Impact of Fidelity and Robustness of Machine Learning Explanations on User Trust

Bo Wang, Jianlong Zhou, Yiqiao Li, and Fang Chen

University of Technology Sydney, Sydney, Australia
Bo.Wang-11@student.uts.edu.au
Jianlong.Zhou@uts.edu.au
Yiqiao.Li-1@student.uts.edu.au
Fang.Chen@uts.edu.au

Abstract. EXplainable machine learning (XML) has recently emerged as a promising approach to address the inherent opacity of machine learning (ML) systems by providing insights into their reasoning processes. This paper explores the relationships among user trust, fidelity, and robustness within the context of ML explanations. To investigate these relationships, a user study is implemented within the context of predicting students' performance. The study is designed to focus on two scenarios: (1) *fidelity-based scenario* — exploring dynamics of user trust across different explanations at varying fidelity levels and (2) *robustness-based scenario* — examining dynamics of in user trust concerning robustness. For each scenario, we conduct experiments based on two different metrics, including self-reported trust and behaviour-based trust metrics. For the fidelity-based scenario, we find that users trust both high and low-fidelity explanations compared to without-fidelity explanations (no explanations) based on the behaviour-based trust results, rather than relying on the self-reported trust results. We also obtain consistent findings based on different metrics, indicating no significant differences in user trust when comparing different explanations across fidelity levels. Additionally, for the robustness-based scenario, we get contrasting results from the two metrics. The self-reported trust metric does not demonstrate any variations in user trust concerning robustness levels, whereas the behaviour-based trust metric suggests that user trust tends to be higher when robustness levels are higher.

Keywords: Human computer interaction · Machine learning explanation · User trust · Fidelity · Robustness.

1 Introduction

Machine learning (ML) finds widespread applications in various domains, playing a pivotal role in numerous contexts. However, the lack of interpretability poses a significant challenge in understanding the inner workings of ML models. Hence, the explanation of machine learning holds the utmost importance. Explaining ML involves elucidating the intricate connections between input and outcome within ML models, facilitating user comprehension of the underlying reasoning. By elucidating the mechanisms of ML models, users can enhance their trust in the model's decisions, gain interpretability of the results, and gain insights into the decision-making process [20]. Moreover, explaining machine learning provides researchers, developers, and decision-makers with opportunities to gain deeper insights and improve the models. Recently, the field of ML explanation has obtained considerable attention from researchers. For instance, in the domain of recommender system [18], image classifier [12], and medicine [7], the researchers demonstrate that users express deeper insights when provided with explanations than systems lacking explanatory capabilities.

Furthermore, the selection of appropriate ML explanation methods with superior performance hinges upon the quality of the explanations. The quality of ML explanations encompasses three crucial aspects: user-related factors (e.g. user trust and satisfaction), explanation-related factors (e.g. fidelity), and model-related factors (e.g. robustness and fairness) [8]. User trust, as a critical aspect in ML explanations, represents one of the primary objectives in the explanatory process. It serves as a measurable criterion for quantifying subjective evaluation and enables assessing the quality of ML explanation methods. Further, fidelity holds significant importance in eXplainable machine learning (XML) as it ensures the provision of reliable explanations that align with the internal mechanisms of the underlying ML model. Recent studies confirm the correlation between explanation fidelity and user trust, emphasizing the need for high-fidelity explanations [11]. However, a fundamental question arises: **Does the user exclusively trust high-fidelity explanations?** To comprehensively address this inquiry, we devise a *fidelity-based scenario*, which builds upon the work of Papenmeier et al. [11] and further explore user trust variation at different levels of fidelity by visualizing explanations from two distinct methods. Alternatively, robustness in XML methods refers to their inherent capability to consistently provide reliable and consistent explanations, even when subjected to diverse perturbations. It is natural to raise the question: **Does robustness affect user trust?** In response to this research question, we design a *robustness-based scenario*, which explores user trust variation at different levels of robustness through visualization of explanations from a single method.

