The Dynamics of Trust in XAI: Assessing Perceived and Demonstrated Trust Across Interaction Modes and Risk Treatments

Mohsen Abbaspour Onari[1,2(✉)], Gregor Baer[1,2], Chao Zhang[3], Isel Grau[1,2], Marco S. Nobile[4], and Yingqian Zhang[1,2]

[1] Information Systems, Eindhoven University of Technology, Eindhoven, The Netherlands
{m.abbaspour.onari,i.d.c.grau.garcia,yqzhang}@tue.nl
[2] Eindhoven Artificial Intelligence Systems Institute, Eindhoven University of Technology, Eindhoven, The Netherlands
[3] Human-Technology Interaction, Eindhoven University of Technology, Eindhoven, The Netherlands
c.zhang.5@tue.nl
[4] Computational Biology, Bioinformatics and Biomedicine, Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy
marco.nobile@unive.it

Abstract. The increasing use of artificial intelligence (AI) models across various fields has raised concerns about whether these models can meet user trust expectations. As a result, researchers are focusing on assessing AI models' performance relative to user expectations to determine trust levels. Evidence suggests that effective interaction with eXplainable AI (XAI) techniques can mitigate over-reliance on AI models and better align user expectations with the actual capabilities of these models in decision-making. In this study, we analyze trust from two perspectives: perceived trust, based on user self-reported trust, and demonstrated trust, which evaluates whether users, when given a choice, prefer to rely on AI or make decisions independently. We also explore how different interactions between human subjects and XAI models, along with varying levels of task risk, influence trust. Our findings reveal that these two types of trust are substantially different; human subjects do not always exhibit trust behavior in actual decision-making tasks, even when they perceive themselves as trusting the AI. Furthermore, we show that an AI model's low error rate in making correct decisions can influence human subjects' mental models, leading them to report a higher tendency to trust the AI. Finally, we conclude that human perceptions of trust are fragile and may change based on ongoing interactions with the model.

Keywords: Trust · User Study · Explainable AI · Decision Making

R. Guidotti et al. (Eds.): xAI 2025, CCIS 2576, pp. 316–334, 2026.
https://doi.org/10.1007/978-3-032-08317-3_15

1 Introduction

The advancement of increasingly complex artificial intelligence (AI) systems has driven the adoption of AI-assisted decision-making. Research in eXplainable AI (XAI) has explored various methods to clarify the complex behavior of machine learning (ML) models. These explainability techniques are essential for justifying the decisions made by black-box models. Enhancing the fairness and trustworthiness of AI systems is frequently cited as a key objective of XAI [10]. However, there are numerous ambiguous aspects of trust that are challenging to formalize using the tools currently available in the AI and human-computer interaction (HCI) literature [12].

The aim of building trust from the user's perspective is to enhance the ability to anticipate behavior in situations involving risk [12]. In this context, XAI serves as a tool that provides users with easier access to the signals that facilitate this anticipation [12]. Prior studies have shown that unintuitive explanations can erode users' trust in an ML model [26]. Therefore, it is essential to design AI-assisted decision-making tools that enable users to interact with the model's explanations, ensuring they can intuitively understand the AI model's logic and ultimately build trust in the system. Another significant challenge is how researchers model and report trust in AI-assisted decision-making. Trust is often conceptualized as a subjective perception of AI, referred to as *perceived trust*, and is typically measured using self-report scales [19]. However, these methods may not accurately capture real trust behavior. The concept of trust reflection within real AI-assisted decision-making tasks, as proposed by Zhang *et al.* [36], is more accurately described as *demonstrated trust* wherein users decide whether to delegate the decision *without* seeing the AI's prediction– thus representing a stricter test of trust. We argue that this approach provides a more reliable measure of trust, as it allows human users to exercise agency in deciding whether to rely on the AI model, rather than merely reporting their perceived trust. Building on this foundation, we aim to take a step further by investigating the role of interaction modes with the XAI model.

In this paper, we aim to investigate the significant disparity between perceived trust and demonstrated trust in AI-assisted decision-making. We will also evaluate how the mode of interaction with the XAI model and the decision-making risk treatment impact human trust. To explore these dynamics, human subjects will be randomly assigned to one of the two XAI modes in the training phase. In the *evaluative AI* mode, participants can manipulate the SHapley Additive exPlanations (SHAP) feature importance plot to observe changes in the AI model's prediction probabilities and feature importance. In contrast, the *non-evaluative AI* mode presents participants with the SHAP feature importance plot and prediction probability for a given sample without any interactive features. At the end of the training phase, participants will report their satisfaction with the efficacy of the explanations and their perceived trust in the XAI modes. During the test phase, participants will be divided into two risk treatment groups: high-risk and low-risk. Each group will be presented with 20 tasks, each involving specific reward and penalty scores. Participants will make

decisions either independently or by delegating them to the AI model, thereby reflecting their demonstrated trust. After completing the decision-making task, participants will again report their perceived trust. Based on this setup, our research seeks to address the following questions:

RQ1: How does the XAI interaction mode affect participants' satisfaction with the effectiveness of the provided explanations during decision-making?
RQ2: Does the evaluative AI mode significantly impact perceived trust during the training phase?
RQ3: How do the XAI mode and risk treatment affect both perceived and demonstrated trust?
RQ4: Does demonstrated trust influence perceived trust after the test phase?
RQ5: Is demonstrated trust significantly different from perceived trust?

The rest of this paper is organized as follows: Sect. 2 reviews the relevant literature. Section 3 introduces our methodology for designing the user study and measuring trust. Section 4 details the conducted user study, while Sect. 5 presents the results of trust measurement. Finally, Sect. 6 concludes the study and highlights directions for future research.

