators and nurses). The system navigates them by predicting what intervention should be taken when an accident happens to a patient at home. Our use case demonstrates the importance of human-centred evaluation methods and potential struggles with mixed methods as detected by differences between qualitative and quantitative approaches. A subjective scale in combination with group interviews was used to evaluate the trust towards the system. The results showed that while users expressed a relatively high trust in the scale, the qualitative insights indicated uncertainty and the need for better explainability to trust the decision support system. In line with the results, we point out the need for better human-centred evaluation methods, as the current subjective scale needs to be complemented by qualitative methods to ensure rich insights.

Keywords

explainable AI (XAI), trustworthy AI, DSS in healthcare, XAI evaluation, explainability needs


1. Introduction


The awareness of the need for explainability² in AI systems is increasingly rising [9], especially in sectors such as healthcare, where outcomes recommended by the system have a large impact affecting humans' lives. Nonetheless, explanations are often lacking or not adapted to end-users' needs [2, 4, 12, 13]. To be able to detect the needs of end-users and enhance the system, evaluation is crucial. However, the evaluation of explainability in AI systems from a human-centred perspective is limited [9, 12, 14]. Developers mainly focus on the technical elements of the explanation, and the user's feedback is often overlooked [9, 11]. Therefore, we focus on evaluation by end-users and relevant stakeholders regarding the explainability and trustworthiness of a platform via existing tools. We show a use case of a healthcare platform with a decision support system (DSS) developed in a Protego³ project. Concretely, we set out to answer the following research question: *What are the limitations of current explainability subjective scales measuring trust in the DSS system based on AI?*

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

✉ kristyna.kacafirkova@vub.be (K. Sirka Kacafirková)

 <https://researchportal.vub.be/en/persons/krist%C3%BDna-sirka-kacaf%C3%ADrkov%C3%A1> (K. Sirka Kacafirková)

 © 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

² In this paper, we refer to explainability as an ability of users to understand how the system works (global explainability), or why a certain outcome was generated (local explainability).

³ <https://www.imec-int.com/en/research-portfolio/protego>

2. Related work

2.1. Evaluating explainability, evaluating trust

When evaluating explainability, it is crucial to acknowledge its link with trust [8]. Trust is perceived as one of the main reasons for implementing XAI explanations [4]. Users seek explanations primarily as an indicator of whether they can trust the outcome of a system or not [1]. However, trust should not be seen as something instantly achievable. It is a long-term process that the user develops continuously when using the system [5]. Using explanations should not aim for over (or under) trusting the AI system but rather an appropriate level of trust [6]. Especially in the healthcare context, too much trust can lead to fatal consequences, such as following an incorrect diagnosis.

Available human-centred evaluation methods for trust and explainability are limited [8, 9]. One of the most used evaluation tools in the human-centred domain are subjective scales [10]. Simply asking a participant if they trust the system via closed questions, however, often does not sufficiently show the reasons behind it [8]. To demonstrate the difficulties, we combined a subjective scale with group interviews to show the importance of using both quantitative and qualitative methods in evaluation.

3. Protego use case: Decision support system in healthcare

3.1. System description

A DSS was developed in collaboration with university researchers, industrial partners, and a local healthcare organisation in Belgium. This system suggests the necessary type of care based on the context of the alarm. A recommendation is based on health and behavioural data (from sensors inside the patient's home), personal information, and data collected by the call operator handling the alarm. Based on this information, the system suggests a next step of action: 1) call an ambulance, 2) send a nurse, 3) send an informal carer, 4) or dismiss the alarm. The call operator then decides what will be the subsequent step. A comprehensive overview of the alarm and its context is sent to the dispatched caregiver (e.g. a nurse), allowing them to follow up on the alarm accurately.