To thoroughly evaluate the impact of explanations on human trust within ML systems, researchers can adopt a combined approach utilizing a self-reported trust scale and behavioural metrics [13]. Thus, our paper assesses user trust from subjective and objective components by developing self-reported and behaviour-based trust metrics. To facilitate this evaluation, we employ a user survey that predicts student performance levels and measures variations in user trust across different levels of fidelity and robustness.

In this study, we make the contributions as follows:

- We evaluate how user trust varies on different explanation methods over different levels of fidelity. Specifically, we conduct a comparative analysis of user trust within the explanations generated by LIME and SHAP, incorporating three distinct levels of fidelity: high, low, and without-fidelity.

- We investigate how user trust fluctuates across different levels of robustness. To evaluate user trust, we employ visualizations of the explanations from LIME, encompassing both high and low levels of robustness.
- We employ a developed self-reported trust questionnaire and behaviour-based trust metrics to measure user trust. These measurement approaches allow us to capture subjective perceptions of trust reported by users, along with objective indicators derived from user behaviours and interactions.

2 Related Work

2.1 User trust

User trust in XML has been identified as a pivotal factor influencing human behaviour in human-machine interactions [21]. Users tend to base their behaviours on the guidance provided by well-performing XML systems when they trust the system. Conversely, if an XML system makes noticeable mistakes, it can lead to mistrust or even complete distrust of the system, causing users to deviate from following its recommendations. Several researchers emphasize that user trust would affect the adoption of ML. On specific, Asan et al. [2] argue that user trust is recognized as one mediator that influences clinicians' use and adoption of ML. Similarly, Shin and Park [15] highlight that user trust plays a crucial role in shaping potential adopters' willingness to undertake the inherent risks involved in adopting algorithm services. Likewise, the results in this paper [16] show that user trust acts as a liaison and interface between heuristic and systematic processing, facilitating ML service adoption. While user trust is inherently a subjective experience, Schmidt and Biessmann [14] build a function that establishes a link between the quality of ML explanation and user trust. This metric aids in identifying whether individuals are more biased toward the predictions made by ML systems.

2.2 Fidelity

Fidelity stands out as a vital property that impacts the quality of ML explanation. Indeed, its importance lies in the fact that a high-fidelity explanation can provide accurate and valuable information, encompassing the identification of important features and their significance to users. On the contrary, a low-fidelity explanation may result in the provision of meaningless insights. In essence, fidelity serves as a metric to measure how well an explanation method can mimic the behaviours of the underlying model [8]. Specially, Moradi and Samwald [9] employ fidelity as a metric to evaluate their proposed explanation method called Confident Itemsets Explanation, demonstrating its superior simulation of the underlying model compared to other ML explanation methods. Numerous researchers have devoted substantial efforts to exploring fidelity in ML explanations. For instance, Dai et al. [6] develop a quantitative metric for measuring fidelity to evaluate the precision of the explanations. Besides, the paper [11] incorporates a user study focusing on textual explanations, which reveals that explanations with low fidelity significantly decrease user trust levels. Also, we recognize the significance of the relationship between fidelity and user trust in this study. However, our primary focus lies in exploring the associations between different explanation methods at varying levels of fidelity and user trust.

2.3 Robustness

The robustness of an explanation method refers to its sensitivity towards minor changes in the input, resulting in both prediction variations and corresponding adjustments in the explanation [3]. Alvarez-Melis and Jaakkola [1] utilize the metrics quantifying the robustness of explanation to evaluate the performance of current explanation methods (e.g. LIME and SHAP). Their findings highlight that while these methods provide explanations, they exhibit notable sensitivity to even slight input variations, thereby affecting their reliability. Moreover, the significance of robustness in XML has been extensively discussed in the literature [4, 19]. Chan and Darwiche [4] introduce their algorithms to maintain the robustness of the Most Probable Explanation, thus facilitating the design and debugging of Bayesian networks in the presence of parameter changes. Furthermore, Tocchetti et al. [19] emphasize the importance of robustness in Graph Neural Networks (GNN) due to their vulnerability to adversarial attacks, where minor input alterations can lead to substantial output impacts. Considering the critical role of robustness, the expectation for the explanation method goes beyond mere reasonability, demanding even greater robustness. Our work focuses on exploring the connection between robustness in explanations and user trust through a user study.