2 Related Work

The increasing use of AI-powered decision aids has sparked a series of experimental studies within HCI communities. These studies aim to understand how humans interact with, rely on, and trust AI models in the context of AI-assisted decision-making [15,18,19]. With the emergence of XAI, researchers have focused on integrating explanations into AI-assisted decision-making. This includes examining the impact of explanations on complementary team performance [3], error detection with explanations [9], application-oriented contexts for fraud detection [2], improving objective performance and subjective usability of model [13], and the impact of model and explanation errors on human decision-making [17,22]. As LLMs gain popularity, explanations play a crucial role in guiding Human-AI collaboration [23], generating counterfactual examples for fairer learning models [21], and facilitating Human-LLM collaborative annotation [33].

Ribeiro *et al.* [28] demonstrated that explanations are valuable across various models for trust-related AI-assisted decision-making tasks. The widespread use of XAI methods has inspired numerous empirical studies examining how humans trust AI models in AI-assisted decision-making. Consequently, researchers have explored various factors influencing trust in XAI models, including example-based explanations for ML classifiers [35], confidence scores and local explanations [36], and the impact of different types of AI assistance [16]. Studies have also examined AI descriptions as algorithmic recommendations [7], dissenting explanations [27], and XAI for skill development in community health workers [24].

Other research has investigated the effects of feature-based explanations on distributive fairness [30], interpretability and outcome feedback [1], and sociotechnical mismatches in AI explainability [6]. Additionally, studies have analyzed the impact of explanations in cases of AI errors [25], unintuitive feature explanations [26], and different treatments such as explanations, model bias disclosure, and proxy correlation disclosure [10]. Unlike previous research, this study examines users' trust behavior across different scenarios–both when they do not interact with an XAI assistant and when they do under two distinct decision-making risk treatments. By assessing both perceived and demonstrated trust, we aim to provide deeper insights into trust behavior in AI-assisted decision-making.

3 Methods

In this section, after reviewing the definitions of trust, we introduce our AI-assisted decision-making approach for designing our experiment to evaluate trust.

3.1 Perceived Trust vs. Demonstrated Trust

Ueno *et al.* [31] defines trust as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party." Jacovi *et al.* [12] extends the Human-AI trust definition as "if H (human) perceives that M (AI model) is trustworthy to contract C, and accepts vulnerability to M's actions, then H trusts M contractually to C. The objective of H in trusting M is to anticipate that M will maintain C in the presence of uncertainty, and consequently, trust does not exist if H does not perceive risk." However, most studies assess perceived trust in AI using self-report scales [19]. Prior research suggests that perceived trust may not reliably reflect actual trusting behaviors [14,29]. Therefore, some scholars have proposed alternative indicators to study trust, such as switch percentage and agreement percentage [36], as well as relying on Cohen's d score [34]. In addition to examining users' perceived trust, this study simulates risky situations to investigate whether participants choose to delegate their decisions to an AI model when they are also capable of making decisions independently (demonstrated trust). We quantify this type of trust as the ratio of the number of decisions delegated to the AI to the total number of decision-making tasks. We believe this approach better aligns with established definitions of Human-AI trust.

3.2 Evaluative AI

In cognitive forcing, the decision-maker is actively engaged in evaluating different options and making trade-offs. Providing explanatory information from the start can help the decision-maker focus on relevant details and make more informed decisions [20]. Buçinca *et al.* [4] and Gajos and Mamykina [8] demonstrated that

cognitive forcing significantly decreases over-reliance compared to basic XAI methods. Miller [20] introduced the concept of evaluative AI, a framework for explainable decision support that resembles cognitive forcing. This framework assists decision-makers in accessing the information they need to evaluate a hypothesis as and when required. Unlike traditional approaches, evaluative AI does not automatically provide recommendations. Instead, it helps users filter out unlikely options, generate new hypotheses, or both. The decision-maker then examines a hypothesis and requests the decision aid to present evidence both supporting and challenging it.

In this study, we utilize the evaluative AI framework for AI-assisted decision-making. For each sample, we present the corresponding SHAP feature importance plot alongside the prediction probability instead of direct recommendations for classifying the instance in a binary classification task. In our evaluative AI mode, human subjects can modify feature values and observe how these changes impact both the SHAP feature importance values and the prediction probabilities. This approach helps human subjects filter out less critical features and focus on those with the greatest influence on the prediction. Additionally, observing how the prediction probability shifts with different feature values can provide human subjects with intuitive insights into the likely classification of the instance. After evaluating various hypotheses, SHAP feature importances, and prediction probabilities, human subjects can make an informed final decision.

4 Human-Subject Experiment

In this section, we present our proposed human-subject experiment designed to measure perceived and demonstrated trust across two different risk levels and interaction modes with XAI models, providing insights into the underlying trust mechanisms.

4.1 Decision Making Task

In our experiment, we ask our human subjects to classify mushroom instances into "Edible" or "Poisonous" classes. The dataset [32] comprises 17 nominal variables and three quantitative variables. It is balanced with respect to the class distribution, with an overall ratio of $e : 0.45$ and $p : 0.55$. During the data preprocessing phase, we first removed variables with more than 50% missing values. Next, we eliminated variables exhibiting high multicollinearity based on the correlation coefficient between features. This process resulted in a final dataset containing 12 variables: nine nominal and three quantitative. We randomly split the dataset into training and test sets using an 80%/20% partition. We tested several ML algorithms, tuning their hyperparameters with GridSearch and assessing performance via 5-fold cross-validation. Among these, AdaBoost emerged as the most suitable model for our study, achieving an accuracy of 0.7989 on the test dataset. Although not perfect, this accuracy is sufficient for

our user study. To effectively simulate a risky situation and assess human subjects' trust in the AI model, we need a model that is not perfect and exhibits some prediction errors. This approach will reveal how human subjects behave in various risky scenarios when they are aware that the model might make mistakes. Next, we implemented SHAP to enhance model explainability by highlighting the importance of each local feature for each instance. In the decision-making task, the primary objective is to classify each mushroom instance as either "Poisonous" or "Edible," with the added risk of penalties for incorrect decisions. If participants are uncertain about their decision, they have the option to delegate it to the AI model. Our goal is to investigate whether, in risky situations where participants are capable of making decisions independently, their choice to delegate decision-making to the AI model reflects trust in its decision.

