Trust in Artificial Intelligence: Exploring the Influence of Model Presentation and Model Interaction on Trust in a Medical Setting

Tina Wünn[1,2], Danielle Sent[1(✉)], Linda W. P. Peute[2], and Stefan Leijnen[1]

[1] Research Group Artificial Intelligence, HU University of Applied Sciences, Utrecht, The Netherlands
danielle.sent@hu.nl

[2] Department of Medical Informatics, Amsterdam UMC Location University of Amsterdam, Amsterdam, The Netherlands

Abstract. The healthcare sector has been confronted with rapidly rising healthcare costs and a shortage of medical staff. At the same time, the field of Artificial Intelligence (AI) has emerged as a promising area of research, offering potential benefits for healthcare. Despite the potential of AI to support healthcare, its widespread implementation, especially in healthcare, remains limited. One possible factor contributing to that is the lack of trust in AI algorithms among healthcare professionals. Previous studies have indicated that explainability plays a crucial role in establishing trust in AI systems. This study aims to explore trust in AI and its connection to explainability in a medical setting. A rapid review was conducted to provide an overview of the existing knowledge and research on trust and explainability. Building upon these insights, a dashboard interface was developed to present the output of an AI-based decision-support tool along with explanatory information, with the aim of enhancing explainability of the AI for healthcare professionals. To investigate the impact of the dashboard and its explanations on healthcare professionals, an exploratory case study was conducted. The study encompassed an assessment of participants' trust in the AI system, their perception of its explainability, as well as their evaluations of perceived ease of use and perceived usefulness. The initial findings from the case study indicate a positive correlation between perceived explainability and trust in the AI system. Our preliminary findings suggest that enhancing the explainability of AI systems could increase trust among healthcare professionals. This may contribute to an increased acceptance and adoption of AI in healthcare. However, a more elaborate experiment with the dashboard is essential.

Keywords: trust · explainability · artificial intelligence · healthcare · dashboard

1 Introduction

Many countries have been experiencing rapidly rising healthcare costs and a shortage of medical staff. At the same time, the growing field of Artificial intelligence (AI) in healthcare aims to extract important information from data and assist in medical decision-making, offering potential solutions for cost and staffing issues, and promising improved

S. Nowaczyk et al. (Eds.): ECAI 2023 Workshops, CCIS 1948, pp. 76–86, 2024.
https://doi.org/10.1007/978-3-031-50485-3_6

healthcare outcomes. However, despite its potential, the adoption of AI in healthcare remains limited [1, 2]. One of the key barriers in implementation is lack of transparency of AI algorithms, which is the level to which the underlying operating rules and inner logic of the technology are understandable to the users [2, 3]. Explainable AI (XAI) is an emerging field in artificial intelligence that deals with methodologies and procedures that provide explainable models of why and how an AI algorithm produces predictions [4]. It addresses the need for transparency and interpretability in AI models, which historically have resembled a 'black box', delivering outputs without a clear understanding by the user of how they were arrived at. This lack of transparency/interpretability can pose significant challenges in sectors such as healthcare, where understanding the decision-making process is necessary for ethical and safety considerations. While XAI has promising prospects for healthcare, uncertainties persist about the kind of explanations that would be most suitable for healthcare professionals and how to present this information to help end users understand the AI [5, 6]. Transparency and explainability have shown to foster trust in AI systems in various contexts [7, 8], but their specific impact in healthcare, a high-risk setting, requires further exploration.

Our objective is to explore the subjects of trust and explainable AI in greater depth. We aim to better understand trust in AI, its relation to explainability, and the resulting implications for the healthcare domain. We strive to accomplish this designing a dashboard prototype that acts as an interface between AI models and end-users to present model characteristics and outputs to the user and to enable the user to interact with the model, with which we conduct an exploratory case study, assessing it for explainability, perceived ease of use, and perceived usefulness, while also determining users' levels of trust in the AI model.

