Hindawi
Human Behavior and Emerging Technologies
Volume 2023, Article ID 4637678, 12 pages
https://doi.org/10.1155/2023/4637678

Research Article

AI Trust: Can Explainable AI Enhance Warranted Trust?

Regina de Brito Duarte,[1] **Filipa Correia**,[2] **Patrícia Arriaga**,[3] **and Ana Paiva**[1]

[1]*INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal*
[2]*Interactive Technologies Institute, LARSyS, Instituto Superior Técnico, Universidade de Lisboa, Portugal*
[3]*ISCTE-Instituto Universitário de Lisboa (IUL), CIS-IUL, Lisboa, Portugal*

Correspondence should be addressed to Regina de Brito Duarte; regina_duarte@outlook.pt

Received 26 May 2023; Revised 15 September 2023; Accepted 9 October 2023; Published 31 October 2023

Academic Editor: Elliot Mbunge

Explainable artificial intelligence (XAI), known to produce explanations so that predictions from AI models can be understood, is commonly used to mitigate possible AI mistrust. The underlying premise is that the explanations of the XAI models enhance AI trust. However, such an increase may depend on many factors. This article examined how trust in an AI recommendation system is affected by the presence of explanations, the performance of the system, and the level of risk. Our experimental study, conducted with 215 participants, has shown that the presence of explanations increases AI trust, but only in certain conditions. AI trust was higher when explanations with feature importance were provided than with counterfactual explanations. Moreover, when the system performance is not guaranteed, the use of explanations seems to lead to an overreliance on the system. Lastly, system performance had a stronger impact on trust, compared to the effects of other factors (explanation and risk).

1. Introduction

The use of artificial intelligence (AI) systems is becoming increasingly common, with applications such as driverless cars, automated health diagnostics, automated financial decisions, and smart buildings. For AI systems to be relied upon, they must be trustworthy to some extent. This trustworthiness is determined by the AI model's ability to adhere to a specific contract [1], such as those related to privacy and data governance, diversity, nondiscrimination and fairness, and accountability [2]. Ultimately, it is the user who must trust the system to take advantage of its capabilities effectively, and this trust may not necessarily be related to the trustworthiness of the system [3]. This has led to an increased focus on AI trust, with research being conducted in both academia and the AI industry. In healthcare, for example, there is already relevant work on AI trust and the problems associated with it [4], such as the disuse of AI tools and inefficient collaboration with clinical decision support tools [5].

Of the many concerns that can be raised about trusting an AI system, we highlight the following: a mismatch between AI trust (how much someone trusts it) and its inherent trustworthiness [6]. This mismatch can lead to either overreliance on the AI system, or underreliance on it [7]. Recent research has shown that even in scenarios where an AI system assists on a human decision-related task and where its suggested outcome is improved, people might still doubt the provided information from it [8]. One of the most used techniques to mitigate mistrust in AI and even increase its trust is the use of explainable artificial intelligence (XAI) models. These models are developed with the goal of generating explanations about how a system derives its predictions/recommendations so that the user can better understand its inner workings [9]. This approach is based on two premises: (1) explanations are required to elucidate the logic of AI systems for better comprehension, and (2) the explanations generated from XAI models are sufficient to enhance AI trust in the system. Premise 1 is still under open debate, as it is unclear whether the future of AI systems

should be based on black box models or not, at least in high-risk domains [10]. Many researchers believe that the best solution is the development and use of AI systems that are intrinsically interpretable themselves [10]. Regarding premise 2, there is still no strong evidence that the explanations generated by XAI models are sufficient to enhance AI trust. There is evidence that certain explanations can help, but it depends on the type and/or context of the XAI model [11].

Our primary objective is to delve into the intricate interplay between XAI models and AI trust, while also examining potential moderating variables. This investigation seeks to gain insight into user trust in AI. To achieve this, we formulated a hypothesis regarding the impact of XAI models on AI trust, taking into consideration the risk level and the performance (or trustworthiness) of the AI system. With this goal in mind, we conducted an experimental study centred around a decision-making task featuring a recommendation system. Through the analysis of various variables, we aimed to disentangle the concepts of the "intrinsic trustworthiness" of a system and the trust attributed to it by users.

This paper addresses the following three research questions:

RQ1. How do different explanations affect user's trust in AI systems?

RQ2. How do the levels of risk of the user's decision-making play a role in the user's trust in AI systems?

RQ3. How does the performance of the AI system play a role in the user's AI trust even when an explanation is present?

2. Related Work

2.1. XAI Models. The utilization of AI systems is on the rise; however, they often operate as complex black box models that defy human comprehension [12]. Explainable AI (XAI) research strives to address this challenge by devising methods to elucidate the reasoning behind AI predictions. The simplest approach for model explanation involves deriving meaning directly from the model itself, leading to interpretable machine learning (ML) models [10]. These models are straightforward and comprehensible in their behaviour. They typically encompass logistic and linear regressions, decision trees, k-nearest neighbours, and rule-based learners [13]. In contrast, post hoc explanations [14] materialize after training the ML model and result from a separate system whose sole function is generating explanations. This category encapsulates diverse techniques that diverge in their mechanisms, objectives, and outputs. In response, efforts have been made to categorize and group these techniques [9, 15]. Among these categorizations, three major classes of XAI post hoc explanations have emerged: feature importance explanations, example-based explanations, and explanations through simplification [15].

2.1.1. Feature Importance Explanations. This category of techniques is aimed at pinpointing the contributions of individual variables to the AI system's predictions [16]. This can occur either on a local scale, involving the significance of each feature for a single observation, or globally, encompassing the importance of each feature for the entire model. Prominent techniques in this group include the Shapley values [17], the LIME technique [18], and saliency maps [19].

2.1.2. Explanations by Example. This category presents outcomes in the form of examples akin to the observation with identical model predictions or counterfactual examples illustrating divergent model predictions. The underlying concept is that users can comprehend the model's rationale by contrasting examples with original observations, thus drawing informed conclusions [20–22].

In this context, counterfactual examples are hypothetical instances demonstrating how a distinct model prediction could have been achieved based on a given observation [23]. They illustrate variations of the same observation with slight alterations that would result in a different outcome.

2.1.3. Explanations by Simplification. This category streamlines the model's rationale by formulating straightforward general rules that elucidate its behaviour. Rule-based learners such as decision trees and genetic programming rule-based extractions can be applied atop the ML model.