4.2 Experimental Treatments

To evaluate how interactions with the XAI model affect both types of trust, we implemented two treatment modes: non-evaluative AI mode and evaluative AI mode in the training phase. In the test phase, human subjects are divided into two groups: high-risk and low-risk treatments. Although the reward for correct decisions is 15 points for both treatments, the high-risk group faces a penalty of 25 points for incorrect decisions, while the low-risk group incurs a 10-point penalty. Consequently, our study includes four treatment conditions to explore how interaction with the XAI model and the level of risk influence participants' trust behavior. The holistic AI-assisted decision-making task in this study is illustrated in Fig. 1.

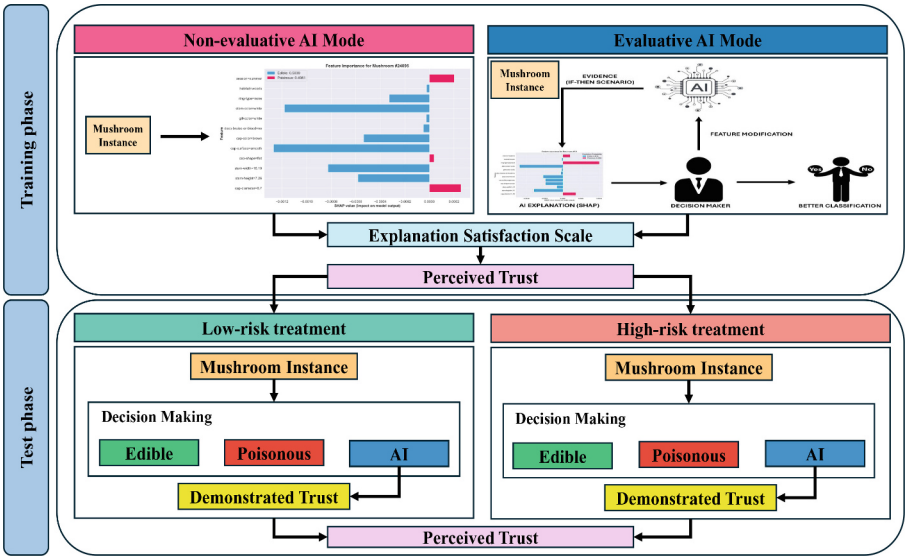

Fig. 1. AI-Assisted Decision-Making Experimental Procedure of This Study

4.3 Experimental Procedure

We conducted our experiment on Prolific, recruiting highly active participants from the US and UK. To assess the impact of both XAI mode and risk treatment on trust, resulting in four experimental groups, we performed a power analysis using a two-way between-subjects ANOVA. We selected an effect size of 0.2, an alpha level of 0.05, and a desired statistical power of 0.9, with a numerator degrees of freedom of 3. This analysis indicated that a sample size of 360 participants would be required to achieve statistically meaningful results. Among the 360 human subjects, 181 were randomly assigned to the evaluative AI model, with 85 participants in the high-risk group and the remaining 96 in the low-risk group. The non-evaluative AI group included 179 participants, with 92 in the high-risk group and 87 in the low-risk group. At the start of the experiment, they were presented with an ethical consent form, and the experiment was terminated if they did not agree to the terms[1]. Human subjects then provided demographic information, including age, gender, and their level of knowledge about mushroom detection and AI.

Following this, human subjects received a brief tutorial on how mushrooms are typically distinguished between poisonous and edible in the real world. They also received a tutorial on XAI and how to interpret SHAP feature importance plots. The training phase started with the presentation of five mushroom instances. After each decision, participants received immediate feedback, including the AI model's prediction. After completing the training task, human subjects rated their satisfaction with the explanations provided during the decision-making process using Explanation Satisfaction Scale (ESS) [11] (See Table 1) on a 5-point Likert scale from disagree strongly to agree strongly. They also assessed their perceived trust based on the proposed trust continuum by [5], which aims to quantify trust (See Table 2). The quantified value can help us to normalize perceived trust to compare with demonstrated trust.

Table 1. Explanation Satisfaction Scale and description

ESS	Description
Understandability	The explanation was understandable.
Sufficiency of details	The explanation had sufficient details.
Completeness	The explanation was complete enough.
Feeling of satisfaction	I am satisfied with the quality of the explanation.
Accuracy	The explanation was accurate enough.
Usability	The explanation was easy to use.
Functionality	In general, the explanation helped me in the decision-making task.

[1] This study has been approved by Ethical Board of the university with reference number: ERB2023IEIS10.

Table 2. Perceived Trust Scale

Linguistic terms	Description	Quantified value
Distrust	I distrust the model.	−1
Undistrust	I have a tendency to distrust the model.	−0.5
Ignorance	I feel ignorant about the model.	0
Untrust	I have a tendency to trust the model.	0.5
Trust	I trust the model.	1

The test phase commenced after human subjects completed the training phase. Based on their risk treatment group, they were informed whether they would face a penalty of 10 or 25 points, with their credit starting at 0. Participants were also informed that their performance in the real experiment would affect their base payment by a maximum of 2 GBP. They then evaluated 20 new mushroom instances (not seen during training) without any additional information or model predictions. The distribution of instances was consistent with the AI model's accuracy, including 16 correctly classified and 4 misclassified instances. During the test phase, human subjects had three decision options: classify as "Edible," classify as "Poisonous," or delegate the decision to the AI model. After making a decision, they were immediately informed if their answer was correct, and their credit was updated accordingly.