3 Hypotheses

To elucidate our research questions, we formulate three hypotheses: H1 and H2 for the fidelity-based scenario, and H3 for the robustness-based scenario.

- H1: The level of user trust is impacted by the fidelity of explanations, wherein high-fidelity explanations uniquely contribute to higher user trust;
- H2: The level of user trust is influenced by distinct explanation methods when the level of fidelity remains constant;
- H3: The level of user trust is contingent upon the robustness of the XML. We posit that the higher robustness of the XML results in higher user trust.

4 Methodology

To investigate our three hypotheses, we conduct a case study focused on predicting student performance levels. We design two rounds, each comprising eight tasks, with consideration for both fidelity (six tasks) and robustness (two tasks).

4.1 Fidelity-based scenario study design

Fidelity denotes the extent to which an explanation accurately reflects the underlying model. To assess hypotheses H1 and H2, we incorporate three defined conditions: high-fidelity, low-fidelity, and without-fidelity. Under these conditions, we present visualizations employing two distinct explanation methods, LIME and SHAP (as delineated in Table 1). We establish the importance of features using the permutation importance method, as illustrated in Figure 1, which serves as a ground truth for generating varying levels of fidelity in explanations.

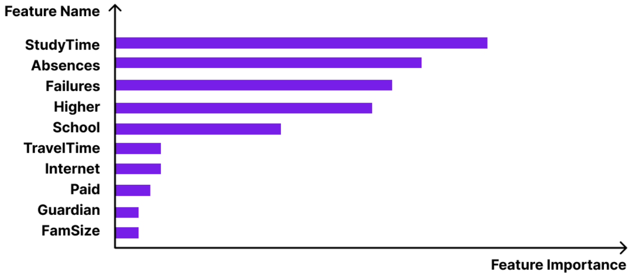


Fig. 1. Ground truth of feature importance.

- High-Fidelity: Under this condition, explanations effectively identify significant features that influence the predictions made by machine learning models. In this study, obtaining high-fidelity explanations involves generating explanations from LIME and SHAP, followed by manual adjustments to align them with the ground truth of feature importance.
- Low-Fidelity: In this condition, the explanations insufficiently identify significant features that impact the predictions made by machine learning models. To create low-fidelity explanations, substantial modifications are applied to the important features, deviating significantly from the ground truth of feature importance.
- Without-Fidelity: In this condition, both explanation methods exhibit none fidelity, effectively equating to a complete absence of any explanations.

4.2 Robustness-based scenario study design

Robustness in XML reflects that similar input should lead to similar explanations, characterized as insensitivity. To validate H3, two conditions are designed, high-robustness and low-robustness, wherein the LIME explanation method is exclusively employed (as illustrated in Table 1).

- High-robustness: In this condition, it is expected that the explanations remain relatively stable under minor feature modifications. In our implementation, achieving high-robustness explanations involves deliberately introducing slight variations in the weight and value of both important and unimportant features within subsequent explanations derived from high-fidelity explanations.

- Low-robustness: In this condition, explanations undergo significant changes when minor feature modifications are made. To deliver participants the low-robustness explanations, the condition is distinguished from the high-robustness condition by visualizing and simulating the explanations with considerable fluctuations of feature weights.

Table 1. Task set up in the experiments.

Methods	Level of Fidelity			Level of Robustness	
	High	Low	None	High	Low
LIME	Task 1.1	Task 2.1	Task 3.1	Task 4.1	Task 4.2
SHAP	Task 1.2	Task 2.2	Task 3.2	——	——

4.3 Metrics

To measure user trust, self-reported and behavior-based trust metrics are used.