In the realm of AI, the most prevalent explanations fall under the feature importance category. This holds true when utilizing explanations to bolster trust. Similarly, example-based explanations, which showcase counterfactual instances, are deemed a method to emulate human thought processes and reconstruct counterfactual scenarios [24]. Consequently, this study centres on these two explanation categories, given their frequent association with AI trust.

2.2. Trust on AI Systems. The primary incentive for employing XAI lies in its potential to bolster user trust in reliable AI systems. However, while prior research has established a positive correlation between explainability and enhanced trust [25, 26], there is also evidence that challenges this notion. A controlled experiment suggested that when users engage in providing feedback to an AI system to enhance its performance, both user trust and their perception of model accuracy decline [27]. In contexts of human-AI collaboration, the perception of an AI system has been found to be influenced by several variables, including communication direction and the nature of the model underpinning the AI system [28]. Beyond the usual human aversion to algorithms [29], research indicates that individuals prefer human decision-making discretion over algorithms that rigidly apply human-derived fairness principles to specific cases [30]. This preference stems from humans' capacity for independent judgment, allowing them to transcend fairness principles as necessary.

These investigations underscore the delicate nature of user trust in AI systems and cast doubt on the assumed benefits of XAI. An additional pertinent study evaluated the utilization of interpretability tools by data scientists as part of their regular workflow [31]. The findings revealed instances of misapplication and unjustified trust due to interpretability tools. The authors proposed that the mere presence of XAI models might lead to unwarranted overtrust. This study

introduces the intriguing notion that XAI explanations could potentially enhance trust in AI systems, irrespective of the AI system's actual performance or trustworthiness.

Moreover, a strand of research has delved into the dynamics of collaboration between AI systems and humans functioning as a unified team. Bansal et al. [32, 33] underscored the significance of user-mental models of AI systems in collaborative settings. They showcased that a parsimonious AI system with an error boundary aligned with the user's mental model could enhance team performance (comprising both AI systems and humans) more effectively than mere model accuracy [32]. This is because humans and AI systems can synergistically operate, with humans comprehending the model's error boundaries and identifying inaccuracies in its predictions. Additionally, the concept of *compatibility* was introduced to describe updates in model performance [33]. These updates should harmonize with prior model versions, ensuring coherency with the user's mental model and maintaining team performance.

Similarly, Wang et al. [34] investigated how human decision-makers in collaborative human-AI settings adopt AI recommendations. Their findings indicate that, in AI-assisted decision-making, human decisions are influenced by their individual judgment and confidence in the decision-making process. People often lean on their own judgment to assess whether to adopt AI recommendations. Moreover, as the stakes of decisions intensify, individuals tend to diminish their faith in AI recommendations' accuracy and rely more heavily on their own judgments.

Studies exploring collaborative human-AI scenarios also examined trust in AI assistants with varying levels of expertise [35]. The outcomes revealed that participants could distinguish between expert and nonexpert AI assistants for a given task, enabling them to calibrate their reliance on AI to optimize team performance.

The consensus is that explainable AI models play a pivotal role in understanding AI system decisions and enhancing user confidence and trust in these systems [6]. Previous research has indeed explored and substantiated the assumption that AI explainability positively impacts trust [25, 36]. Additionally, evidence suggests that counterfactual explanations, due to their natural contrastive attributes aligning with human causal reasoning, offer a valuable means of explaining models [24, 37]. However, despite this, the most widely adopted XAI technique continues to be feature importance explanations.

Looking at the risk involved in a decision-making task, Jacovi et al. [1] posit that trust in AI cannot exist without an element of risk, while Mayer et al. [38] introduced a social trust model that perceives risk as an outcome of trust. These distinct perspectives on risk within trust scenarios can be harmonious. Nevertheless, a consensus on how risk impacts trust in AI when explanations are incorporated remains absent. These studies contribute insights into trust in AI and users' expectations. Factors such as interactions, system communication, and the system's mental model are pivotal in decision-making tasks. However, to our current knowledge, the impact of incorporating explanations into decision-making scenarios on trust in AI remains uncharted territory in cases where the system's trustworthiness is uncertain, and task stakes are high.

3. Hypotheses

The following hypotheses state our expectations regarding the effect of explanations on AI trust in decision-making scenarios, taking into account the stakes of the decision and the trustworthiness of the AI model. The hypotheses were based on the body of literature on AI trust and XAI presented in the previous section.

H1. Trust in AI systems is higher when explanations of the system are present (specifically explanations of the importance of features and counterfactual explanations) compared to the absence of explanations

H2. The presence of explanations enhances trust in AI systems, regardless of the AI performance

H3. Trust in AI is lower when the level of risk is high

4. Method

To address our hypotheses, we created a web-based game as an experimental platform. This game features a recommender system (AI) that assists players in making a series of decisions. The context of the game and the AI model was centred around the problem of detecting the edibility of mushrooms, a scenario commonly employed in AI applications. Both the game and the AI model were built from the ground up. In the upcoming sections, we outline the development process of both the game and the AI model.

4.1. Game Development. The *mushroom game* is a virtual game in which players evaluate the edibility of mushrooms, determining whether to consume them or avoid them. On the screen, information is presented on the characteristics of mushrooms and a cartoon depiction of them to aid in decision-making. For additional guidance, an AI recommender system suggests whether the mushroom should be eaten or avoided.

The game encompasses a series of 12 mushrooms, each requiring a decision from the player. Thus, there are 12 distinct decision-making tasks to complete. Following each decision (to eat or avoid the mushroom), the player receives feedback on its soundness, indicating whether it was a favourable choice or not. Upon completing the sequence, the player receives a final score reflecting the count of edible and poisonous mushrooms consumed. The ultimate objective of the game is to consume as many mushrooms as possible while avoiding any poisonous ones.

4.2. AI Model Development. The AI recommender system utilized in the game was developed to predict the edibility of mushrooms. To create this AI model, data was collected from the *Mushroom's dataset*, available in the *UCI Machine Learning Repository* [39]. This dataset contains 8124 entries, encompassing 22 categorical variables and a binary target variable indicating the mushroom's edibility. These variables span various aspects of the mushroom, including odour, habitat, and physical characteristics of the cap, gill, stalk, and veil.