Upon completing the real experiment, participants were informed that they would receive a base payment of 2.5 GBP regardless of their performance. To simulate a risky situation, they were told their final reward or penalty would be adjusted by up to 2 GBP based on their performance. An additional 1 GBP bonus was awarded to the top 13 participants with the highest credit balance. Finally, participants were asked another question about their perceived trust level to assess any shifts in trust before and after the test phase.

5 Evaluations

This section has been divided into different subsections, each addressing one or more research questions defined in this study.

5.1 Impact of Evaluative AI on ESS

Addressing RQ1, we evaluated the impact of XAI mode on human subjects' satisfaction with the efficacy of the provided explanations. Given that ESS (the dependent variable) has a meaningful order, we employed the MANOVA test, which is designed to assess whether there are statistically significant differences in multiple dependent variables across different groups. MANOVA is particularly useful when dealing with multiple correlated dependent variables, allowing us to analyze the effect of the independent variable (XAI mode) while accounting for these relationships.

The MANOVA test revealed no significant effect of XAI mode on ESS. The results are summarized in Table 3 and illustrated in Fig. 2. The key observation is that, across both XAI modes, ESS is generally high, with participants somewhat agreeing that the explanations are helpful. However, it is noteworthy that for the evaluative AI mode, there is a higher concentration of responses in the "I agree somewhat" scale. In contrast, for the non-evaluative AI mode, responses are more evenly distributed across different satisfaction scales. Thus, our assumption that evaluative AI would lead to higher satisfaction with explanations is rejected.

Table 3. Results of the MANOVA Test for ESS

Intercept	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.0570	7.0	352.0	831.5447	0.0
Pillai's trace	0.9430	7.0	352.0	831.5447	0.0
Hotelling-Lawley trace	16.5364	7.0	352.0	831.5447	0.0
Roy's greatest root	16.5364	7.0	352.0	831.5447	0.0
C(XAI Mode)	**Value**	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Wilks' lambda	0.9803	7.0	352.0	1.0126	0.4219
Pillai's trace	0.0197	7.0	352.0	1.0126	0.4219
Hotelling-Lawley trace	0.0201	7.0	352.0	1.0126	0.4219
Roy's greatest root	0.0201	7.0	352.0	1.0126	0.4219

5.2 Impact of Evaluative AI on Perceived Trust After Training Phase

To address RQ2, at the end of the training phase, human subjects rated their perceived trust in relation to the XAI modes, as shown in Table 2. A one-way ANOVA (see Table 4) reveals no significant differences in perceived trust levels between the XAI modes at this stage. Both groups demonstrated a high tendency to trust the guidance provided by the XAI models (Although it is still not perfect trust). However, as indicated in Fig. 3, the proportion of "Ignorance" responses is lower for the non-evaluative AI mode. Additionally, Cohen's d score

Table 4. One-way ANOVA Results with Cohen's d for Perceived Trust After Training Phase

	df	sum_sq	mean_sq	F	PR(>F)	Cohen's d
XAI Mode	1.0	1.1148	1.1148	1.1850	0.2770	-0.1148
Residual	358.0	336.7851	0.9407			

suggests a slight increase in perceived trust for the non-evaluative AI mode, although this difference is not statistically significant. Therefore, our assumption that evaluative AI will lead to higher perceived trust after the training phase is rejected.