3.2. Data collection

A Proof-of-Concept (PoC) based on the evaluation of mock-up versions was developed and evaluated in two phases. For the first PoC group interviews, a pre-defined scenario of a diabetes patient was used to evaluate the PoC with seven stakeholders⁴: 2 medical call operators (W, 32, E-4, DW-4; W, 28, E-3, DW-3), 2 home nurses (W, 31, E-3, DW-3; W, 43, E-5, DW-10) and 3 ethical board members. A refined PoC was then presented, using a pre-defined scenario of a heart patient, in a final group interview with six stakeholders: one medical call-operator (W, 46, E-8, DW-15), one home nurse (M, 39, E-10, DW-10) and 4 ethical board members. Participants were invited to fill out a questionnaire at the end of each interview. For the first iteration, Cahour's and Forzy's [3] trust scale was used with additional questions specific to the platform. The second iteration also included Hoffman's trust scale [7], which is one of the most used scales for XAI evaluation to see if there is a significant difference when the scale is adapted to XAI. All participants signed informed consent. The sessions were organized at the research centre premises, interviews were audio recorded, and notes were taken. The notes were subsequently analysed through a thematic analysis. Questionnaire data were summarized using GoogleSheets, ChartExpo.

⁴ Where possible, we included data about: gender (W-woman, M-man, X-other), Age (in years), Experience in a role (E-number of years), digital working experience (DW – number of years of using a computer in their job)

4. Results: Evaluation and insights from user’s feedback

4.1. Subjective scores on trust, predictability, reliability, and decision support

The questionnaire data indicates that the overall level of trust in the system (for all participants) is neutral (3) to high (4-5), see Figure 1. In general, home nurses expressed higher trust than call operators, the ethical board were also less trustful than the professionals. Questions regarding the user interface showed that the interpretation of percentages of suggested actions was perceived as less understandable and thus lowered the trust score, which was also expressed during interviews.

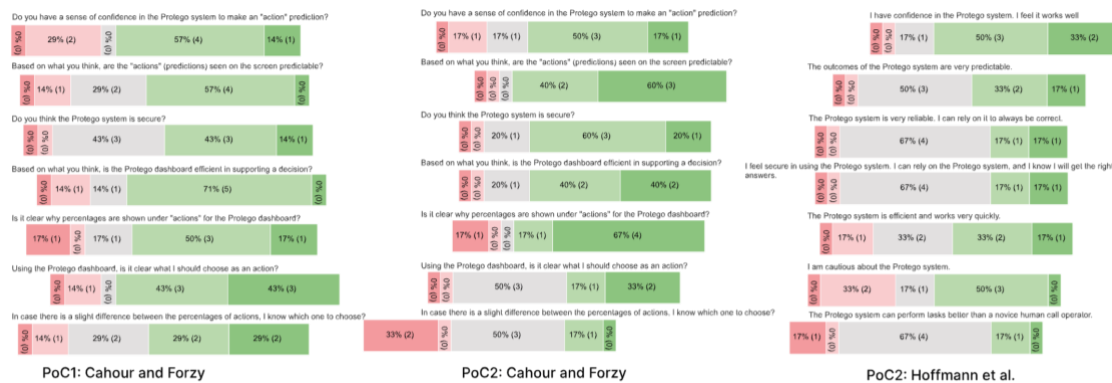


Figure 1: Overview of scale results from two PoC

4.2. The need for global explainability and simpler explanations

A sufficient explanation of “why the percentages were generated” was missing, affecting users’ trust in the outcome. However, as participants said, instead of making the user interface more transparent with more detailed information, they would be more interested in instructions on how the system works and when it is updated. Call operators expressed a need to understand the influence of responding to system questions on the actual recommendation. This might indicate that there was a high need for a global explanation of the recommender system rather than an explanation of each outcome (local explanation).

Participants explained that during the process of action, there is often no time to explore further what data means or why something was recommended. From a nurse’s point of view, in an emergency, most of the information collected from the sensors is simply irrelevant and unnecessary. Immediate action in such a situation is more important. **Call operators** also expressed that they would decide on follow-up action according to other factors that can be sensed only by previous experience, for example, nervousity in voice, difficulty with breathing etc., which AI cannot determine. On the other hand, call operators also indicated that personally reaching a different conclusion than the percentages show would make them doubt their skills and knowledge, and they would start to lose trust in themselves.

5. Discussion and conclusion

The subjective scales indicated rather higher trust and explainability in the system, even though no typical XAI explanation methods were integrated. However, as we found out during interviews, participants expressed an unfulfilled need for global explainability and were seeking explanations of how the system works. More than a need for local explanations, participants were

seeking to understand the logic behind the system and how answering questions in different ways can affect the outcome.