To suit the game's context, the AI model was constructed with a limited set of features for two primary reasons: (1) to maintain a realistic AI model with a nonzero error rate and (2) to present only a manageable number of features to players, reducing cognitive load. Consequently, a feature selection process was performed based on the significance of features in a decision tree baseline model. The evaluation of the importance of the characteristic revealed that odour stood out as the most influential attribute, possessing a substantial predictive capacity to determine the edibility of mushrooms in most observations. Surprisingly, certain characteristics commonly deemed significant for classification, such as cap colour and gill colour, held relatively low importance. Given the game's objectives, our aim was to present mushroom features that were both easily comprehensible and resonated with laypeople's perceptions. Thus, attributes like odour were retained to enhance the AI model's comprehensibility. Similarly, colour attributes were valuable to align with people's mental models of mushrooms. Following this rationale, certain relevant features were omitted, while others initially considered less relevant were retained. Ultimately, the final model incorporated only 8 features: odour, cap colour, cap shape, cap surface, gill colour, gill size, gill spacing, and ring number.

4.2.1. Trustworthy AI Model: Good Performance. The final model was a decision tree classifier trained with 70% of the dataset and with a depth of 12 nodes. The accuracy of the model was 96% in the test set. This level of performance was desired to have some examples misclassified, yet still obtain a good performance. This model was used as the trustworthy AI recommender system in the game.

4.2.2. Untrustworthy AI System: Poor Performance. In the context of the user study, it was imperative to have an "untrustworthy" AI model that mirrored its performance accurately. In pursuit of this goal, the creation of the second AI model entailed manipulating the target variable, edibility, within the dataset. Specifically, 30% of the target observations were altered to their opposing values. Subsequently, we retained the same variables as utilized in the high-performance model and proceeded to train a decision tree classifier. The outcome of this data manipulation was an accuracy of 60% when evaluated on the test set. This model was intentionally designed to represent a lower level of performance, thereby establishing an "untrustworthy" counterpart for the user study.

4.3. User Study. In the study, we wanted to assess AI trust in a decision-making task, where users had the help of an AI system. Consequently, we asked the participants to play the *mushroom game* as an experiment and collected their behaviour when playing.

4.4. Independent Variables. The user study was structured to manipulate three key factors within the mushroom game: the type of explanations of the XAI model, the level of risk involved, and the performance of the AI model.

4.4.1. Type of XAI Model. To test our initial hypothesis concerning the influence of explanations on decision-making, we established three task conditions: a control condition without any AI recommendation explanations and two other conditions wherein users received distinct explanation types during the decision-making process. In one condition, users were presented with a local feature importance explanation generated through the widely recognized LIME technique [18]. In the third condition, users encountered an example-based explanation using the DICE technique, which provides state-of-the-art counterfactual explanations [22]. The variations in the mushroom task, along with the manipulation of XAI types, are illustrated in Figure 1.

4.4.2. The Risk Level. We further manipulated the stakes within the mushroom game across two levels: low and high. In the high-risk scenario, participants were informed that consuming a single poisonous mushroom would result in sickness and consuming a second poisonous mushroom would lead to their character's demise (game over). In contrast, the low-risk scenario conveyed that participants would fall ill only after consuming three poisonous mushrooms.

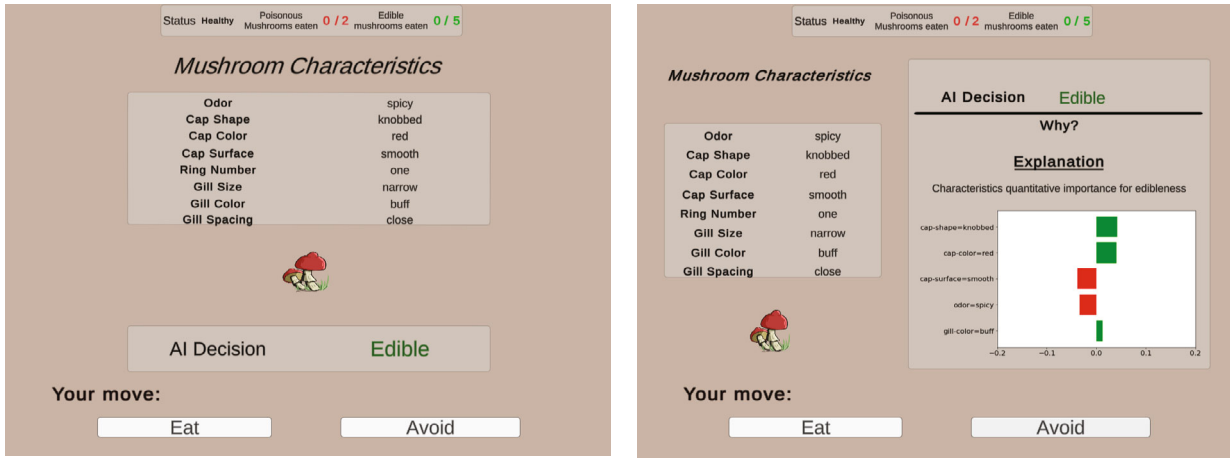
4.4.3. Model Performance. Lastly, we controlled the performance of the AI recommender system to serve as an indicator of the model's reliability. This manipulation involved two performance levels, "high performance" and "low performance," which were designed to reflect the accuracy of the AI system. The construction of these models was detailed in the previous section.

4.5. Outcomes. Given the intricate nature of assessing trust in AI systems [1], our approach encompassed a blend of self-reported and behavioural trust measurements for the AI system. This was supplemented by evaluations of AI understandability and XAI quality. Consequently, following each game instance, participants completed a survey employing the Likert scale responses to gauge these dimensions. To evaluate subjective trust in the AI system, we employed two measures:

(1) The multidimensional measure of trust (MDMT) [40], which gauges indirect subjective trust through participants' assessments of eight-related traits associated with the AI system (reliable, predictable, consistent, skilled, capable, competent, precise, and transparent). Responses were provided on a scale from 0 (not at all) to 7 (very) and were averaged

(2) Direct subjective trust, using a single item that directly inquired about participants' perceived trust in the AI system, utilizing the same Likert-type response format

Additionally, we assessed the system's understandability (AI understandability) using two items ("How well did the game help you understand how the AI system works?" and "How would you rate your understanding of how the AI system works?"). The quality of the XAI model's explanations (XAI quality) was measured with four items, such as "The



(a) Control layout

(b) LIME layout

(c) DICE layout

Figure 1: Layouts of the mushroom game presented on the screen to the participants corresponding to each XAI model condition. In the mushroom game, participants could see the mushroom characteristics, a cartoon picture of the mushroom, and the AI recommendation. In the case of the two conditions with the presence of explanations, participants were also provided with explanations of the AI recommender system.

information presented on the screen for making decisions had sufficient details" and "The information presented on the screen for making decisions was satisfying." Similar to the subjective trust evaluation, these measures utilized the same response format, with averaged values calculated for each measure.

