HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230126

Towards Responsible AI: Developing Explanations to Increase Human-AI Collaboration

Regina DE BRITO DUARTE [a,1]
[a] *Instituto Superior Técnico, Lisbon*

Abstract. Most current XAI models are primarily designed to verify input-output relationships of AI models, without considering context. This objective may not always align with the goals of Human-AI collaboration, which aim to enhance team performance and establish appropriate levels of trust. Developing XAI models that can promote justified trust is therefore still a challenge in the AI field, but it is a crucial step towards responsible AI. The focus of this research is to develop an XAI model optimized for human-AI collaboration, with a specific goal of generating explanations that improve understanding of the AI system's limitations and increase warranted trust in it. To achieve this goal, a user experiment was conducted to analyze the effects of including explanations in the decision-making process on AI trust.

Keywords. Human-AI Interaction, Human-AI Collaboration, AI Trust, Explainable AI

1. Introduction

Artificial Intelligence (AI) systems have become increasingly important in recent years, performing tasks and making decisions that were previously only possible for humans. With advancements in AI research and development, highly accurate systems have been created that can even outperform human capabilities. However, it is not yet time to blindly trust autonomous intelligent systems and delegate all high-stakes decisions to them. Currently, human-AI collaboration is crucial for achieving the best possible outcomes in most tasks. Examples of this collaboration can be seen in various areas, such as driving a car with a driverless system, in healthcare, where doctors use automated diagnostic systems to assist with patient diagnosis, or even in the financial industry, where credit decisions are made with the aid of a scoring system.

Research on Human-AI collaboration is already underway, with a focus on developing mental models of AI systems that allow humans to understand the system's error boundaries. The goal is to achieve the best possible team performance [1]. This approach enables users to have calibrated trust in the AI system, knowing when to trust and when not to trust it [2]. However, humans are often unaware of the error boundary of the model, which can result in uncalibrated trust. This can lead to a mismatch between the trust that

[1] Corresponding Author: Regina de Brito Duarte, reginaduarte@tecnico.ulisboa.pt

humans have in the AI (AI trust) and the trustworthiness of the AI itself (the intrinsic characteristics of the model that made it trustworthy). This mismatch can cause either over-reliance on the AI system or under-reliance on it [3].

One of the most commonly used techniques for calibrating trust and increasing knowledge of the error boundary of AI systems is Explainable Artificial Intelligence (XAI) models. These models are designed to generate explanations about how a system makes its predictions or recommendations so that users can gain a better understanding of its inner workings. However, empirical studies have not consistently found evidence of the effectiveness of explanations in calibrating trust. In fact, some studies suggest that explanations can increase unwarranted trust, leading to an over-reliance on the system even when it is incorrect [4,5].

This is because XAI models are primarily optimized to verify AI input-output relations, regardless of the context. This objective does not always align with the goals of Human-AI collaboration, which are to enhance team performance and calibrate trust appropriately. Designing and developing XAI models that improve appropriate trust is still a challenge in the field of AI. However, it is a crucial step towards achieving responsible AI, where humans are aware of the strengths and limitations of AI systems and can make better decisions in a variety of tasks.

2. Related Work

2.1. XAI Models Background

AI systems are becoming largely used, but they are also predominantly black-box models in a way that are too complex for a human to understand its decisions. XAI research tries to solve that problem, coming up with ways to explain the predictions made by the AI. The most simple way to explain a model is to retrieve meaning from the model itself. When that is present, I have **interpretable ML models**. These are the simplest ML models, easy enough to understand their behavior. Often, these models are linked with Logistic and linear regressions, decision trees, k-nearest neighbor, and rule-based learners. In contrast, **post-hoc explanations** are generated after the training of the ML model, and they constitute the output of a second system whose sole purpose is the creation of the explanation. **Post-hoc explanations** encapsulate numerous techniques that differ in the way they work; their goals and outputs. This is why some efforts were made to classify and group all those techniques [6,7]. From all the categorizations, three major classes of XAI post-explanations emerge: Feature importance explanations, Example base explanations and Example by simplification explanations.