2 Preliminaries

2.1 Trust

For a definition of trust, we use the one presented by Madsen and Gregor [9] who describe trust as 'the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid'. Jacovi et al. [10] define trust in the context of AI as a combination of a human's perception that the AI is trustworthy and the acceptance of vulnerability to its actions. In this case, the physician is aware that he is vulnerable to the risk of relying on the AI model, understanding possible adverse consequences. It is important to stress that the AI itself does not actually have to be trustworthy, the user only has to perceive it as being such. An AI model can be completely untrustworthy (e.g., always giving incorrect diagnoses), but if the user believes that it is trustworthy, they can still trust the AI. In other words, the correlation between trustworthiness and perceived trust can be very low. Here we can differentiate between warranted and unwarranted trust; trust is warranted if it is the result of trustworthiness, and otherwise it is unwarranted. It is also noteworthy that trust is very dynamic; it changes over time and over contexts, and once it is established it does not mean that it will stay [7].

Trust between two humans and trust between a human and AI depend on different factors [3]. These factors can be grouped in four categories: user (e.g., age, gender, understanding of technology), environment (e.g., task difficulty, perceived risks/benefits, task characteristics), model performance and traits (e.g., reliability, explainability, validity), and model presentation (e.g., transparency, appearance, ease of use) as shown in Fig. 1. It is difficult to know the relative influence that each factor has on trust, however, factors concerning technology seem to be more influential than factors related to the environment or the user [11]. Explainability, and concepts related to it such as transparency and reflections of reliability of AI, plays a significant role in establishing trust [8]. It might help with aligning users' expectations with the actual performance of the system, which is important for forming warranted trust.

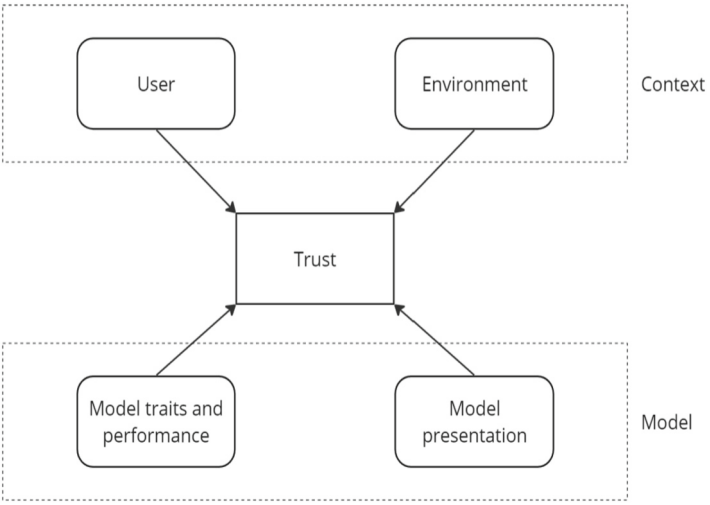

Fig. 1. Categories of factors influencing trust

Trust has been widely acknowledged to play an important role in user acceptance of technology. It has been incorporated many times into frameworks such as the Technology Acceptance Model (TAM), Unified Theory of Acceptance and Use of Technology (UTAUT) [12]. These frameworks help researchers understand and predict technology acceptance and adoption behaviours in various contexts. Abbas et al. (2018) have extended the Technology Acceptance Model (TAM) from a healthcare technology perspective [13]. They integrate trust as a factor that influences perceived ease of use, perceived usefulness, and behavioural intention. The TAM model described three categories of factors influencing trust: human-, organisational- and technology factors. We slightly adapted their model, renaming the first two categories and splitting the latter into two different categories to include the factors influencing trust in the way we have categorised them and arrive at the extended TAM in Fig. 2.

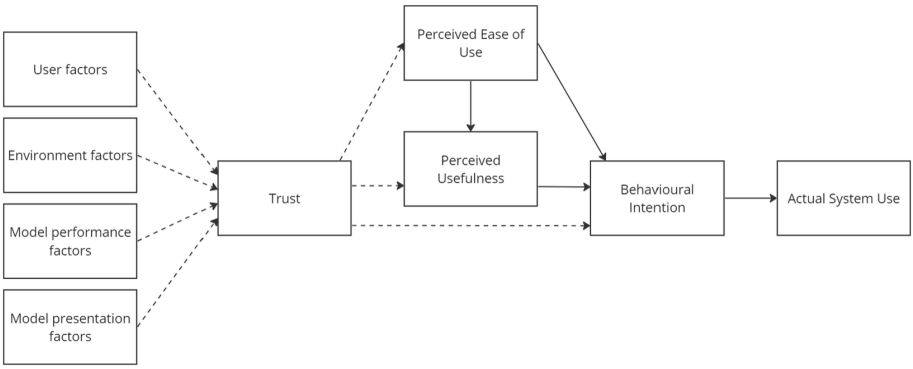

Fig. 2. Extended Technology Acceptance Model

Ideally, trust is assessed in a 'natural' situation and environment and not in a contrived experimental setup. As previously established, trust is a construct influenced by numerous factors, thereby developing differently across varied settings. This is, however, often not possible due to the high-risk context for healthcare settings, typically concerning patient safety and/or privacy. For example, measuring trust in a hospital decision-support system intended for the use in life-critical situations cannot be done in a real-life situation due to ethical concerns.