Additionally, we gathered data on participants' behaviour. We assessed behavioural trust by calculating the percentage of instances in which participants agreed with the recommendations provided by the AI system and subsequently made decisions based on those recommendations. We measured Delegation by considering whether participants chose to delegate the decision-making to the AI or not, representing this as a binary variable.

4.6. Participants and Procedure. We recruited a total of 215 participants from Prolific, a crowdsourcing platform designed for extensive data collection. To ensure data quality, we specifically selected participants who were proficient in English. After eliminating responses that indicated a lack of attention (using attention check questions), we retained data from 211 participants. Within this sample, 114 identified as women, 94 as men, and 3 as nonbinary. The participants reported an average age of 27 years, ranging from 19 to 61 years. For participation in the study, each respondent was fairly compensated and received a payment of 2.50 GBP. On average, participants took around 12.12 minutes to complete the task, translating to a median hourly compensation rate of 11.41 GBP.

The allocation of participants to the XAI model and risk conditions followed a between-subject approach, while AI system performance was manipulated within subjects. Consequently, each participant was randomly assigned to a specific XAI manipulation and risk condition. In addition, participants participated in the game twice, undergoing both high and low AI performance conditions in a counterbalanced sequence. Therefore, participants underwent the following procedure. They began the experiment upon voluntarily providing their consent. Then answered general demographic questions, covering age, gender, nationality, and education level. Subsequently, participants received instructions and rules for the mushroom game, followed by

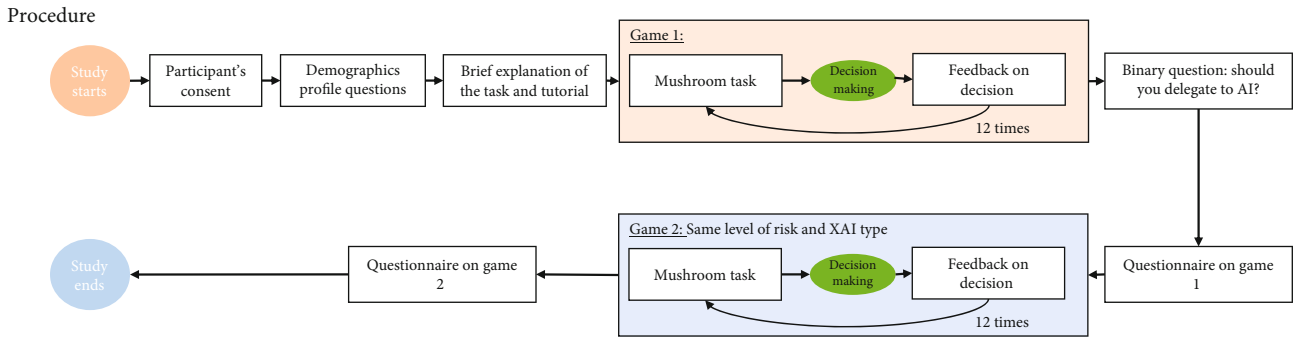


Figure 2: Experiment workflow.

Table 1: Number of participants in each condition of risk and XAI model.

Risk	XAI condition			Total
	LIME	DICE	Control	
High	33	33	31	97
Low	35	39	40	114
Total	68	72	71	211

a straightforward tutorial to acquaint them with the gameplay. Upon completing the tutorial, they proceeded to the first game. In the last stage of the game, participants indicated whether they would delegate the upcoming action to the AI recommender system. Once the mushroom game ended, a final score was displayed, prompting participants to participate in a survey that gauged their views about the game and the AI recommendation system. Afterward, participants engaged in a second round with a different AI performance level and subsequently filled out a survey pertaining to this round as well. Figure 2 illustrates the experimental setup.

5. Results

To examine our hypotheses, we employed two main approaches. Firstly, we utilized a mixed-design analysis of variance (ANOVA) through SPSS Statistics 26. This involved a $3 \times 2 \times 2$ design, considering the factors of the XAI model, risk level, and model performance, and applied to five key dependent variables. Secondly, we employed a chi-squared test to analyse the dependent variable *Delegation*. The significance level was set at 0.05, and for adjustments due to multiple testing, we employed a Bonferroni alpha correction.

Our primary focus is on the outcomes that pertain to our hypotheses. Additionally, a succinct overview of the study's findings can be found in a preceding paper that reports on initial results [41].

Table 1 shows how the participants ($N = 211$) were distributed on the two between-subject factors: risk and XAI model type.

5.1. Self-Reported Measures. The effects on self-reported measures *direct MDMT*, *direct subjective trust*, *AI understandability*, and *XAI quality* were similar.

Significant main effects of model performance were observed on the variables *MDMT*, *direct subjective trust*, *AI understandability*, and *XAI quality*. Model performance exhibited the most substantial influence on all four of these dependent variables. As anticipated, reported levels of trust in the AI system displayed a notable increase when the AI system's performance was high, in contrast to instances of lower AI system performance. Similarly, both the comprehensibility of the AI system and the perceived quality of explainable AI (XAI) were heightened in cases where the AI model's performance was high. Detailed means, standard errors, and statistical outcomes are presented in Table 2.

Furthermore, the XAI factor exhibits a significant main effect on MDMT, AI understandability, and XAI quality. However, in the case of the direct subjective trust measure, the XAI factor does not show a significant main effect. Detailed statistical results are presented in Table 3.

The results of pairwise tests specify that participants exposed to the LIME XAI condition (explanations incorporating feature importance) reported higher levels of subjective trust (measured by MDMT) and greater understanding of the AI system and perceived the explanations as having higher quality compared to those in the other two conditions (DICE with counterfactual explanations and the control group). However, the comparison between the DICE and control conditions did not show statistically significant differences between these three variables. Mean values and standard deviations can be located in Table 3, with statistically significant pairs of groups ($p < 0.05$) denoted by a pair of * or † symbols.