Feature importance Explanations This group of techniques identifies the contributions of each feature for the predictions of the AI system. This can be done either locally, understanding the importance of each feature for only one observation, or globally, understanding the importance of each feature for the model as a whole. Popular techniques as the Shapley values, the Lime technique and saliency maps belong to this category [8,9,10].

Explanations by example This group of techniques presents its output as a set of examples similar to the observation with the same model prediction or counterfactual examples with a different model prediction. The idea is that the user can understand the reasoning behind the model by comparing examples with original observations, and then make conclusions as a result [11,12,13]. Counterfactual examples, in this context, are hypothetical examples that show how to obtain a different model prediction given a certain observation. It shows versions of the same observation with slightly different characteristics that would have had a different outcome.

Explanations by simplification This category simplifies the model reasoning by coming up with simple general rules to understand its behavior. Rule-based learners as decision trees and genetic programming rule-based extractions can be used above the ML model.

In the AI field, feature importance explanations are the most used. This applies also when using explanation to increase trust. Yet, example-based explanations that show counterfactual examples are thought as a way to simulate the way humans think as they are contrastive and recreate a counterfactual world [14].

2.2. *Trust on AI systems*

A key motivation of XAI is to increase the user's trust in trustworthy AI. Although there is prior work that relates positively the use of explainability with the increasing of trust [15,16], there are also evidences that contradicts that assumption. A controlled experiment suggested that when users provide feedback to an AI system to improve its performance, user trust and user model accuracy perception decreases [17]. In a Human-AI collaborative setting study, it has been shown that the perception of the AI system can be dependent on many variants, such as the direction of communication and the type of model that is behind the AI system [18]. Furthermore, besides the typical human algorithmic aversion, there is also evidence that humans prefer human decision-making discretion to algorithms that blindly apply human-created fairness principles to specific cases. The reason is that humans have the free will to transcend fairness principles if needed [19].

These studies showcase the sensibility of the user's trust in AI systems, and they question the real benefits of XAI. Another relevant study assessing the importance of XAI on users' trust is an experiment that studied data scientists' use of interpretability tools in their daily work [4]. The study found that the interpretability tools are sometimes being misused and even ensuring unwarranted trust in data scientists practitioners. As data scientists and machine learning practitioners are a special case of users with knowledge of AI and ML, the study is even more pertinent.

2.3. *Human-AI collaboration*

There is a line of research that evaluates the collaboration between AI systems and humans as being part of the same team. Bansal et al. emphasize the importance of the user's mental models of AI Systems in a collaborative setting [1,20]. They highlight that the error boundary of an AI system that is parsimonious and in line with the user's mental model can be more effective and increase team performance (AI systems and humans) than model accuracy. The reason is that humans and AI systems can work in a complementary manner so that the human knows the error boundaries of the model and knows when it predicts wrongly. Additionally, they introduce the notion of *compatibility* to describe updates to the model performance. The idea is that these updates should be compatible with prior versions of the model so that the new version can be coherent with the mental model of the user. Hence, the team performance does not deteriorate. Similarly, Wang et al. examined whether and when human decision-makers adopt the AI model's recommendation in a Human-AI collaborative setting [21]. Their results highlight that in AI-assisted decision-making, human decision-makers' utility evaluation and action selection are influenced by their judgment and confidence in the decision-making task. That is, humans are prone to make use of their own judgment in a decision-making trial to gauge whether to adopt the AI recommendation. Furthermore, when the stakes of the decisions become larger, people tend to lower their belief in AI recommendation's correctness and rely more on their own judgment in AI-assisted decision-making.

Similar studies on Human-AI collaborative settings investigated how people trust an AI assistant with a different level of expertise [22]. The results demonstrated that participants were able to perceive when the assistant was an expert or non-expert within the same task and calibrate their reliance on AI to improve team performance. Additionally, communicating expertise through the linguistic properties of the explanation text was effective, where embracing language increased reliance and distancing language reduced reliance on AI.

3. Research Questions and Challenges

This research is concentrated in the **development of an XAI model that is optimized for Human AI collaboration**. This means developing explanations that serve human-AI teams by contributing to the understanding of the model's limitations and enhance *warranted* trust in the system. Only that way, human-AI collaboration can be optimized.