2.2 Explainable AI

Explainable Artificial Intelligence (XAI) is a field that is concerned with the development of new methods that explain and interpret AI models [14]. Local methods provide explanations that are restricted to single predictions, while global methods explain the whole model. An overview of categories of XAI is presented in Fig. 3.

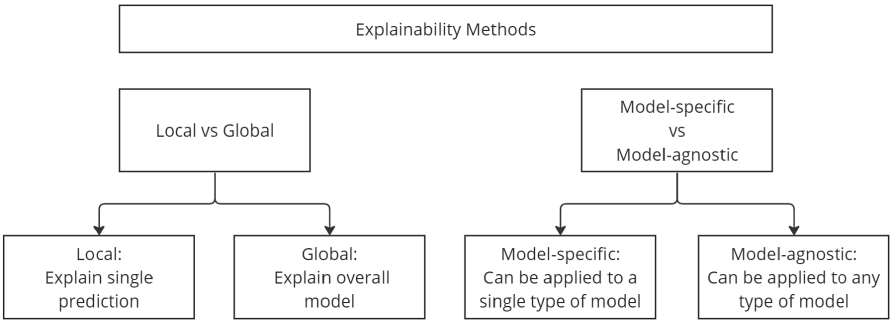

Fig. 3. Categories of XAI

The primary objective of Explainable AI (XAI) in healthcare is to assist medical professionals in understanding the underlying mechanisms that lead to specific results, thereby facilitating clear and comprehensible communication of medical decisions to

patients. However, existing XAI techniques are often designed with AI developers in mind, focusing on system or model evaluation rather than providing insights that healthcare professionals would find useful. This disconnect can make it challenging for medical practitioners to interpret the data, especially considering their possible limited technical expertise.

It has been argued that people attribute human-like traits to artificially intelligent agents and expect explanations about their behaviour to mirror those of humans [15]. Therefore, people would expect explanations about the behaviour of an AI system to be similar to an explanation about the behaviour of a human. Previous research has found that humans tend to form 'contrastive explanations', meaning they tend to explain the cause of an event in comparison to a counterfactual (another event that did not happen) rather than the event itself [16] : humans do not explain the event itself, but why it happened instead of some other, hypothetical, event.

2.3 Explainability and Trust in Artificial Intelligence

Developing trust(-worthiness) is one of the key motivations for XAI. The goal of explainable AI is for the user to develop warranted trust. This can be achieved by increasing the trustworthiness of the AI system itself, increasing the trust of the user in a trustworthy AI and increasing the distrust of the user in a non-trustworthy AI. Therefore, the goal of XAI is to target three concepts in our conceptualisation of trust: trustworthiness, warranted trust and warranted distrust as can be seen in Fig. 4.

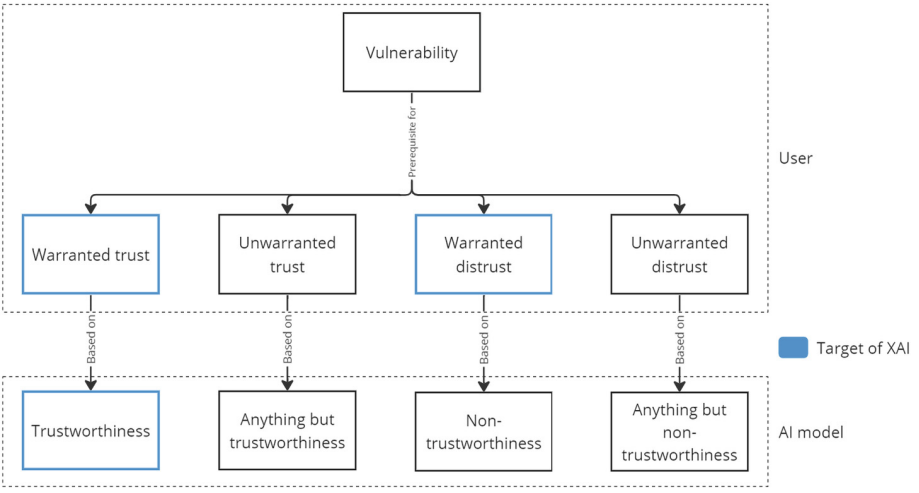

Fig. 4. Conceptualisation of trust, with the targets of XAI highlighted.