None of the four self-reported dependent variables exhibited significant interactions between the XAI model and the other independent variables in the $3 \times 2 \times 2$ ANOVAs. This consistent pattern persisted when examining various levels of risk and AI system performance. The findings presented in Figure 3 illustrate that the pattern of subjective trust in the AI system, whether assessed directly or indirectly through the MDMT scale, remained consistent across different XAI conditions, irrespective of the model's performance.

Table 2: ANOVA results for the main effects of performance on the dependent variables. All the dependent variables show statistically significant results on AI performance. When the AI system has high performance, trust, understandability, and XAI quality perception are also high compared to low AI performance.

Dependent variable	High performance ± (SE)	Low performance ± (SE)	$F(1,205)$	p value	η_p^2
MDMT	5.34 ± 0.85	4.05 ± 0.97	172.83	<0.001	0.46
Direct subjective trust	4.84 ± 0.11	3.46 ± 0.12	125.64	<0.001	0.38
Behavioural trust	0.87 ± 0.01	0.8 ± 0.01	37.71	<0.001	0.16
AI understandability	4.53 ± 0.12	3.92 ± 0.13	45.34	<0.001	0.18
XAI quality	4.85 ± 0.10	4.11 ± 0.11	92.03	<0.001	0.31

Table 3: ANOVA results for the main effect of the XAI type on the dependent variables. Among these dependent variables, MDMT, AI understandability, and XAI quality exhibit a significant main effect related to the XAI model. In particular, direct subjective trust demonstrates trends analogous to other self-reported measures: trust levels were elevated when LIME explanations were presented. Pairwise statistically significant comparisons are shown by the symbols * and †. Furthermore, a strong correlation is observed between the two subjective trust measures, MDMT and direct subjective trust ($r = 0.80$, $p < 0.001$).

Dependent variable	DICE ± (SE)	LIME ± (SE)	Control ± (SE)	$F(2,205)$	p value	η_p^2
MDMT	$4.60 \pm 0.13^{*}$	$4.98 \pm 0.14^{*\dagger}$	$4.50 \pm 0.13^{\dagger}$	3.50	0.032	0.033
Direct subjective trust	4.02 ± 0.17	4.45 ± 0.18	3.97 ± 0.17	2.26	0.10	0.022
Behavioural trust	0.83 ± 0.02	0.84 ± 0.02	0.84 ± 0.02	0.23	0.79	0.002
AI understandability	$3.99 \pm 0.20^{*}$	$4.70 \pm 0.20^{*\dagger}$	$4.00 \pm 0.20^{\dagger}$	4.10	0.018	0.038
XAI quality	$4.40 \pm 0.16^{*}$	$5.01 \pm 0.17^{*\dagger}$	$4.04 \pm 0.17^{\dagger}$	8.61	<0.001	0.078

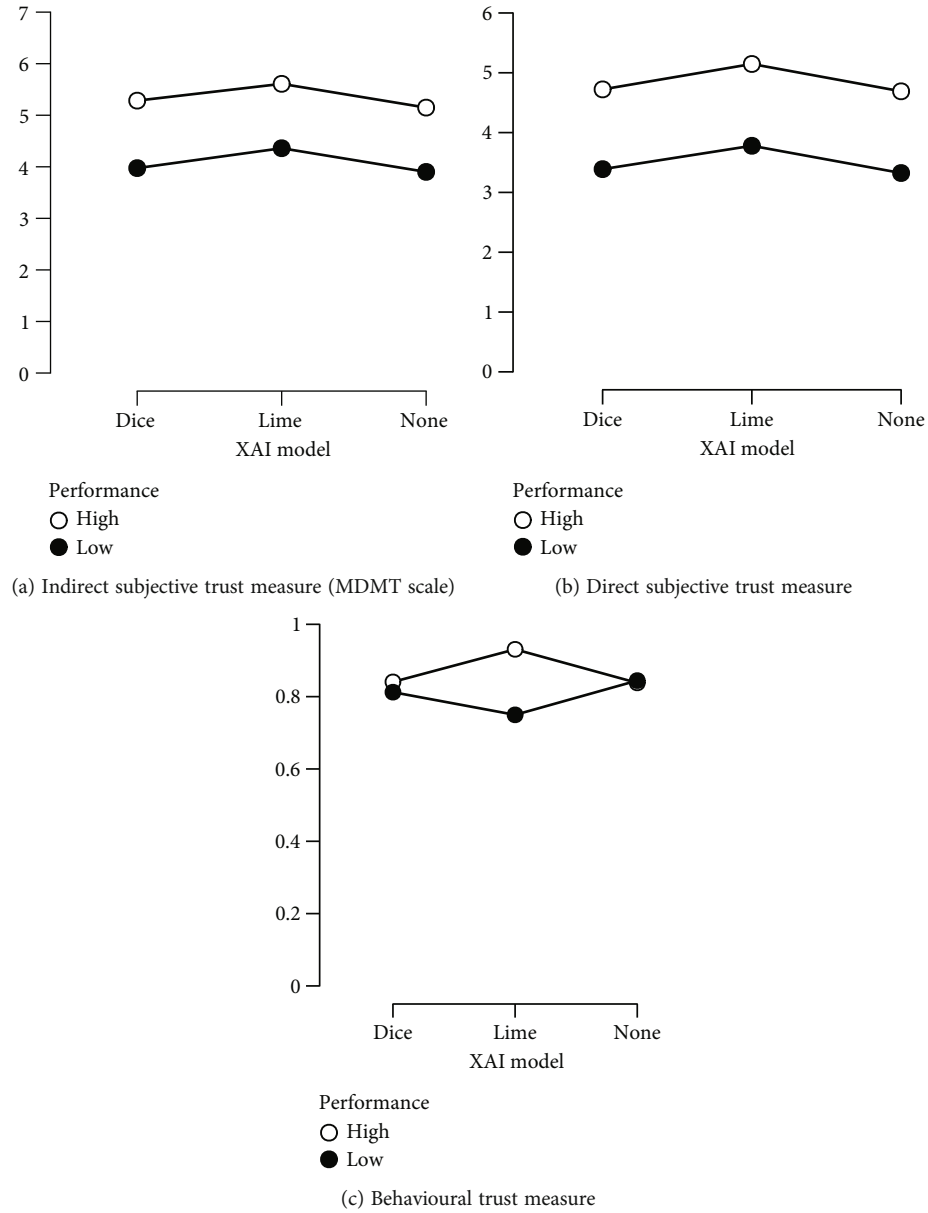
Regarding the impact of risk, a consistent pattern emerged across all four self-reported measures, revealing significant interactions between risk and performance. Participants displayed less trust, less comprehension, and lower evaluations of AI explanation quality in the high-risk condition compared to the low-risk condition, but solely when the AI system's performance was low. Conversely, when the AI system's performance was high, no noteworthy differences were observed in any of the dependent variables. Detailed mean values, standard errors, and statistical outcomes can be found in Table 4.