To develop these type of explanations, several preliminary research questions have to be investigated in order to guarantee an optimal Human-AI collaboration:

1. How current XAI models affect AI trust?
2. How to measure AI trust correctly and appropriately given a certein context?

In the first research question, I want to understand how the XAI models affect AI trust when a user has to do a decision-making task with the assistance of an AI system. The literature on this topic is not extensive. Moreover, the positive effect of explanations on AI trust is still in question. The gap in the understanding of this relationship is even bigger when one considers different contexts of decision-making tasks (e.g. high stakes vs low stakes and high vs low model performance).

The second research question is related with the right measurement of AI trust. Are we measuring trust on AI systems effectively? Which is the best way to measure AI trust? I want to investigate new metrics of AI trust that can measure more accurately the trust of the humans on the AI system. Currently, AI trust is being measured by subjective questionnaires that more often than not measure perceived trust rather than demonstrated trust [23]. Moreover, even the trust measures that assess the behavior of the users can be also misleading, as the settings to ensure trust may not be there [2]. Based on these evidences, a more systematic way for measuring AI trust given the appropriate context is missing.

To elaborate on this research, I have conducted a randomized experiment to investigate the relationship between explanations and trust in AI when humans are aided by AI systems in a decision-making task. The methodology and findings of this experiment are presented in the following sections.

4. First experiment: Understanding the relation between XAI and AI trust

This experiment was designed to investigate whether the presence of explanations of AI predictions during the decision-making task could increase trust in the AI system. As external variables may influence this relationship, I aimed to explore if the stakes of the decision task and the performance of the AI model played a significant role in the relation between explanations and AI trust. With these objectives in mind, I tested three main hypotheses:

H1: *Trust on AI systems is higher when explanations of the system are present (specifically feature importance explanations and counterfactual explanations) comparing to the absence of explanations.*

H2: *The presence of explanations enhance trust on AI systems, regardless of AI performance.*

H3: *AI trust is lower when the level of risk is high.*

4.1. Task description

Participants were asked to complete a decision-making task where they had to determine if a mushroom was edible or poisonous. The task included various characteristics of the mushroom such as cap color, cap shape, odor, etc., and a visual representation of the mushroom. Additionally, an AI model predicting the edibility of the mushroom was presented to the participant. Using this information, the participants had to decide if the mushroom was edible or poisonous by clicking on the appropriate button on the screen. The task was presented as a game where the participants had to repeat the process several times, with the objective of consuming edible mushrooms and avoiding poisonous ones.

4.2. Study Design

I designed the mushroom task in a way that I could manipulate three factors: the type of explanations (XAI model) presented; the risk level; and the AI model performance.

Type of XAI model. To test the first hypothesis related to the presence of explanations in the decision-making, I created 3 conditions in the task: a **control condition**, in which no explanation of the AI recommendation was provided, and two other conditions with different types of explanations presented to the user at the time of the decision. In one condition, a local feature importance explanation was presented to the user. The explanations were produced by **LIME**, a well-known technique to obtain local explanations [9]. In the last condition, an example-based explanation was presented to the user. In this case, **DICE** technique was used to produce state-of-art counterfactual examples [11]. More information about these techniques can be seen in the Supporting Material.

The risk level. I manipulated the stakes of the mushroom game on two factors: low or high. In the low-risk condition, participants received information that they got sick after eating 3 poisonous mushrooms (the risk associated to the decision is low, indicating they will still be alive at the end). In the high risk condition, users were informed that they would get sick after eating one poisonous mushroom, and they would die after eating the second poisonous mushroom.

Model performance. Finally, I manipulated the performance of the AI recommender system as a proxy for the trustworthiness of the model. I developed 2 levels: good and bad performance. I defined an AI system with poor performance, with accuracy rate of 60% and another with a good performance, with accuracy of 96%. The development of these two models is detailed in the Supporting Material.