Self-reported trust metric. The self-reported user trust (subjective user trust) level is assessed through a series of five subjective questions, with the first three questions derived from the measurements [17] and the last two questions formulated based on the metrics [10]. The self-reported user trust level is assessed on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

- I trust the results from the ML explanation system.
- The ML explanation system is trustworthy.
- I believe that the results from the ML explanation system are reliable.
- I believe that the ML explanation system can explain well the reasons behind students' performance.
- I believe that the ML explanation system can provide a detailed explanation for each student's performance.

Behaviour-based trust metric. We introduce this approach that takes into account the participants' behaviours in fidelity and robustness-based scenario experiments. It quantifies behaviour-based trust (objective user trust) as the frequency of appropriate decisions made by participants out of the total decisions undertaken. A higher average frequency indicates that participants exhibit higher user trust level in completing the tasks.

5 Experiment

5.1 Dataset

The student performance dataset in secondary education is employed as the foundational data source for this study. The original dataset, sourced from the UCI machine learning repository [5], contains 649 instances and comprises 30 nominal attributes which determine student numeric grades from 0 to 20. To streamline the complexity of this study, the dataset is reduced to 10 attributes, namely: school, guardian, paid, higher, famsize, traveltime, studytime, failures, absences, and internet. Furthermore, students' grades are categorized into distinct performance levels (A^+, A, B, C, D, and F). XGBoost is employed for performance level prediction, utilizing 70% of data for training and the rest for testing.

5.2 Participants

30 participants (15 males, 14 females, and 1 participant who prefers not to disclose their gender) were invited to participate in this study through our social networks. The participants' ages were distributed as follows: 18-24 years (6 participants), 25-34 years (24 participants), and 34-50 years (1 participant). Among the participants, 15 had completed their master's degree, followed by 12 participants who held bachelor's degrees, 2 participants were current Ph.D. students, and 1 who had completed an honours degree.

5.3 Experimental procedure

The experiment is deployed on the Qualtrics platform. Participants begin with a welcome page that introduces the researchers and outlines the study's objectives. Following this, to proceed with the study, participants are required to provide their explicit consent from the consent form page. Prior to commencing the study tasks, participants are afforded the opportunity to familiarize themselves with the process through example and feature information pages.

Subsequently, the main study commences, where participants are presented with a random task. Each task involves one student instance with ten features, accompanied by its corresponding performance level and the explanations generated by a specific explainability method. Depending on the task types (fidelity or robustness), the requirements to complete the tasks are different. For fidelity-based tasks, participants are asked to make predictions with the supply of new student instances based on the ground truth of feature importance and provided explanations. In contrast, robustness-based tasks require participants to make predictions based on the new student instance, considering two successive minor modifications in the student instance and updated visualizations of the explanations. Upon the completion of each task, participants rate their level of user trust concerning the explanations using a 5-point Likert scale questionnaire.

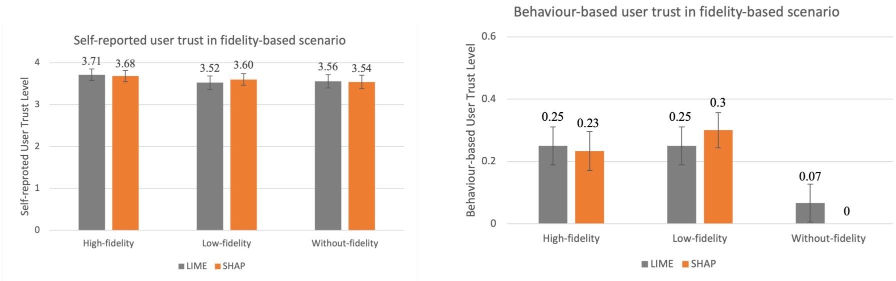
Following the completion of the 16 tasks across the two rounds, participants are then required to fill out a demographic questionnaire, providing information on their gender, age, education level, and familiarity with machine learning explanations. The distribution of 16 tasks is randomized to ensure the prevention of bias in the collected results. The entire study is estimated to take participants between 15 to 25 minutes to complete.