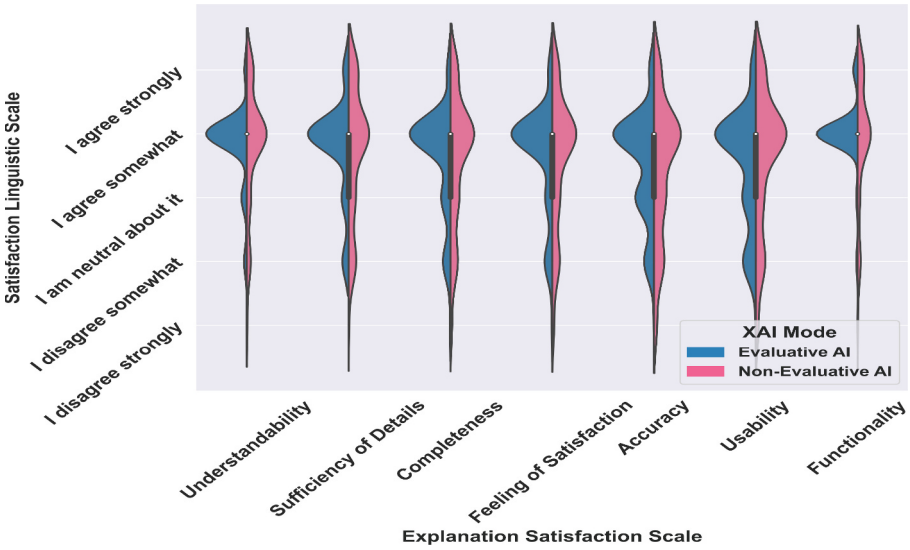

Fig. 2. Violin Plot of ESS

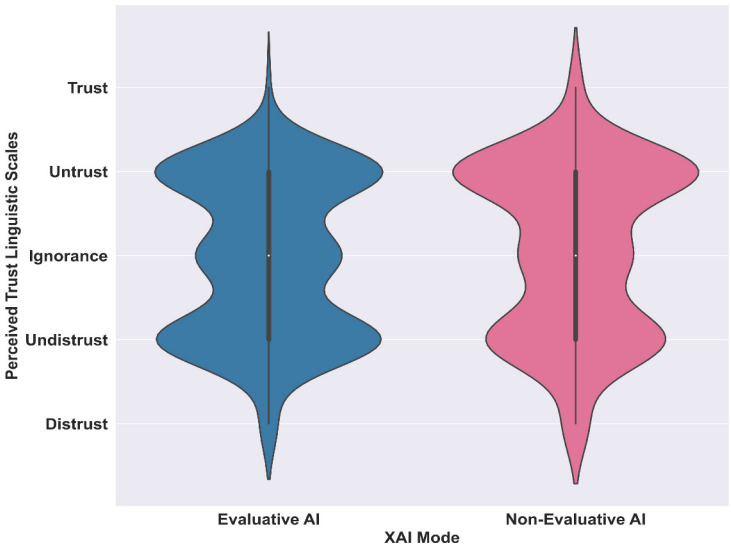

Fig. 3. Violin Plot of Perceived Trust after Training Phase

5.3 Trust in Test Phase

This section addresses RQ3 and RQ4, focusing on identifying which factors most significantly influence human subjects' trust both in terms of perceived trust and demonstrated trust. Visualizing the three types of trust measured in this experiment, as shown in Fig. 4, reveals that most participants either felt undistrust or untrust in the model by the end of the training phase. During the test phase, they largely avoided delegating decisions to the AI model. However, by the end of the experiment, there was a noticeable tendency for participants to perceive trust in the model. This shift prompts further investigation into how demonstrated trust influences perceived trust after the test phase, despite the fact that participants predominantly did not show demonstrated trust towards the model. We began by conducting a two-way ANOVA, as detailed in Table 5, to assess whether the XAI mode, risk treatment, or their interaction have a statistically significant effect on demonstrated trust. The results indicate that none of these factors significantly impact demonstrated trust. The F-values and corresponding p-values higher than 0.05 indicate that the variations in demonstrated trust cannot be attributed to differences in XAI mode or risk treatment. This suggests that demonstrated trust is relatively stable across different experimental conditions. Further analysis is supported by the interaction plot presented in Fig. 5. The plot visually confirms the lack of significant interaction effects between XAI mode and risk treatment on demonstrated trust. The parallel lines across the XAI modes suggest that the combined influence of XAI mode and risk treatment does not lead to meaningful differences in demonstrated trust. Therefore, we cannot support the hypothesis that different XAI modes or risk treatments have a significant impact on demonstrated trust. Our findings indicate that human subjects, regardless of XAI mode or risk treatment, tend to rely primarily on their own decisions.

Table 5. Two-way ANOVA Results for Demonstrated Trust

	df	sum_sq	mean_sq	F	PR(>F)
XAI Mode	1.0	0.2150	0.2150	2.3164	0.1289
Risk Treatment	1.0	0.1960	0.1960	2.1115	0.1470
XAI Mode:Risk Treatment	1.0	0.0028	0.0028	0.0308	0.8606
Residual	356.0	33.0559	0.0928		

The following two-way ANOVA investigates whether XAI mode and risk treatment impact perceived trust after the test phase. Table 6 reveals that the only significant factor influencing perceived trust after the test phase is the interaction between XAI mode and risk treatment (p-value=0.0472<0.05). This finding is further supported by the interaction plot presented in Fig. 6, which illustrates how different combinations of XAI modes and risk treatments influence perceived trust. The plot shows that the impact of XAI mode on trust is

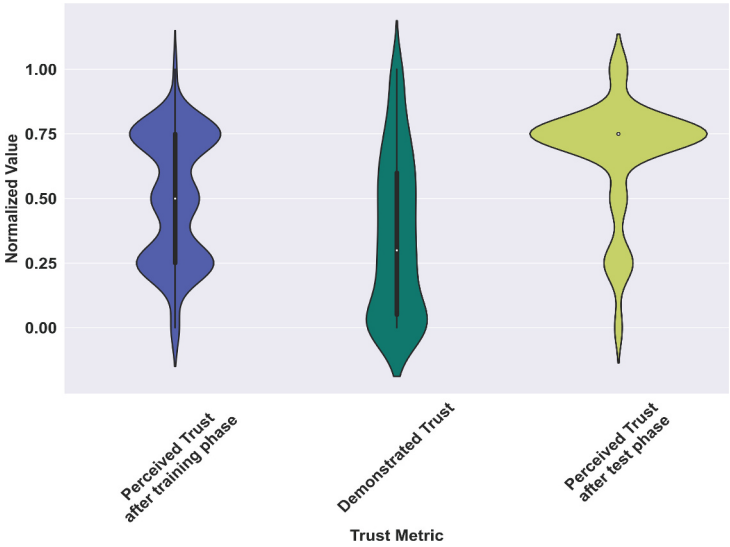

Fig. 4. Violin Plot for All Measured Trust

modulated by the level of risk treatment, suggesting that the effectiveness of AI explanations varies depending on the contextual risk factors.

Table 6. Two-way ANOVA Results for Perceived Trust

	df	sum_sq	mean_sq	F	PR(>F)
XAI Mode	1.0	0.0277	0.0277	0.5748	0.4488
Risk Treatment	1.0	0.1303	0.1303	2.7000	0.1012
XAI Mode:Risk Treatment	1.0	0.1913	0.1913	3.9635	0.0472
Residual	356.0	17.1893	0.0482		

To investigate the significant interaction effect observed in the two-way ANOVA, we conducted post-hoc Tukey HSD analysis for both XAI modes and risk treatments. The results, presented in Table 7, indicate that despite the significant interaction effect identified by the ANOVA, the Tukey HSD comparisons do not show significant differences between the specific groups, as the p-values for all comparisons exceed 0.05. This suggests that while there is a significant interaction, the pairwise comparisons alone do not capture the underlying complexities or additional factors influencing the results. To further explore and understand the nature of this interaction effect, we proceeded with additional ANCOVA analysis.