The experiment had a mixed 3 (XAI model) X 2 (Risk) X 2 (Performance) design. The Type of XAI models and the risk level were between-subjects, whereas the performance of the AI system was a counterbalanced within-subject variable. In other words, each participant was assigned to only one type of XAI model condition, either in a low or in a high risk condition. Moreover, each participant was assigned to both good or bad AI performance by playing the game twice, with the order of good/bad performance of the system being counterbalanced.

4.3. Procedure

Participants began the experiment after providing their consent to participate voluntarily. They were then asked general demographic questions regarding their age, gender, nationality, and education level. Next, the instructions and rules for the mushroom game were presented, along with a simple tutorial to help participants become familiar with the game. Following the tutorial, they began the first game and were required to consume at least 5 edible mushrooms to win. Each game consisted of 12 tasks with distinct mushrooms, and after each task, they received feedback on their decision indicating whether the mushroom was edible or not. In the final round of the game, participants were asked if they would delegate the next action to the AI recommender system or not. Once the mushroom game was completed, a final score was presented, and participants were asked to respond to a questionnaire to assess their perceptions of the game and the AI recommender system. The questionnaire included several items to measure AI trust. Additionally, two attention check questions were included to ensure reliable results. Finally, participants played a second time with different AI performance and filled out a questionnaire regarding the second game. Figure 1 provides a visual representation of the experiment to aid in understanding.

4.4. Measures

The goal was to quantify trust in AI recommendations. Despite the complexities of studying and measuring trust in AI systems [2], several variables were used as proxies to measure trust. Four self-report measures were assessed: **MDMT Trust**, **Single-item Trust**, **AI Understandability**, and **XAI Quality**. The first two measures assessed subjective trust in the system through a questionnaire based on the Multi Dimensional Measure of Trust (MDMT) scale [24] and a single question asking for participants' perception of trust, respectively. The other two measures assessed system understandability and the

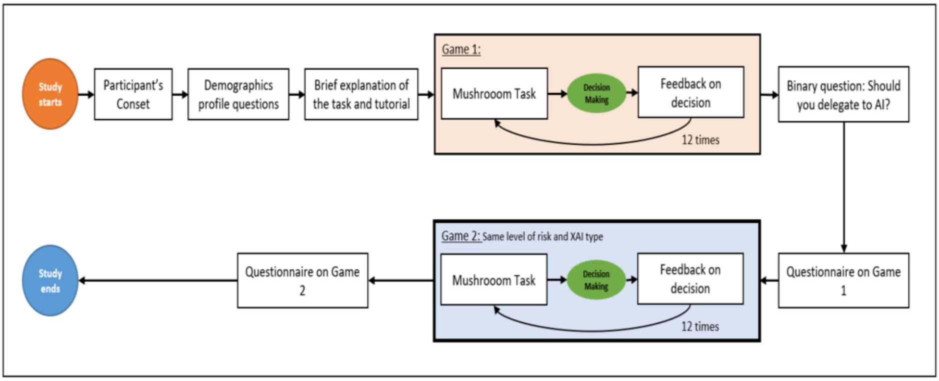

Figure 1. Experiment Workflow of the first study to understand the relation between XAI explanations and AI trust

perceived quality of the XAI model's explanation. These self-report measures were derived from the questionnaires after each game.

Two variables related to participant behavior were also assessed. The first, **Behavioral Trust**, measured the percentage of times participants agreed with the Recommender system and made decisions accordingly. The second, **Delegation**, was a binary variable indicating whether participants would delegate the decision to the AI.

4.5. Participants

I recruited a total of 215 participants from Prolific, a crowdsourcing platform that facilitates large-scale data collection from participants. To ensure data quality, I only included participants with a fluent level of English. After excluding responses from inattentive participants (based on the attention check questions), I was left with data from 211 participants. Out of this population, 114 identified as female, 94 as male, and 3 as non-binary. The average reported age was 27 years, ranging from 19 to 61 years. The participants were compensated with a payment of 2.50 GBP for their participation in the study. On average, it took participants approximately 12.12 minutes to complete the task, resulting in a median hourly rate of 11.41 GBP.