3 Method

3.1 Dashboard

If an AI-based decision-support tool would be implemented in healthcare, the interaction between the tool and the user would likely happen through a dashboard. In the following section, we present an illustrative example where we develop a dashboard prototype for the display of the output of DESIRE, an AI model that has been developed for optimising patient discharge after major surgery [17]. It predicts if a patient can be safely released from the hospital two days post-surgery. The objective of our dashboard is to present the recommendation/output by the DESIRE model along with additional information to make the AI model explainable to the user, and, consequently, increase warranted trust in the model.

The dashboard had several different options. First, basic information about the AI model was presented, such as its purpose, how it was trained, using which data, and a list of input features. This acts as global explanation. Four screens were intended for local explanations: 'Input features', 'Change values', 'Counterfactuals', and 'Similar patients'. The Input features screen displayed the input features of the model, their case-specific values, and the relative importance of the five most influential input features (Fig. 5a). This allows users to identify any abnormal values that might have arisen from faulty data entry, and that could explain potentially unexpected model predictions. By highlighting the top five influential input features, users gain a clearer understanding of the main drivers impacting the AI model's decisions. The screen Change values had the option to change input values or 'grey out' input values and recompute the advice for the given patient. The user could thus interactively explore the behaviour of the model and the impact of input feature variations on its predictions. It allows to test the robustness of the predictions under different hypothetical scenarios, which relates to the concept of contrastive explanation. The screen Counterfactuals displayed the minimal change needed in patient values that lead to a different prediction (Fig. 5b). This allowed users to compare their patient's prediction to a hypothetical contrary prediction, which also relates to contrastive explanation. For easier interpretation, the features where values have been modified in the counterfactual scenario are highlighted. This emphasises features that contribute to the shift in prediction. The last screen Similar patients showed patients who share similarities with the current patient, along with their respective predictions. The goal of this screen is to gain a broader understanding of how the model performs across a range of comparable cases. Inconsistencies in an AI model's recommendations for comparable patients, where healthcare professionals would anticipate consistent advice, can reveal discrepancies in model performance. At each of the screens a certainty score of the prediction was presented, which serves as an indicator of how certain the model is about its advice.

3.2 Participants and Setting

Participants in this study were healthcare professionals specialising in the surgical field. Participant recruitment was done through convenience sampling at the Dutch university medical centres Amsterdam UMC and Erasmus MC. Due to time constraints, only four healthcare professionals were willing to participate.

First, participants completed a questionnaire capturing characteristics such as age, gender, experience working and studying in the healthcare field, experience using AI, and their general attitude towards AI. Participants were shown static images of the five dashboard screens and after each one they were asked to fill out a questionnaire. The questionnaire was designed to measure two key aspects: the perceived explainability of the AI model and the level of trust in its predictions. The first screen that was shown to the participants was the plain screen containing basic information about the AI model. The order of presentation of the remaining screens was randomised to minimise the influence of order effects. Subsequently, participants were shown all the screens containing a local explanation again. They were asked to rank them, according to how useful they found them for their decision-making process. Perceived explainability was evaluated with the validated Explanation Satisfaction Scale [18]. Trust was measured using an adapted version of the Recommended Scale for explainable AI (XAI) [18]. Items that were not suitable in our context were dropped from the scale, either because they are only relevant after considerable use of a system, or they are not applicable for non-interactive screens. The four items taken from the Recommended Scale relate to how a user feels about an AI model and how trustworthy they perceive it to be. We added a fifth item asking the participant if they would follow the advice that the AI model gives. This incorporates the user's acceptance of vulnerability to the AI model's actions in the trust measurement. Additionally, two items from the Technology Acceptance Model (TAM) questionnaire were adapted to our context and added to assess Perceived Usefulness (PU) and Perceived Ease of Use (PEU). All items were rated on a five-point Likert scale. To analyse the results of the questionnaire, a trust and an explainability score were calculated. The scores were determined by quantifying and aggregating the answers to the respective questionnaire items.