Analyzing the ANCOVA results shown in Table 8 highlights several important points regarding the influence of XAI modes and risk treatments on perceived

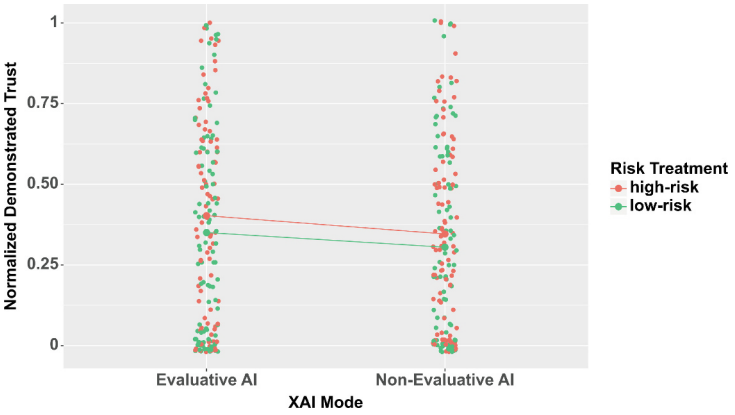

Fig. 5. Interaction Plot of Demonstrated Trust

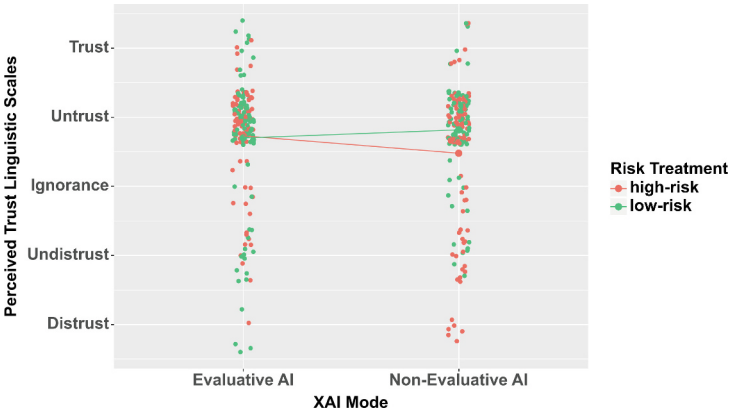

Fig. 6. Interaction Plot of Perceived Trust

Table 7. Post-Hoc Analysis: Tukey HSD Results

	XAI Modes	Risk Treatment
Group 1	Evaluative AI	High-risk
Group 2	Non-evaluative AI	Low-risk
Mean Diff.	−0.0176	0.0388
Adj. p-value	0.4518	0.0959
Lower CI	−0.0634	−0.0069
Upper CI	0.0283	0.0845
Reject	False	False

trust after the test phase. Although the effect of risk treatment is not statistically significant, it approaches significance, suggesting it may have a subtle impact on perceived trust. The most noteworthy finding is the significant effect of demonstrated trust on perceived trust after the test phase (RQ4).

Table 8. ANCOVA Test Results

	sum_sq	df	F	PR(>F)
XAI Mode	0.0123	1.0	0.2572	0.6123
Risk Treatment	0.1566	1.0	3.2754	0.0711
XAI Mode:Risk Treatment	0.1876	1.0	3.9234	0.0483
Demonstrated Trust	0.2132	1.0	4.4584	0.0354
Residual	16.9761	355.0		

5.4 Impact of Demonstrated Trust on Perceived Trust

To address the unusual observation that perceived trust increased even when demonstrated trust by the human subject was not significant, we conducted further analysis. First, we examined the impact of credit balance on trust. A T-test revealed a significant p-value of 0.0037, indicating a strong relationship between credit balance and demonstrated trust. The Pearson correlation coefficient of 0.3488 suggests a moderate positive correlation between these two variables. Figure 7 is another piece of evidence that shows human subjects with higher demonstrated trust end up with higher credit balances at the end of the experiment, though there is some variability at certain balance levels. In the same way, the p-value and Pearson correlation coefficient between credit balance and perceived trust after the test phase were 0.0750 and 0.1772, respectively. This confirms that the main reason for the increased perceived trust after the test is not the credit balance, which could otherwise serve as a motivating factor to enhance trust.

The reason behind these results largely stems from the AI model's role in decision-making. As depicted in Fig. 8, human participants relied on their own knowledge to classify mushrooms in 65% of the decision-making tasks. This aligns with earlier conclusions that the XAI mode and risk treatment did not significantly impact demonstrated trust, and this plot further supports that finding. There is no significant difference between the correct decisions made by participants influenced by the XAI mode, suggesting that evaluative AI did not substantially help participants gain more knowledge in mushroom classification. However, evaluative AI did reduce the number of incorrect decisions. Notably, the error rate among human participants in correctly classifying mushrooms is quite high at 44.8%. This increased reliance on their own decision-making correlates with a higher error rate in mushroom classification. This is particularly striking given that 88.8% of participants reported having little to no prior knowledge of