Furthermore, they tend to trust their experience more than the system. Even though the user interface offered detailed information via percentages and sensors, these were not useful in the work context of call operators and nurses, which is in line with Barda et al.'s study [2]. Too much information can be overwhelming and not always easy to interpret in a timely manner, which might also be the case if some XAI explanations had been implemented. This might indicate that new simpler explanations for users with non-technical backgrounds are needed to be explored. Nevertheless, it is important to point out that our study used a limited size of sample and was not rich considering the types of stakeholders. The evaluation is also in the initial phase, more iterations and capturing the trust over the time are needed. The use of pre-defined scenarios can also not sufficiently reflect the situation in the real world.

Besides, our study also showed that the evaluation of explainability and trust of AI systems from a human-centred perspective is still limited [9] and has room for improvement. For example, ethical board members pointed out that the scale is too generic and hard to assess while filling in the questionnaire. Based on the results, we demonstrate that a qualitative approach to end-users' perception is necessary. Future work could focus on creating and validating scales reflecting additional dimensions, such as users' usefulness, understandability, and satisfaction [10] that can affect users' trust.

Acknowledgements

This work is part of the PROTEGO project, which is an imec.icon research project funded by imec, Innoviris and Agentschap Innoveren & Ondernemen.

References

- [1] Amann, J. et al.: To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLoS Digital Health*. 1, 2, e0000016 (2022). <https://doi.org/10.1371/journal.pdig.0000016>.
- [2] Barda, A.J. et al.: A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Med Inform Decis Mak*. 20, 1, 1–16 (2020). <https://doi.org/10.1186/s12911-020-01276-x>.
- [3] Cahour, B., Forzy, J.-F.: Does projection into use improve trust and exploration? An example with a cruise control system. *Saf Sci*. 47, 9, 1260–1270 (2009).
- [4] Das, A., Rad, P.: Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. arXiv preprint arXiv:2006.11371. (2020).
- [5] Davis, B. et al.: Measure Utility, Gain Trust: Practical Advice for XAI Researchers. In: *Proceedings - 2020 IEEE Workshop on TRust and EXpertise in Visual Analytics, TREX 2020*. pp. 1–8 Institute of Electrical and Electronics Engineers Inc. (2020). <https://doi.org/10.1109/TREX51495.2020.00005>.
- [6] Ferrario, A., Loi, M.: How Explainability Contributes to Trust in AI. In: *ACM International Conference Proceeding Series*. pp. 1457–1466 Association for Computing Machinery (2022). <https://doi.org/10.1145/3531146.3533202>.
- [7] Hoffman, R.R. et al.: Metrics for Explainable AI: Challenges and Prospects Institute for Human and Machine Cognition. arXiv preprint. (2018).
- [8] Jacovi, A. et al.: Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In: *FACCT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. pp. 624–635 Association for Computing Machinery, Inc (2021). <https://doi.org/10.1145/3442188.3445923>.
- [9] Kim, S.S.Y. et al.: “Help Me Help the AI”: Understanding How Explainability Can Support Human-AI Interaction. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. pp. 1–17 (2023). <https://doi.org/10.1145/3544548.3581001>.
- [10] Lopes, P. et al.: XAI Systems Evaluation: A Review of Human and Computer-Centred Methods, (2022). <https://doi.org/10.3390/app12199423>.
- [11] Morley, J. et al.: From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Sci Eng Ethics*. 26, 4, 2141–2168 (2020). <https://doi.org/10.1007/s11948-019-00165-5>.
- [12] Sperrle, F. et al.: Should We Trust (X)AI? Design Dimensions for Structured Experimental Evaluations. In: arXiv preprint arXiv:2009.06433. (2020).
- [13] Waa, J. van der et al.: Interpretable confidence measures for decision support systems. *International Journal of Human Computer Studies*. 144, (2020). <https://doi.org/10.1016/j.ijhcs.2020.102493>.
- [14] Zhou, J. et al.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics (Switzerland)*. 10, 5, 593 (2021). <https://doi.org/10.3390/electronics10050593>.