5. Preliminary Results - First Study

To test the hypotheses, I conducted mixed Analyses of Variance (ANOVAs) using the SPSS Statistics 26 with a 3 (XAI model) X 2 (Risk level) X 2 (Model Performance) design on the five dependent variables and a Chi squared test for the dependent variable *Delegation*. The alpha level was set at 0.05, and a Bonferroni alpha correction was used for multiple testing adjustments. For the sake of brevity, I will focus on statistically significant results, and selectively report non-statistically significant results to address my specific hypotheses.

5.1. Self-Reported Measures

The effects on the self-reported metrics *MDMT Trust, AI Understandability, XAI Quality* and the *Single-Item Trust* were similar, therefore, I report them grouped in one section.

Dependent Variable	DICE	LIME	Control	Good Performance	Bad performance
MDMT Trust	4.60 ± 0.13*	4.98 ± 0.14*†	4.50 ± 0.13†	5.34 ± 0.85	4.05 ± 0.97
Single-item Trust	4.02 ± 0.17	4.45 ± 0.18	3.97 ± 0.17	4.84 ± 0.11	3.46 ± 0.12
AI Understandability	3.99 ± 0.20*	4.70 ± 0.20*†	4.00 ± 0.20†	4.53 ± 0.12	3.92 ± 0.13
XAI Quality	4.40 ± 0.16*	5.01 ± 0.17*†	4.04 ± 0.17†	4.85 ± 0.10	4.11 ± 0.11
Behavioral Trust	0.83 ± 0.02	0.84 ± 0.02	0.84 ± 0.02	0.87 ± 0.01	0.8 ± 0.01

Table 1. Means and standard errors for the main effects of the XAI Model Type (in the left columns) and Model Performance (in the Right columns) on the dependent variables.

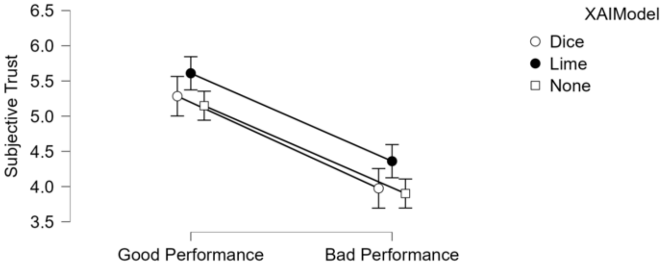

Figure 2. Mean of MDMT Trust Scale by XAI Type and AI system performance.

There was a significant main effect of Model Performance on *MDMT Trust*, $F(1,205) = 172,83, p < .001, n_p^2 = .459,$ *AI Understandability*, $F(1,205) = 45,342, p < .001 n_p^2 = .181$, *XAI Quality*, $F(1,205) = 92,03, p < .001, n_p^2 = .310$ and *Single Item Trust*, $F(1,205) = 125,646, p < .001, n_p^2 = .380$. The model performance was the factor that had the most impact in all these four dependent variables. As predicted, trust levels on the AI system were significantly higher when the AI system performance was good compared with a bad AI system performance. The understandability of the AI system and the perceived XAI quality were also higher when the performance of the AI model was good. The means and standard errors are displayed in the Table 1.

Additionally, the XAI model type factor showed significant main effects on the MDMT Trust ($F(2,205) = 3.50, p < .032, n_2^p = .033$), AI Understandability ($F(2,205) = 4.096, p < .018, n_2^p = .038$), and XAI Quality ($F(2,205) = 8,613, p < .001, n_2^p = .078$) measures. Pairwise comparisons revealed that participants in the LIME XAI condition (explanations with feature importance) reported higher subjective trust, better understanding of the AI system, and perceived the explanations as having higher quality than those in the other two conditions (DICE with counterfactual explanations and the control group). However, there were no significant differences in these three variables between the DICE and control conditions. Table 1 presents the means and standard deviations, with the pairs of groups that were statistically significant at $p < .05$ indicated with * or † symbols.

The 3X2X2 ANOVAs did not reveal any significant interactions between the XAI model and the other independent variables for each of the four dependent variables. This pattern remained consistent when considering the different levels of risk and AI system performance. Figure 2 shows that the patterns for the XAI Model independent variable were similar regardless of the model performance on the *MDMT Trust* measure.