5.2. Behavioural Trust and Delegation. In relation to the assessment of *behavioural trust*, which gauges the extent of compliance with AI recommendations, a significant main effect of model performance was identified, as presented in Table 2.The rate of alignment with AI recommendations is higher when the AI system's performance is higher, compared to when it is lower.

The influence of the XAI model on *behavioural trust* revealed a distinctive pattern compared to the findings obtained using self-reported measures. While the primary effect of the XAI condition did not reach statistical significance, a notable interaction emerged between the XAI model and AI system performance ($F(2,205) = 25.40$, $p < 0.001$, $\eta_p^2 = 0.20$). As depicted in Figure 3(c), when AI system performance was elevated, participants exposed to the LIME condition (0.93 ± 0.02) exhibited greater trust by adhering more closely to system recommendations than participants in other conditions (0.84 ± 0.02 in DICE and control). However, in cases where performance was lower, participants in the LIME condition (0.75 ± 0.02) displayed significantly reduced *behavioural trust* compared to the other two conditions (0.81 ± 0.02 in DICE and 0.84 ± 0.02 in the control condition). Once again, *behavioural trust* between the DICE and control conditions did not exhibit statistically significant differences. These outcomes mirror the trends observed in *MDMT*, *AI understandability*, and *XAI quality*.

Unlike the self-reported measures, the results for *behavioural trust* did not reveal any significant effects of risk ($F(1,205) = 0.566$, $p = 0.453$, $\eta_p^2 = 0.003$). Participants' compliance with recommendations did not differ based on risk, even when the model's performance was low. Consistent with the self-reported findings, the factor that most significantly impacted *behavioural trust* was model performance. Participants demonstrated a stronger tendency to adhere to system recommendations when the AI system's performance was higher rather than lower. Mean values and standard errors can be found in Table 2.

Regarding *Delegation*, an effect of risk was identified specifically when model performance was low. The inclination to automate decisions by allowing the AI system to decide appeared to be contingent on risk, but only when the AI system's performance was low. Delegation was more pronounced in instances of low risk ($X(1) = 9.52$, $p = 0.002$).



(a) Indirect subjective trust measure (MDMT scale)

(b) Direct subjective trust measure

(c) Behavioural trust measure

Figure 3: The average values of the three trust measures in relation to the XAI condition and AI system performance are presented. In terms of subjective trust measures, it becomes evident that when the AI model's performance is reduced, reported trust diminishes in comparison to instances where the system performs better. When considering each type of explanation, participants who were exposed to LIME explanations consistently reported greater trust compared to the other type of explanations (DICE and no explanations), regardless of the performance of the AI system's performance.

Table 4: Results of ANOVA for the risk and performance interaction on the dependent variables.

Dependent variable	High performance		Low performance		$F(2,205)$	p value	η_p^2
	High risk	Low risk	High risk	Low risk			
MDMT	5.29 ± 0.13	5.39 ± 0.12	3.68 ± 0.14*	4.41 ± 0.13*	10.02	0.002	0.047
Direct subjective trust	4.62 ± 0.16	5.06 ± 0.15	2.93 ± 0.18*	3.98 ± 0.17*	5.99	0.015	0.028
Behavioural trust	0.86 ± 0.02	0.86 ± 0.01	0.79 ± 0.02	0.81 ± 0.02	0.130	0.72	0.001
AI understandability	4.45 ± 0.18	4.62 ± 0.16	3.65 ± 0.19*	4.19 ± 0.17*	4.33	0.039	0.021
XAI quality	4.77 ± 0.15	4.93 ± 0.13	3.83 ± 0.16*	4.39 ± 0.15*	6.51	0.011	0.031

* indicates significant results at the significance level of 0.05. In all self-reported measures, when the performance of the AI system is low, trust levels are higher when the risk is low compared to a high-risk scenario.

No other significant effects were observed for the variable of Delegation.

6. Discussion

This user experiment was conducted to evaluate how XAI explanations affect the level of trust in the AI system when it is making decisions in tasks assisted by AI. In the next section, we will dive into the findings of our research.

6.1. AI Trust in Presence of Explanations. This research sought to investigate the effect of providing explanations (counterfactual and feature importance) in decision-making on trust in AI systems. When feature importance explanations (specifically LIME explanations) were provided, users exhibited higher subjective trust and understanding of the AI system. However, the measure of behavioural trust yielded contrasting outcomes based on AI performance levels. For good AI performance, users following LIME explanations adhered more to AI recommendations, but for lower performance, users deviated from AI recommendations. This suggests that users comprehended LIME explanations, which highlighted the importance of different variables in predictions and the associated uncertainty. When using a low-performance model, users faced uncertain recommendations, leading them to rely more on their own judgments, indicating their continued understanding and trust in the system. These findings align with previous research underlining the effectiveness of presenting prediction confidence intervals in calibrating trust in AI systems [42].

Contrary to the expected effect, the presentation of counterfactual explanations generated by DICE did not lead to increased trust, and this was similar to the absence of any explanation (control condition). This finding challenges the common belief that counterfactual explanations, which simulate AI reasoning, are easily understandable and enhance trust [24]. While previous research had indicated that experts feel more competence with AI systems when using counterfactual explanations [25], our study's outcomes differed. Notably, our experiment involved nonexperts, who might lack the technical background to grasp counterfactual examples without prior knowledge, potentially diminishing the explanation's effectiveness.

The results suggest that our initial hypothesis lacks complete support, as only feature importance explanations led to higher trust compared to no explanations. Referencing Zhang et al.'s study [42] on explainability and confidence intervals' impact on trust calibration, it was observed that presenting confidence intervals with AI predictions results in more calibrated user trust, unlike the effect of local feature explanations that did not show this effect. This observation raises doubts about the actual influence of explainability on trust.