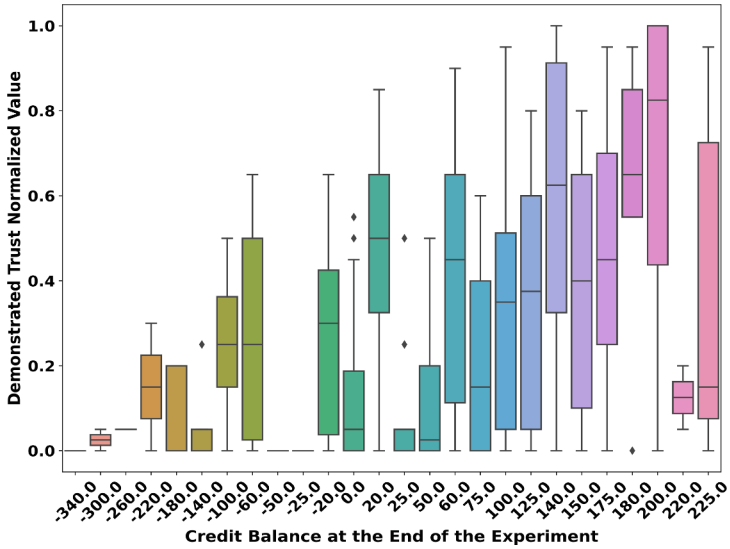

Fig. 7. Relationship Between Credit Balance and Demonstrated Trust

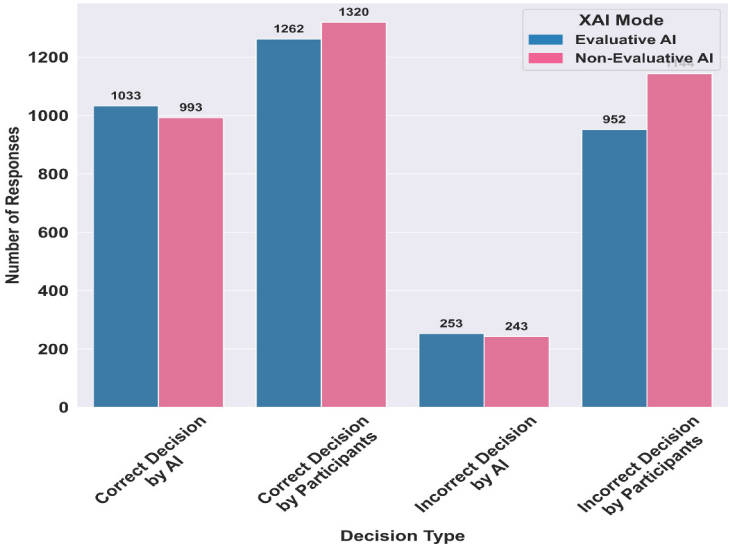

Fig. 8. Number and Type of Responses in Risk Treatment

mushroom identification, indicating a generally low demonstrated trust rate in the decision-making tasks. Conversely, in 35% of the decision-making tasks, participants chose to delegate the decision to the AI model. The AI model, with an error rate of 19.67% and a performance close to its expected prediction accuracy of 80%, was able to correctly identify the mushroom instances. This suggests

that, within the context of the experiment, the AI model was more accurate than the human participants. In conclusion, the lower rate of incorrect decisions made by the AI likely influences participants' mental models of perceived trust, leading them to report higher levels of trust after the decision-making task. The most significant finding of this study is the clear distinction between demonstrated trust and perceived trust–two fundamentally different aspects of trust that should be studied independently (RQ5). According to the proposed trust definitions, demonstrated trust aligns more closely with a human's practical understanding of trust. On the other hand, perceived trust is not only fragile and easily swayed, but it also fails to accurately reflect real-world behavior, where trust is demonstrated by the actual delegation of decisions to AI models.

6 Conclusion and Discussion

In this study, we highlighted the substantial difference between demonstrated trust and perceived trust in AI-assisted decision-making. Our results indicate that there is no significant difference in human subjects' satisfaction with the efficacy of explanations provided by different XAI modes; satisfaction rates are relatively high in both cases. We also found that perceived trust during the training phase does not significantly differ across XAI modes, with human subjects generally displaying mostly undistrust and untrust toward XAI. Another key finding is the lack of a significant effect of XAI modes and risk treatment on both demonstrated and perceived trust. However, despite human users largely not delegating decisions to the AI model during the test phase, this interaction significantly impacts perceived trust, leading most subjects to develop a tendency to trust the model. It turns out that, beyond satisfaction with credit balance at the end of the experiment, the higher accuracy of the AI model leads to increased perceived trust. This finding shows that perceived trust is fragile and that human mental models can be influenced by the noticeable performance of an AI model, even when they do not delegate their decisions to it.

This paper has some limitations that we plan to address in future studies. First, we intend to include an additional control group without the risk of penalization to better understand how risk influences human subjects' tendency to rely on their own decisions rather than the AI's. Second, we plan to design a novel experiment that improves human subjects' decision-making accuracy to align more closely with that of the AI. This will enable us to study how such a setup impacts decision delegation to AI and overall trust. Third, we aim to compare human reliance on AI with perceived trust and demonstrated trust. Unlike perceived trust, which is based on subjective self-reports, reliance reflects objective behavior in response to AI system's recommendations. This study could be highly insightful for modeling human mental models during interactions with AI. It aims to distinguish differences in behavior based on subjective perceptions, objective actions, and demonstrated trust.

Acknowledgements. G. Baer, C. Zhang, and I. Grau are supported by the European Union's HORIZON Research and Innovation Program under grant agreement

No. 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI). Disclosure of Interests All authors declare that they have no conflicts of interest.