6 Results

This section presents the results obtained from the experiments conducted in both the fidelity and robustness scenarios. To acquire these results, we perform a series of statistical tests, including one-way ANOVA tests, Tukey's HSD post-hoc tests, and paired t-tests to analyse the variations in user trust within these two scenarios.

6.1 Correlations between user trust and fidelity.

In order to examine the validity of H1 and H2, respectively, we initially implement one-way ANOVA tests and Tukey's HSD post-hoc tests to assess user trust variations in the different fidelity levels for each explanation. Subsequently, we conduct paired t-tests to further evaluate user trust differences when comparing two explanation methods. The implementation of Tukey's HSD post-hoc tests is more effective in reducing Type I errors when conducting multiple comparisons within a condition compared to paired t-tests.



(a) Variations in user trust across three fidelity levels based on the self-reported trust metric for both LIME and SHAP

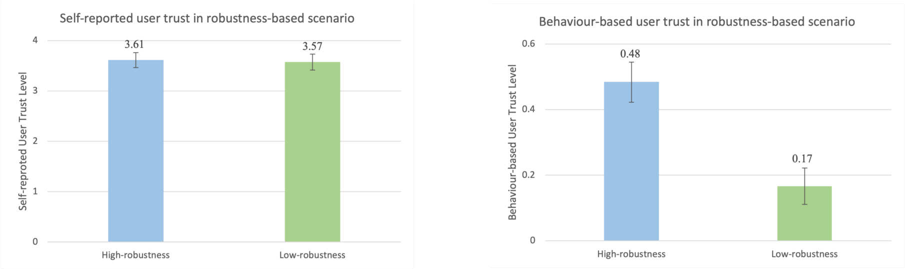
(b) Variations in user trust across three fidelity levels based on the behaviour-based trust metric for both LIME and SHAP

Fig. 2. Fidelity-based scenario: self-reported and behaviour-based trust metrics.

In the fidelity-based scenario experiment, variations in user trust level are first investigated among three fidelity levels when explanations are derived from the same methods. Self-reported user trust results indicate that participants trust high, low, and without-fidelity explanations. However, behaviour-based user trust results show trust in both high and low-fidelity explanations, while no trust in without-fidelity explanations. Figure 2(a) depicts the average subjective user trust levels across two different explanations at high, low, and without-fidelity levels (error bars correspond to standard errors and it is the same in other figures). One-way ANOVA tests at a 5% significance level (it is the same in other tests) indicate no significant user trust differences for both of the explanation methods based on the self-reported trust metric, thereby rejecting H1.

Nonetheless, one-way ANOVA tests find that there are significant user trust differences in the LIME ($F(2, 87) = 3.839$, $p = .025$) and SHAP ($F(2, 87) = 10.500$, $p < .000$) respectively when examining the behaviour-based user trust (see Figure 2(b)). Then Tukey's HSD post-hoc tests are performed to explore the pair-wise user trust differences between the fidelity conditions for each explanation method. In the explanations generated by LIME and SHAP, Tukey's HSD post-hoc tests find that participants have higher user trust when the explanations have high-fidelity than those without-fidelity (LIME: $p = .048$; SHAP: $p = .003$). Participants also show higher user trust when the explanations have low-fidelity compared to those without-fidelity (LIME: $p = .048$; SHAP: $p < .000$). However, participants show the same user trust when the explanations are with high-fidelity compared to those low-fidelity (LIME: $p > .05$; SHAP: $p > .05$). These findings reject H1 by implying that users trust high and low-fidelity explanations while untrusting ones without fidelity.

On the other hand, variations in user trust level are then analysed between explanations from two different methods when keeping fidelity consistent. No statistically significant differences in user trust among the consistent fidelity levels have been observed when comparing the two explanation methods, as evidenced by both self-reported and behaviour-based trust metrics (see Figure 2(a) and Figure 2(b), respectively). The paired t-tests do not reveal any statistically significant differences in user trust across different explanation methods at each fidelity level. These outcomes, based on both self-reported and behaviour-based trust metrics, lead to the rejection of H2.