The study also investigated the impact of XAI explanations when the model's trustworthiness is in doubt. The hypothesis was that XAI explanations would boost trust regardless of the system's performance. Figures depicting indirect and direct trust measures indicated that the effect of XAI model type remained consistent regardless of high or low system performance. This suggests that even when dealing with untrustworthy AI systems, users exhibited more trust when presented with feature importance explanations compared to seeing only basic decision-making information. Despite lower trust levels for systems with poor performance, the presence of feature importance elevated trust. Consequently, subjective trust was directed towards the system due to the presence of specific explanations, rather than the system's reliability. This finding is aligned with another research where data scientist practitioners overtrusted the XAI methods without a careful evaluation of their explanations [31].

Our hypothesis suggested that higher risk would lower AI trust. The results revealed lower trust in the participants when the risk was high in all measures, but significance was seen only in low-performing AI systems. While risk does impact AI trust, model trustworthiness is a stronger factor; risk's significance diminishes with guaranteed model trustworthiness. The initial hypothesis is partially supported, as risk mainly affects low-performing models. High AI model performance is crucial, especially in high-stakes contexts. Model trustworthiness emerges as the primary predictor of AI trust, in line with previous studies in human-robot interaction scenarios [43]. This trend extends to perceived AI understandability and XAI quality: users report better understanding and higher quality explanations from high-performing systems. This connection, though unexpected, can be attributed to an unambiguous high-performing AI system being easier to comprehend and explain. Another possibility is that a well-performing model might create a misleading impression of understanding. The exact contributor remains unclear.

6.2. General Implications and Future Work. This user experiment has contributed to the growing research literature on the effect of explanations of AI trust. The study evidenced the contextual dependency of the effect of explanations on AI trust in AI-assisted decision-making. The actual impact of explanations can be influenced by various factors, including task difficulty [44], decision-maker expertise, decision design [45], and the cognitive load required to comprehend the explanation [46] and also the type of explanations and the way it is presented.

It is also plausible that systematic errors might arise when assessing explanations in decision-making tasks, thereby impeding the development of trust. These errors encompass factors such as a lack of curiosity during decision-making, absence of context, confirmatory search, misinterpretation of explanations, or the formation of habitual responses [47]. These aspects underscore the challenge of accurately assessing the true impact of explanations and provide insight into the reasons behind conflicting results. The effect of explanation on trust hinges not solely on the explanations themselves but also on the genuine understanding of the explanations, which is complex to evaluate.

Furthermore, the findings about overtrust in untrustworthy scenarios have substantial implications. They suggest that users might align with an AI system primarily due to the

presence of explanations, creating the illusion of increased trustworthiness. In critical domains, these implications could ripple through numerous decisions, potentially resulting in significant injustices and malpractices.

For future work, it is crucial to increase the level of research on the various factors influencing the impact of explanations on the decision-making process and AI trust. This step is essential to enable the effective use of explanations in practical scenarios. Additionally, research and development of XAI techniques to effectively convey the trustworthiness of AI-assisted models are crucial. This approach helps mitigate issues of overtrust in cases involving untrustworthy AI systems.

7. Limitations

One aspect that we need to acknowledge in our current study is the limited scope of XAI types that we focused on. Specifically, we concentrated our assessment on two primary categories of XAI models: counterfactual explanations and feature importance explanations. While these classifications represent substantial subsets of XAI models, we recognize that there are other types that we did not explore in this research.

Furthermore, our experimental design was centred around a decision-making task involving participants from an online recruitment platform, whom we assumed to be individuals without specialized expertise in the particular task. This approach has two important implications. First, we considered the participants' familiarity with assessing mushroom edibility as that of lay users. Second, we acknowledge that our findings might not extend directly to scenarios that involve collaboration between experts and AI. Notably, other studies suggest that counterfactual explanations might be particularly advantageous for decision-making when considering expert knowledge [25].

Lastly, we operated on the premise that an AI system achieving 60% accuracy is indicative of poor performance and lack of trustworthiness. However, this point invites further discussion. It is worth exploring whether 60% accuracy truly corresponds to an untrustworthy scenario and whether our findings regarding this accuracy level can be extrapolated to scenarios with even lower performance.

8. Conclusion

The primary objective of this research was to examine how explainable artificial intelligence (XAI) models influence users' confidence in AI systems when making decisions. This investigation took into account factors such as risk and the reliability of the model as potential influencing factors. The outcomes partly confirmed the original expectations, suggesting that explanations have the capacity to strengthen trust in AI. However, this effect varies depending on the specific kind of explanation given. Visual explanations highlighting the importance of features are more effective for individuals without expertise in the field compared to explanations involving counterfactual reasoning.

An intriguing discovery from the research was that explanations can increase trust in AI even when the AI model itself is not reliable. This observation raises apprehensions regarding the potential misapplication of explanations. They might inadvertently contribute to fostering trust in the AI system rather than instigating a healthy sense of scepticism. Moreover, the investigation unveiled a decline in trust, particularly in situations involving high risk, especially if the AI model lacks reliability. The trustworthiness of the model was identified as a pivotal element with a notable influence on raising the trust and comprehensibility of the AI system.

The practical implications drawn from these findings underscore the importance of establishing a trustworthy model as a means to boost both trust and comprehensibility. Additionally, it highlights the necessity of formulating explanations that are precisely tailored to fine-tune trust levels within the system. The study also underscores the need for more in-depth research to devise ways to present explanations that are comprehensible and contribute to a well-calibrated level of trust, steering clear of excessive reliance on AI systems.

Data Availability

Study materials and anonymized study data will be shared via OSF (link for accessing data: https://osf.io/e46qj/?view_only=dca1bd8882484f2daaf1d4720651d229 (data) and https://osf.io/vkrax/?view_only=260d56f910ce4fdab18ff903ebec8a36 (materials and results)).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the INESC-ID (UIDB/50021/2020), as well as the projects CRAI C628696807-00454142 (IAPMEI/PRR), TAILOR H2020-ICT-48-2020/952215, and HumanE AI Network H2020-ICT-48-2020/952026, and by the LARSyS (UIDP/50009/2020 and LA/P/0083/2020).

References

[1] A. Jacovi, A. Marasovi'c, T. Miller, and Y. Goldberg, "Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI," in *FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 624–635, New York, NY, USA, 2021.

[2] European Commission, Content Directorate-General for Communications Networks, and Technology, *Ethics guidelines for trustworthy AI*, Publications Office, 2019.