References

1. Ahn, D., Almaatouq, A., Gulabani, M., Hosanagar, K.: Impact of model interpretability and outcome feedback on trust in AI. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–25 (2024)
2. Amarasinghe, K., et al.: On the importance of application-grounded experimental design for evaluating explainable ML methods. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 20921–20929 (2024)
3. Bansal, G., 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, pp. 1–16 (2021)
4. Buçinca, Z., Malaya, M.B., Gajos, K.Z.: To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc. ACM Hum. Comput. Interact. **5**(CSCW1), 1–21 (2021)
5. Cho, J.H., Chan, K., Adali, S.: A survey on trust modeling. ACM Comput. Surv. (CSUR) **48**(2), 1–40 (2015)
6. Ehsan, U., Liao, Q.V., Passi, S., Riedl, M.O., Daumé, H.: Seamful XAI: operationalizing Seamful design in explainable AI. Proc. ACM Hum. Comput. Interact. **8**(CSCW1), 1–29 (2024)
7. Figueiredo, M.C., Ankrah, E., Powell, J.E., Epstein, D.A., Chen, Y.: Powered by AI: examining how AI descriptions influence perceptions of fertility tracking applications. Proc. ACM Interact. Mobile Wearable Ubiquitous Technol. **7**(4), 1–24 (2024)
8. Gajos, K.Z., Mamykina, L.: Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In: Proceedings of the 27th International Conference on Intelligent User Interfaces, pp. 794–806 (2022)
9. González, A.V., et al.: Do explanations help users detect errors in open-domain QA? An evaluation of spoken vs. visual explanations. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1103–1116 (2021)
10. Goyal, N., Baumler, C., Nguyen, T., Daumé III, H.: The impact of explanations on fairness in human-AI decision-making: protected vs proxy features. In: Proceedings of the 29th International Conference on Intelligent User Interfaces, pp. 155–180 (2024)
11. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Front. Comput. Sci. **5**, 1096257 (2023)
12. 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, pp. 624–635 (2021)
13. Kuhl, U., Artelt, A., Hammer, B.: Let's go to the alien zoo: introducing an experimental framework to study usability of counterfactual explanations for machine learning. Front. Comput. Sci. **5**, 1087929 (2023)

14. Kunkel, J., Donkers, T., Michael, L., Barbu, C.M., Ziegler, J.: Let me explain: impact of personal and impersonal explanations on trust in recommender systems. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2019)
15. Li, Z., Lu, Z., Yin, M.: Modeling human trust and reliance in AI-assisted decision making: a Markovian approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 6056–6064 (2023)
16. Li, Z., Lu, Z., Yin, M.: Decoding AI'S nudge: a unified framework to predict human behavior in AI-assisted decision making. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 10083–10091 (2024)
17. Liu, F., Lv, J., Cui, S., Luan, Z., Wu, K., Zhou, T.: Smart "error"! Exploring imperfect AI to support creative ideation. Proc. ACM Hum. Comput. Interact. **8**(CSCW1), 1–28 (2024)
18. Lu, Z., Li, Z., Chiang, C.W., Yin, M.: Strategic adversarial attacks in AI-assisted decision making to reduce human trust and reliance. In: IJCAI, pp. 3020–3028 (2023)
19. Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M., Ma, X.: are you really sure? Understanding the effects of human self-confidence calibration in AI-assisted decision making. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–20 (2024)
20. Miller, T.: Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 333–342 (2023)
21. Mishra, A., Nayak, G., Bhattacharya, S., Kumar, T., Shah, A., Foltin, M.: LLM-guided counterfactual data generation for fairer AI. In: Companion Proceedings of the ACM on Web Conference 2024, pp. 1538–1545 (2024)
22. Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., Perer, A.: The impact of imperfect XAI on human-AI decision-making. Proc. ACM Hum. Comput. Interact. **8**(CSCW1), 1–39 (2024)
23. Mozannar, H., Lee, J., Wei, D., Sattigeri, P., Das, S., Sontag, D.: Effective human-AI teams via learned natural language rules and onboarding. Adv. Neural Inf. Process. Syst. **36** (2024)
24. Okolo, C.T., Agarwal, D., Dell, N., Vashistha, A.: "if it is easy to understand then it will have value": examining perceptions of explainable AI with community health workers in rural India. Proc. ACM Hum. Comput. Interact. **8**(CSCW1), 1–28 (2024)
25. Pafla, M., Larson, K., Hancock, M.: Unraveling the dilemma of AI errors: exploring the effectiveness of human and machine explanations for large language models. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–20 (2024)
26. Qu, J., Arguello, J., Wang, Y.: Why is "problems" predictive of positive sentiment? A case study of explaining unintuitive features in sentiment classification. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 161–172 (2024)
27. Reingold, O., Shen, J.H., Talati, A.: Dissenting explanations: leveraging disagreement to reduce model overreliance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 21537–21544 (2024)
28. 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 (2016)

29. Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., Höllerer, T.: I can do better than your AI: expertise and explanations. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 240–251 (2019)
30. Schoeffer, J., De-Arteaga, M., Kuehl, N.: Explanations, fairness, and appropriate reliance in human-AI decision-making. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–18 (2024)
31. Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., Seaborn, K.: Trust in human-AI interaction: scoping out models, measures, and methods. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts, pp. 1–7 (2022)
32. Wagner, D., Heider, D., Hattab, G.: Mushroom data creation, curation, and simulation to support classification tasks. Sci. Rep. **11**(1), 8134 (2021)
33. Wang, X., Kim, H., Rahman, S., Mitra, K., Miao, Z.: Human-LLM collaborative annotation through effective verification of LLM labels. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–21 (2024)
34. Wang, X., Yin, M.: Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In: Proceedings of the 26th International Conference on Intelligent User Interfaces, pp. 318–328 (2021)
35. Yang, F., Huang, Z., Scholtz, J., Arendt, D.L.: How do visual explanations foster end users' appropriate trust in machine learning? In: Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 189–201 (2020)
36. Zhang, Y., Liao, Q.V., Bellamy, R.K.: 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, pp. 295–305 (2020)