(a) Variations in user trust in the high and low-robustness based on the self-reported trust metric for LIME.

(b) Variations in user trust in the high and low-robustness based on the behaviour trust metric for LIME.

Fig. 3. Robustness-based scenario: self-reported and behaviour-based trust metrics.

6.2 Correlations between user trust and robustness.

To evaluate the validity of H3, we perform paired t-tests to quantify user trust differences with respect to the robustness level based on both the self-reported and behaviour-based user trust levels.

In the robustness-based scenario, variations in user trust level are examined between the high and low levels of robustness when a single explanation method is employed. The results of the scenario demonstrate no significant differences in user trust between the high and low robustness levels through the analysis of the self-reported user trust levels. Conversely, analysing the behaviour-based user trust levels shows a positive correlation between user trust and robustness level. Figure 3(a) illustrates the average subjective user trust level for the high and low-robustness levels, based on the self-reported trust metric. The results of the paired t-tests, suggesting no statistically significant differences in user trust levels between the high and low robustness levels, reject our H3 based on the self-reported trust metric.

However, Figure 3(b) depicts the mean objective user trust levels for the high and low-robustness levels, respectively. Another paired t-test is conducted to analyse the objective user trust levels, revealing statistically significant differences in user trust levels between the high and low robustness conditions ($t = 3.842$, $p < .000$). These results suggest a positive correlation between user trust and robustness, indicating that higher robustness levels lead to higher user trust. As a result, our H3 is confirmed based on the behaviour-based trust metric.

7 Discussion

This study investigates the user trust differences in the fidelity and robustness of ML explanation under a specific condition, respectively. In the fidelity-based scenario, the analysis of the results of the self-reported trust metric indicates that participants do not exhibit significant differences in trust toward explanations with different levels of fidelity. On the contrary, the analysis of the results of the behaviour-based trust metric shows that participants trust both high and low-fidelity explanations, while a majority of them do not trust the absent explanations when fidelity levels are comparable. Furthermore, both self-reported and behaviour-based trust metrics indicate that there are no significant differences in user trust among the consistent fidelity levels when comparing the two explanation methods. In the robustness-based scenario, our findings show no significant differences in user trust between high and low-robustness explanations, as determined through the self-reported trust metric. However, we obtain a positive correlation between robustness levels and user trust based on the behaviour-based trust metric.

Our findings carry significant implications for the evaluation of the quality of ML explanations. For example, if a unified method for measuring the quality of ML explanations is established, fidelity and robustness could not only serve as contributing factors to this method but also complement the evaluation of user trust. Moreover, participants significantly trust both high and low-fidelity explanations compared to without-fidelity explanations. This suggests that explanations incorporating fidelity can effectively build user trust in ML applications. However, trust in high and low fidelity may pose a challenge for users as they attempt to discern the reliability of the provided explanations. Additionally, the observed positive correlation between user trust and robustness, as assessed through the behaviour-based trust metric, can be applied in the implementation of algorithms to enhance user trust in ML systems.

8 Conclusion and Future Work

The quality of ML explanations significantly influences the selection of effective explanation methods, encompassing user-related, explanation-related, and model-related aspects. Among these aspects, user trust holds particular importance and serves as a key objective in the explanatory process. Both fidelity and robustness also play essential roles in shaping the quality of explanations and ultimately impacting user trust. Through an investigation of the relationships between user trust, fidelity, and robustness, we contribute to a comprehensive understanding of these factors in the context of ML explanations. Our findings indicate that 1) participants trust both high and low-fidelity explanations than without-fidelity explanations in the behaviour-based user trust, contrary to the results from self-reported user trust; 2) no significant differences in user trust exist when comparing explanations with consistent fidelity level; 3) a higher robustness level positively influences user trust in the behaviour-based user trust analysis rather than in the self-reported user trust analysis. Moving forward, our future work aims to establish a comprehensive metric that combines subjective perceptions and objective properties to effectively evaluate the quality of ML explanations. Additionally, we recognize the need for improvement in the visualization of explanations in our user study.