The results for all four self-reported measures were consistent regarding the effects of risk, showing significant interactions between risk and performance. Participants expressed less trust, understanding, and provided lower evaluations of the quality of AI explanations in the high-risk condition compared to the low-risk condition, but only when the performance of the AI system was poor. When the AI system performed well, there were no significant differences in any of the dependent variables.

5.2. Behavioral Trust and Delegation

Regarding the *Behavioral Trust* measure, which assesses the level of compliance with AI recommendations, a significant main effect of model performance was found ($F(1,205) = 37.71, p < .001, \eta_p^2 = 0.155$). Specifically, when the AI had good performance, the rate of concordance with the AI recommendation was higher.

In contrast to the self-reported measures, the effect of XAI model type on *Behavioral Trust* showed a different pattern. Although the main effect of XAI model was not statistically significant, there was a significant interaction between XAI model and AI system performance ($F(2,205) = 25.40, p < 0.001, \eta_p^2 = 0.20$). As shown in Figure 3 and similar to the results on *MDMT Trust, AI Understandability and XAI Quality*, participants in the LIME condition (0.93 ± 0.02) displayed more trust by following the system recommendation compared to participants in the other conditions (0.84 ± 0.02 in DICE and Control). However, this effect was only significant when the AI system performance was good. When the performance was bad, the *Behavioral Trust* of participants in the LIME condition (0.75 ± 0.02) was significantly lower than that of the other two conditions (0.81 ± 0.02 in DICE and 0.84 ± 0.02 in the Control condition). Once again, the behavioral trust in the DICE and Control conditions was not statistically different.

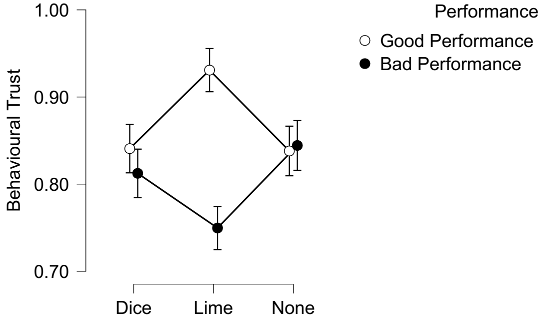

Figure 3. Means of the Behavioral trust by XAI model and AI system performance; error bars represent 95% confidence intervals.

In contrast to the self-reported measures, the results for the *Behavioral Trust* did not indicate significant effects of risk. Participants' behavior in following the recommendations did not vary based on risk, even when the AI model performed poorly.

Similar to the self-reported measures, the most significant impact on *Behavioral Trust* was the model performance. Participants were more likely to follow the AI system's recommendations when its performance was good. The means and standard errors are presented in Table 1.

Finally, regarding *Delegation*, I observed an effect of risk only when the model performance was poor. The willingness to delegate decision-making to the AI system was dependent on risk, but only when the AI system performed poorly. Delegation was higher when the risk was low ($X(1) = 9.52, pvalue = 0.002$), and no other significant effects were found for Delegation.

6. Discussion and Research Directions

6.1. The effects of the presence of explanations

My first hypothesis states that user's trust on AI systems is higher when explanations, specifically counterfactual explanations or feature importance explanations, are present in the decision-making process.

The results show that there is significantly higher trust and understandability of the AI system when LIME XAI model explanations are presented. However, the results on the ”Behavioral Trust” measure may contradict this evidence because the effect of the XAI type variable was dependent on the level of AI performance. Participants in the LIME condition followed the AI recommendations more often compared to the other conditions when the model performance was good. Yet, when the AI performed badly, the participants receiving the LIME explanations followed the AI recommendation less.

This pattern might seem odd at first, but this result provides evidence that users understood the explanations presented by the feature importance explanations. As the explanations presented in the LIME condition represent a quantitative contribution of each mushroom characteristic for the prediction, users saw a high or low positive or negative contribution associated with each mushroom characteristic.