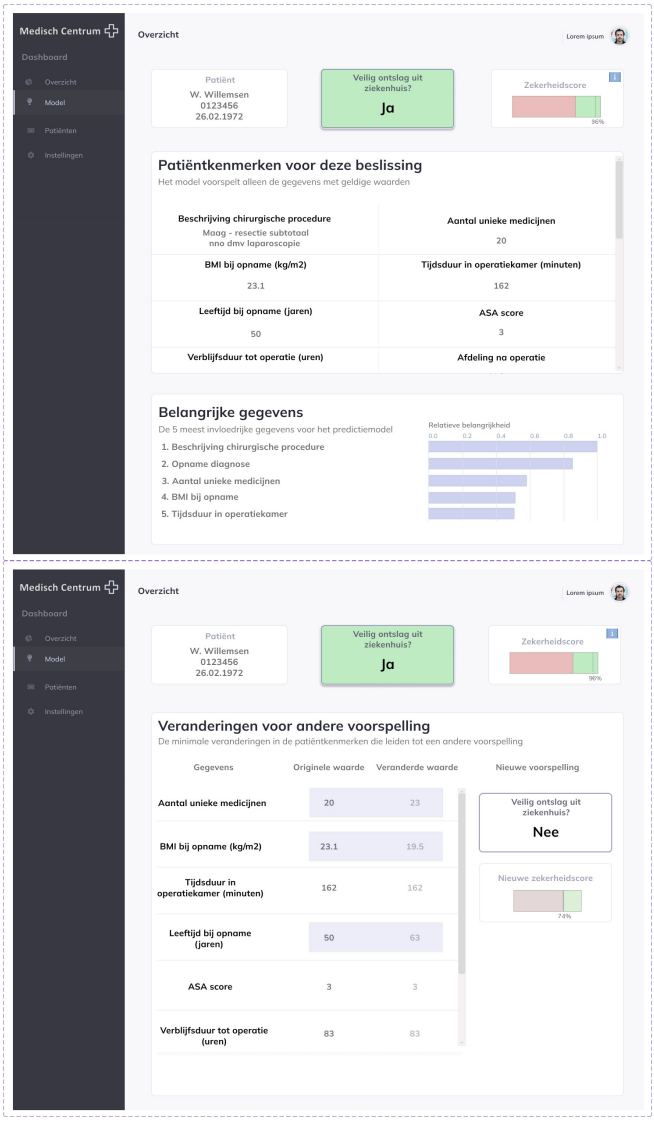

Fig. 5. Two screens of the dashboard. Figure 5a represents the input features and their relative importance. Figure 5b represents the counterfactuals and the values that will lead to a different conclusion. Translations: Veilig ontslag uit ziekenhuis = Safe discharge from hospital, Zekerheidscore = Certainty score, Patiëntkenmerken voor deze beslissing = Patient characteristics for this decision, Belangrijke gegevens = Important data, Veranderingen voor andere voorspelling = Changes for different prediction

4 Results

Based on the questionnaire concerning participant characteristics, there appears to be a predominantly positive attitude towards AI among the participants. Furthermore, the majority indicated having at least some level of experience with AI in both their daily lives and professional environments. The screen showcasing input features was found to receive the highest level of trust, while the counterfactual screen received the lowest score. No significant difference between them with regards to elicited trust was found. Similarly, for the explainability scores for the different screens, some variation can be seen, but no significant differences. The similar patients screen earned the highest score, while the change values screen received the lowest. Interestingly, the change values screen also exhibited the largest standard deviation, indicating more diverse opinions among the participants regarding its explainability compared to the other screens.

The input features screen received the highest score for PU, and both the input features and Plain screen share the highest score for PEU. Overall, participants ranked the similar patients screen as the most useful screen for decision-making. The change values and counterfactual screen share the lowest rank. Participants were most divided about the input features screen, with two assigning it the highest rank and two ranking it at the lowest.

5 Discussion

No significant difference in levels of trust and explainability was observed between the screens. This could imply that the individual participants' experiences and reactions remained comparably consistent, irrespective of the distinct components presented on each screen. Alternatively, it could suggest that individual variations existed, but these balanced out when analysed at the collective group level, resulting in no substantial differences. It is important to note that this does not prove that there is no real difference between the screens; we might not be able to detect it, either due to the small sample size or due to a small difference between screens.