References

1. Alvarez-Melis, D., Jaakkola, T.S.: On the Robustness of Interpretability Methods (Jun 2018), http://arxiv.org/abs/1806.08049, arXiv:1806.08049 [cs, stat]
2. Asan, O., Bayrak, A.E., Choudhury, A.: Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. Journal of Medical Internet Research **22**(6), e15154 (Jun 2020). https://doi.org/10.2196/15154, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada
3. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics **8**(8), 832 (Aug 2019). https://doi.org/10.3390/electronics8080832
4. Chan, H., Darwiche, A.: On the Robustness of Most Probable Explanations (Jun 2012), http://arxiv.org/abs/1206.6819, arXiv:1206.6819 [cs]
5. Cortez, P.: Student Performance. UCI Machine Learning Repository (2014), DOI: https://doi.org/10.24432/C5TG7T
6. Dai, J., Upadhyay, S., Aivodji, U., Bach, S.H., Lakkaraju, H.: Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post hoc Explanations. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. pp. 203–214 (Jul 2022). https://doi.org/10.1145/3514094.3534159, arXiv:2205.07277 [cs]

7. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? (Dec 2017), http://arxiv.org/abs/1712.09923, arXiv:1712.09923 [cs, stat]
8. Löfström, H., Hammar, K., Johansson, U.: A Meta Survey of Quality Evaluation Criteria in Explanation Methods (Mar 2022), http://arxiv.org/abs/2203.13929, arXiv:2203.13929 [cs]
9. Moradi, M., Samwald, M.: Post-hoc explanation of black-box classifiers using confident itemsets. Expert Systems with Applications **165**, 113941 (Mar 2021). https://doi.org/10.1016/j.eswa.2020.113941, arXiv:2005.01992 [cs]
10. Pan, Y., Froese, F., Liu, N., Hu, Y., Ye, M.: The adoption of artificial intelligence in employee recruitment: The influence of contextual factors. The International Journal of Human Resource Management **33**(6), 1125–1147 (Mar 2022). https://doi.org/10.1080/09585192.2021.1879206
11. Papenmeier, A., Englebienne, G., Seifert, C.: How model accuracy and explanation fidelity influence user trust (Jul 2019), http://arxiv.org/abs/1907.12652, arXiv:1907.12652 [cs]
12. Ribeiro, M.T., Singh, S., Guestrin, C.: ”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. pp. 1135–1144. ACM, San Francisco California USA (Aug 2016). https://doi.org/10.1145/2939672.2939778
13. Sanneman, L., Shah, J.A.: The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems. International Journal of Human–Computer Interaction **38**(18-20), 1772–1788 (Dec 2022). https://doi.org/10.1080/10447318.2022.2081282
14. Schmidt, P., Biessmann, F.: Quantifying Interpretability and Trust in Machine Learning Systems (Jan 2019), http://arxiv.org/abs/1901.08558, arXiv:1901.08558 [cs, stat]
15. Shin, D.: Role of fairness, accountability, and transparency in algorithmic affordance. Computers in Human Behavior (2019)
16. Shin, D.: How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. Computers in Human Behavior (2020)
17. Shin, D.: The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies **146**, 102551 (Feb 2021). https://doi.org/10.1016/j.ijhcs.2020.102551
18. Tintarev, N.: Explaining recommendations. Ph.D. thesis, University of Aberdeen, UK (2009)
19. Tocchetti, A., Corti, L., Balayn, A., Yurrita, M., Lippmann, P., Brambilla, M., Yang, J.: A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities (Oct 2022), http://arxiv.org/abs/2210.08906, arXiv:2210.08906 [cs]
20. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics **10**(5), 593 (Jan 2021). https://doi.org/10.3390/electronics10050593
21. Zhou, J., Verma, S., Mittal, M., Chen, F.: Understanding Relations Between Perception of Fairness and Trust in Algorithmic Decision Making (Sep 2021), http://arxiv.org/abs/2109.14345, arXiv:2109.14345 [cs]