Hence, the explanation also showed the uncertainty of the prediction. Accordingly, when users in the LIME condition played the game with the model's bad performance, they were presented with a recommendation that reflected the uncertainty of the model. Therefore, the user could make their own judgment independently of the recommendation. Consequently, the user understood the process and continued to trust the system. These results are also consistent with previous research indicating that presenting confidence intervals of the predictions is useful in calibrating trust in AI systems [25].

Nonetheless, the comparison between the counterfactual explanations produced by DICE and the control condition did not yield significant differences. Hence, the presence of counterfactual explanations alone was not enough to elicit higher levels of trust. This result contradicts the common assumption that counterfactual explanations are easy for humans to understand as they simulate human reasoning [14]. Previous research suggested that the presence of counterfactual explanations led to higher perceived understandability and competence of the AI system in an AI-Human collaboration task involving expert humans [16], contradicting my findings. However, my experiment involved non-expert users, whose lack of knowledge about the problem could have made counterfactual explanations less effective since understanding a counterfactual example without any prior knowledge is difficult.

These results suggest that the first hypothesis is not fully supported, as only the feature importance explanations resulted in higher trust compared to no explanations at all. In a study conducted by Zhang et al. on the effect of explainability and confidence intervals on trust calibration, it was found that when confidence intervals of AI predictions are presented, user trust in the system is more calibrated, but the same is not applicable to the effect of local feature explanations [25]. These findings raise questions about the actual effects of explainability on trust. It is possible that systematic errors may occur when evaluating explanations presented in decision-making tasks, which can prevent the establishment of trust [26]. Factors such as lack of curiosity about the explanation during decision-making, lack of context, confirmatory search, misinterpretation of the explanation, or formation of habits can contribute to this. These factors highlight the challenge of assessing the real effect of explanations and provide insights into why there are several conflicting results. The effect of explanations on trust depends not only on the explanations themselves but also on the participants' actual understanding of the explanations, which is difficult to evaluate. The results of this study suggest that participants understood the feature importance explanation better than the counterfactual ones.

I was also interested in examining the effect of XAI explanations on model trustworthiness when it is uncertain. My hypothesis was that the presence of XAI explanations would enhance trust regardless of the system's performance. Figure 2 shows that the effect of the XAI model type was independent of whether the AI system performed well or poorly. Thus, even for untrustworthy AI systems, users appear to trust them more when presented with feature importance explanations compared to basic information used for decision-making. Despite the fact that trust was significantly lower when the system had poor performance, trust was higher when feature importance explanations were present. Therefore, subjective trust is placed in the system not because it is dependable, but because specific explanations of the system are given. These results have significant implications. A user may agree with an AI system simply because the presence of explanations makes the AI system appear more trustworthy. In high-stakes domains, this implication can affect several decisions and lead to severe consequences of injustice and malpractice.

6.2. AI trust decreases with a high level of risk, but only in untrustworthy scenarios

I hypothesized that risk would have a negative influence on AI trust. The results indicate that participants reported lower trust when the risk was high on all four subjective measures. However, these results were only statistically significant when the performance of the AI system was poor (i.e., when the AI system was deemed untrustworthy). I can argue that risk level does indeed affect AI trust. Yet, model trustworthiness appears to be a more critical factor. When the trustworthiness of the model is guaranteed, the level of risk becomes less significant. Therefore, my initial hypothesis is only partially supported, as the effect of risk is only visible under conditions where the model's performance is poor. In high-stakes domains, ensuring a good AI model performance is essential.

6.3. Research Direction

The results of this experiment provided preliminary insights into the ambiguity of the relationship between AI trust and XAI explanations. For the purpose of Human-AI collaboration, explanations should be clear enough to calibrate trust in the AI system without increasing undesired trust. Currently, this goal is not yet met.

As I move forward with this research topic, the next step is to gain insights into causal theory and develop an XAI model that explains the behavior of the AI system based on its causal relationships. Causal explanations are commonly used when humans explain behavior [14]. Therefore, these types of explanations may be more aligned with Human-AI collaboration objectives than the previous ones. Once such a model is developed, I will conduct a similar user experiment to determine if the assumption that causal explanations are better suited for enhancing AI trust holds true.