[3] L. Kastner, M. Langer, V. Lazar, A. Schomacker, T. Speith, and S. Sterz, "On the relation of trust and explainability: why to engineer for trustworthiness," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pp. 169–175, Notre Dame, IN, USA, 2021.

[4] F. Gille, A. Jobin, and M. Ienca, "What we talk about when we talk about trust: theory of trust for AI in healthcare," *Intelligence-Based Medicine*, vol. 1-2, article 100001, 2020.

[5] M. Jacobs, J. He, M. F. Pradier et al., "Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens," in *CHI'21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, New York, NY, USA, 2021.

[6] A. Ferrario and M. Loi, "How explainability contributes to trust in AI," in *FAccT'22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1457–1466, New York, NY, USA, 2022.

[7] Q. V. Liao and S. S. Sundar, "Designing for responsible trust in AI systems: a communication perspective," in *FAccT'22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1257–1268, New York, NY, USA, 2022.

[8] A. Erlei, F. Nekdem, L. Meub, A. Anand, and U. Gadiraju, "Impact of algorithmic decision making on human behavior: evidence from ultimatum bargaining," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, no. 1, pp. 43–52, 2020.

[9] A. B. Arrieta, N. Diaz-Rodriguez, J. Del Ser et al., "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[10] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[11] P. Hemmer, M. Schemmer, M. Vossing, and N. Kuhl, "Human-AI complementarity in hybrid intelligence systems: a structured literature review," in *Twenty-fifth Pacific Asia Conference on Information Systems*, Dubai, UAE, 2021.

[12] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini, "Meaningful explanations of black box AI decision systems," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 9780–9784, 2019.

[13] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[14] D. Slack, A. Hilgard, S. Singh, and H. Lakkaraju, "Reliable post hoc explanations: modeling uncertainty in explainability," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9391–9404, 2021.

[15] Q. Vera Liao, D. Gruen, and S. Miller, "Questioning the AI: informing design practices for explainable AI user experiences," in *CHI'20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, New York, NY, USA, 2020.

[16] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Applied Sciences*, vol. 3, no. 2, pp. 1–12, 2021.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., pp. 4765–4774, Curran Associates, Inc., 2017.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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, San Francisco, CA, USA, 2016.

[19] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[20] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proceedings of the 36th International Conference on Machine Learning*, pp. 2376–2384, Long Beach, California, USA, 2019.

[21] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and counterfactual explanations for black box decision making," *IEEE Intelligent Systems*, vol. 34, no. 6, pp. 14–23, 2019.

[22] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *FAT'20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, New York, NY, USA, 2020.

[23] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

[24] T. Miller, "Explanation in artificial intelligence: insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

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

[26] P. Pearl and L. Chen, "Trust building with explanation interfaces," in *IUI'06: Proceedings of the 11th International Conference on Intelligent User Interfaces*, pp. 93–100, New York, NY, USA, 2006.

[27] D. Honeycutt, M. Nourani, and E. Ragan, "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*, vol. 8, no. 1, pp. 63–72, 2020.

[28] Z. Ashktorab, C. Dugan, J. Johnson et al., *Effects of communication directionality and AI agent differences in human-AI interaction*, Association for Computing Machinery, New York, NY, USA, 2021.

[29] E. Jussupow, I. Benbasat, and A. Heinzl, "Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion," *28th European Conference on Information Systems - Liberty, Equality, and Fraternity in a Digitizing World, ECIS 2020*, Marrakech, Morocco, 2020https://aisel.aisnet.org/ecis2020_rp/168.

[30] J. Jauernig, M. Uhl, and G. Walkowitz, "People prefer moral discretion to algorithms: algorithm aversion beyond intransparency," *Philosophy & Technology*, vol. 35, p. 2, 2022.

[31] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. W. Vaughan, "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," in *CHI'20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, New York, NY, USA, 2020.

[32] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: the role of mental models in human-AI team performance," *Proceedings of the AAAI*

Conference on Human Computation and Crowdsourcing, vol. 7, no. 1, pp. 2–11, 2019.

[33] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz, "Updates in human-AI teams: understanding and addressing the performance/compatibility tradeoff," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 2429–2437, 2019.

[34] X. Wang, Z. Lu, and M. Yin, "Will you accept the AI recommendation? Predicting human behavior in AI-assisted decision making," in *WWW'22: Proceedings of the ACM Web Conference 2022*, pp. 1697–1708, New York, NY, USA, 2022.

[35] Q. Zhang, M. L. Lee, and S. Carter, "You complete me: human-AI teams and complementary expertise," in *CHI'22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–28, New York, NY, USA, 2022.

[36] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI," *International Journal of Human-Computer Studies*, vol. 146, article 102551, 2021.

[37] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: a review," 2020, http://arxiv.org/abs/2010.10596.

[38] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995.

[39] D. Dua and C. Graff, *UCI Machine Learning Repository*, 2017.

[40] B. F. Malle and D. Ullman, "Chapter 1- A multidimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interaction*, C. S. Nam and J. B. Lyons, Eds., pp. 3–25, Academic Press, 2021.

[41] R. D. B. Duarte, "Towards responsible AI: Developing explanations to increase human-AI collaboration," in *HHAI 2023: Augmenting human intellect*, pp. 470–482, IOS Press, 2023.

[42] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," in *FAT'20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 295–305, New York, NY, USA, 2020.

[43] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: comparing automatically generated explanations," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pp. 109–116, Christchurch, New Zealand, 2016.

[44] H. Vasconcelos, M. Jorke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna, "Explanations can reduce overreliance on ai systems during decision-making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, pp. 1–38, 2023.

[45] K. Z. Gajos and L. Mamykina, "Do people engage cognitively with AI? Impact of AI assistance on incidental learning," in *IUI'22: 27th International Conference on Intelligent User Interfaces*, pp. 794–806, New York, NY, USA, 2022.

[46] M. Schemmer, P. Hemmer, M. Nitsche, N. Kuhl, and M. Vossing, "A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making," in *AIES'22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 617–626, New York, NY, USA, 2022.

[47] M. Naiseh, D. Cemiloglu, D. Al Thani, N. Jiang, and R. Ali, "Explainable recommendations and calibrated trust: two systematic user errors," *Computer*, vol. 54, no. 10, pp. 28–37, 2021.