Of particular interest is that the two screens based on the concept of contrastive explanation, namely Change values and Counterfactual, appeared to underperform across various metrics: they received the lowest trust levels, the lowest explainability scores, and were assigned the lowest rankings by participants. This finding suggests either that this kind of explanation is not ideal for making AI explainable in this specific context and that participants may not consider it particularly useful in their decision-making process, or it could indicate that our efforts to integrate this concept within a dashboard interface were not fully successful. In our results, we observed a strong positive correlation between explainability and trust. While we observed a strong positive correlation between explainability and trust in the AI system, it is important to remember that correlation does not imply causation. This relationship suggests that as explainability increases, so does trust. However, it does not necessarily mean that higher explainability causes an increase in trust. It is possible that other factors not accounted for in this study contribute to this relationship. Considering the small sample size, the results of this study should be interpreted with caution. Future studies with larger sample sizes would help

increase the reliability of the results. It would also be beneficial to involve participants with a diverse range of experiences with AI to increase the representativeness of the findings.

For a future, and larger, experiment, we intend not to only ask for trust directly, but ask the physicians to what degree they intend to accept (or not) the advice of the AI model, to make a direct connection between the action and trust in the system. It might also be interesting to use more than one patient case (as we did in this study) but have several cases such that the decision to be made is a different one during the experiment.

To be able to capture differences in perceived trust and explainability, we are also interested whether a co-called conjoint analysis with screens of our dashboard could provide us with more valuable results. During such an experiment, users are 'forced' to choose between two screens as to which of these is perceived to increase trust most.

Additionally, our study treated trust as a static quality, assessed at one point in time, rather than a dynamic process. Trust is likely to change with increasing interaction and experience with the system. Our conclusions can therefore only be applied to the initial exposure to a system. Future research could incorporate longitudinal designs with repeated measurements to capture the evolving nature of trust.

References

1. Peterson, E.D.: Machine learning, predictive analytics, and clinical practice: can the past inform the present? JAMA **322**(23), 2283–2284 (2019)
2. He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X., Zhang, K.: The practical implementation of artificial intelligence technologies in medicine. Nat. Med. **25**(1), 30–36 (2019)
3. Hoff, K.A., Bashir, M.: Trust in automation: integrating empirical evidence on factors that influence trust. Hum. Factors **57**(3), 407–434 (2015)
4. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inform. Fusion **58**, 82–115 (2020)
5. Liao, Q.V., Pribic, M., Han, J., Miller, S., Sow, D., Question-driven design process for explainable AI user experiences. arXiv preprint arXiv:2104.03483 (2021)
6. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J. Biomed. Inform. **113**, 103655 (2021)
7. Hoffman, R., Mueller, S.T., Klein, G., Litman, J.: Measuring trust in the XAI context. Technical Report, DARPA Explainable AI Program (2018)
8. Glikson, E., Williams Woolley, A.: Human trust in artificial intelligence: review of empirical research. Acad. Manag. Ann. **14**(2), 627–660 (2020)
9. Madsen, M., Gregor, S., Measuring human-computer trust. In: 11th Australasian Conference on Information Systems. Citeseer, vol. 53, pp. 6–8 (2000)
10. Jacovi, A., Marasovic, A., Miller, T., Goldberg, Y., Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 624–635 (2021)
11. Hancok, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., De Visser, E.J., Parasuraman, R.: A meta-analysis of factors affecting trust in human-robot interaction. Hum. Factors **53**(5), 517–527 (2011)
12. Ghazizadeh, M., Lee, J.D., Ng Boyle, L.: Extending the technology acceptance model to assess automation. Cogn. Technol. Work **14**, 39–49 (2012)

13. Abbas, R.M., Carroll, N., Richardson, I.: In technology we trust: extending TAM from a healthcare technology perspective. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI), pp. 348–349. IEEE (2018)
14. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**(1), 18 (2020)
15. De Graaf, M.M.A., Malle, B.F.: How people explain action (and autonomous intelligent systems should too). In: 2017 AAAI Fall Symposium Series (2017)
16. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
17. Van de Sande, D., et al.: Predicting need for hospital-specific interventional care after surgery using electronic health record data. Surgery **170**(3), 790–796 (2021)
18. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)