Once I have gained an understanding of how current XAI techniques and causal explanations affect AI trust and contribute to the development of mental models of the AI systems, my goal is to explore a new type of XAI techniques optimized directly for AI collaboration objectives. This approach will address AI trust and human-AI collaboration directly in the development of the XAI system, rather than as a post-development benefit. To accomplish this objective, I intend to use reinforcement learning with human feedback to develop a model with explanations tailored to end users and team collaboration.

References

[1] Bansal G, Nushi B, Kamar E, Lasecki WS, Weld DS, Horvitz E. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. 2019 Oct;7(1):2-11. Available from: https://ojs.aaai.org/index.php/HCOMP/article/view/5285.

[2] Jacovi A, Marasović 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; 2021. p. 624-35.

[3] Liao QV, Sundar SS. Designing for Responsible Trust in AI Systems: A Communication Perspective. arXiv preprint arXiv:220413828. 2022.

[4] Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 1–14. Available from: https://doi.org/10.1145/3313831.3376219.

[5] Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, et al. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021. p. 1-16.

[6] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82-115.

[7] Liao QV, Gruen D, Miller S. Questioning the AI: informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020. p. 1-15.

[8] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 4765-74. Available from: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[9] Ribeiro MT, 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, San Francisco, CA, USA, August 13-17, 2016; 2016. p. 1135-44.

[10] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:13126034. 2013.

[11] Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; 2020. p. 607-17.

[12] Guidotti R, Monreale A, Giannotti F, Pedreschi D, Ruggieri S, Turini F. Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems. 2019;34(6):14-23.

[13] Goyal Y, Wu Z, Ernst J, Batra D, Parikh D, Lee S. Counterfactual visual explanations. In: International Conference on Machine Learning. PMLR; 2019. p. 2376-84.

[14] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. 2019;267:1-38. Available from: https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[15] Pu P, Chen L. Trust building with explanation interfaces. In: Proceedings of the 11th international conference on Intelligent user interfaces; 2006. p. 93-100.

[16] Naiseh M, Al-Thani D, Jiang N, Ali R. How the different explanation classes impact trust calibration: The case of clinical decision support systems. International Journal of Human-Computer Studies. 2023;169:102941.

[17] Honeycutt D, Nourani M, Ragan E. Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. 2020 Oct;8(1):63-72. Available from: https://ojs.aaai.org/index.php/HCOMP/article/view/7464.

[18] Ashktorab Z, Dugan C, Johnson J, Pan Q, Zhang W, Kumaravel S, et al. In: Effects of Communication Directionality and AI Agent Differences in Human-AI Interaction. New York, NY, USA: Association for Computing Machinery; 2021. Available from: https://doi.org/10.1145/3411764.3445256.

[19] Jauernig J, Uhl M, Walkowitz G. People Prefer Moral Discretion to Fair Algorithms: Algorithm Aversion Beyond Intransparency. Available at SSRN 3857292. 2021.

[20] Bansal G, Nushi B, Kamar E, Weld DS, Lasecki WS, Horvitz E. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. Proceedings of the AAAI Conference on Artificial Intelligence. 2019 Jul;33(01):2429-37. Available from: https://ojs.aaai.org/index.php/AAAI/article/view/4087.

[21] Wang X, Lu Z, Yin M. Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. 2022.

[22] Zhang Q, Lee ML, Carter S. You Complete Me: Human-AI Teams and Complementary Expertise. In: CHI Conference on Human Factors in Computing Systems; 2022. p. 1-28.

[23] Miller T. Are we measuring trust correctly in explainability, interpretability, and transparency research? arXiv preprint arXiv:220900651. 2022.

[24] Malle BF, Ullman D. Chapter 1 - A multidimensional conception and measure of human-robot trust. In: Nam CS, Lyons JB, editors. Trust in Human-Robot Interaction. Academic Press; 2021. p. 3-25. Available from: https://www.sciencedirect.com/science/article/pii/B9780128194720000010.

[25] Zhang Y, Liao QV, Bellamy RK. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; 2020. p. 295-305.

[26] Naiseh M, Cemiloglu D, Al Thani D, Jiang N, Ali R. Explainable recommendations and calibrated trust: two systematic user errors. Computer. 2021;54(10):28-37.