Explaining the Unexplainable: The Impact of Misleading Explanations on Trust in Unreliable Predictions for Hardly Assessable Tasks

Mersedeh Sadeghi
sadeghi@cs.uni-koeln.de
University of Cologne
Cologne, Germany

Daniel Pöttgen
danielpoettgen@icloud.com
University of Cologne
Cologne, Germany

Patrick Ebel
ebel@uni-leipzig.de
ScaDS.AI, Leipzig University
Leipzig, Germany

Andreas Vogelsang
vogelsang@cs.uni-koeln.de
University of Cologne
Cologne, Germany

ABSTRACT

To increase trust in systems, engineers strive to create explanations that are as accurate as possible. However, if the system's accuracy is compromised, providing explanations for its incorrect behavior may inadvertently lead to misleading explanations. This concern is particularly pertinent when the correctness of the system is difficult for users to judge. In an online survey experiment with 162 participants, we analyze the impact of misleading explanations on users' perceived and demonstrated trust in a system that performs a hardly assessable task in an unreliable manner. Participants who used a system that provided potentially misleading explanations rated their trust significantly higher than participants who saw the system's prediction alone. They also aligned their initial prediction with the system's prediction significantly more often. Our findings underscore the importance of exercising caution when generating explanations, especially in tasks that are inherently difficult to evaluate. The paper and supplementary materials are available at https://doi.org/10.17605/osf.io/azu72

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Software and its engineering** → **Extra-functional properties**.

KEYWORDS

explainability, trust, machine learning, XAI

ACM Reference Format:
Mersedeh Sadeghi, Daniel Pöttgen, Patrick Ebel, and Andreas Vogelsang. 2024. Explaining the Unexplainable: The Impact of Misleading Explanations on Trust in Unreliable Predictions for Hardly Assessable Tasks. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), July 1–4, 2024, Cagliari, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3627043.3659573

UMAP '24, July 1–4, 2024, Cagliari, Italy

ACM ISBN 979-8-4007-0433-8/24/07…$15.00
https://doi.org/10.1145/3627043.3659573

1 INTRODUCTION

The explainability-trust hypothesis [36] claims that "Explainability is a suitable means for facilitating trust in a stakeholder." If we better understand how a system produces its outputs and the explanation for a given output fits with our expectations of a good decision, this explanation presents a reason to trust the system. Although this seems intuitively appealing, especially for opaque machine learning (ML) systems (e.g., [1, 22, 47]), recent psychological studies indicate that explanations do not necessarily facilitate trust [14, 15, 37, 53]. Considering the overall landscape, some studies find no statistically significant correlation between explanations and trust [17, 54], while others demonstrate that trust increases with seemingly random or placebo explanations [4, 8, 23, 42]. Empirical evidence supports that explanations aid human cognition in identifying system failures, promoting what is termed calibrated trust [57]. However, a counter-narrative warns against excessive reliance on machine explanations, as users may endorse AI results despite errors [4, 61].

Our research focuses on this critical gap: a lack of consensus on the impact of explanations on trust in intelligent systems, despite extensive scholarly investigation. While the initial contradictions within these mixed findings might appear perplexing, we introduce a theoretical framework that aims to reconcile the apparent disparities, and more importantly, sheds light on some unexplored aspects of the explainability-trust relationship. Specifically, we examine the effect of explanations on user trust as a function of several variables. We believe that the different results observed in different research studies regarding the effect of explanations on trust can be attributed to differences in the experimental settings, namely different values for the following variables.

- **Reliability of the system**: Most ML systems are not perfect and may produce incorrect outputs. Trust is especially important in such situations. Explanations can help users understand whether or not a system performed reliable reasoning to come up with an output. On the other hand, users may not trust a system even if it solves a task perfectly.

- **Genuineness of an explanation**: Also known as the fidelity of an explanation, refers to the accuracy, faithfulness, or truthfulness of the explanation, i.e., the extent to which the explanation reflects the genuine reasoning process of a machine learning (ML) system.
- **Assessability of the task**: If a user can easily differentiate a correct from an incorrect output, an explanation is unlikely to increase trust in the system. However, if a user is uncertain about the correct output, explanations may justify system outputs and, thus, increase trust in the system.
- **Explanation fit**: Explanations can come in different forms. If a user cannot comprehend the explanation or cannot use it to evaluate the system, the explanation might not change their trust in the system.
- **Measurement of trust**: Several studies suggest that perceived and demonstrated trust may not be correlated [40, 51, 60, 69]. Although users *say* that they trust a system (i.e., perceived trust), they may not *act* accordingly (i.e., demonstrated trust).

Some of these factors have already been identified and controlled in previous experiments in the literature. However, from an engineering point of view, it is most interesting to study the relationship between explanations and trust in a context where:

(1) A system is unreliable: It may produce incorrect outputs.
(2) An explanation is misleading: It is designed in a way that best justifies the output of a system (regardless of whether the output is correct or not).
(3) A task is hardy-assessable: A user cannot easily differentiate between a correct and an incorrect output of a performed task. Nevertheless, the cognitive complexity of a task for humans differs from its level of assessability. For instance, solving mazes may pose high cognitive demands, yet verifying the correctness of a solution remains relatively straightforward for humans. On the other hand, optimizing moves during a chess game, particularly for non-expert players, can introduce both complexity and hard assessability.

While points (1) and (2) have been individually examined in previous studies, they become especially interesting in a combined exploration, particularly when considered alongside point (3). This scenario addresses the challenge of providing faithful explanations due to opaque models—instead, explanations based on surrogate models are employed to justify the system's reasoning. Simultaneously, we deal with a system that may produce incorrect results (i.e., an unreliable system), in situations where human users find it challenging to evaluate the correctness of the system's output (i.e., a hardly assessable task)

In this paper, we investigate an unreliable system (i.e., a system with possibly incorrect predictions) solving hardly assessable tasks. A variant of the system provides additional explanations that are supposed to substantiate the system's predictions. Consequently, the explanations are misleading since they justify the system's output without knowing whether it is correct. In an online survey experiment with 162 participants, we analyze the impact of such misleading explanations on the perceived and demonstrated trust of users. We investigate the following research questions:

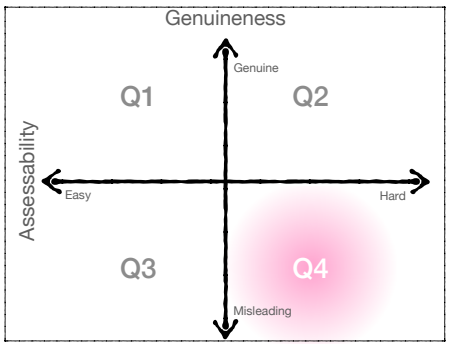

Figure 1: Assessability-Genuineness Grid

RQ1: Do misleading explanations in hardly assessable tasks increase the probability that users align their prediction with that of an intelligent system?
RQ2: Do misleading explanations in hardly assessable tasks increase the confidence of user predictions?
RQ3: Do misleading explanations in hardly assessable tasks increase users' perceived trust?

Although our (artificial) system was incorrect in 50% of the cases, participants using the variant with misleading explanations (i.e., the treatment group) aligned their initial prediction significantly more often with the prediction of the system, were significantly more confident about their final prediction, and rated trust in the system significantly higher when compared to participants of the control group who did not receive any explanations.

Based on the results, we conclude that explanations, even if they explain incorrect behavior, can induce unwarranted trust. In our experiment, a loose relation of the explanations to some underlying data was enough to create this effect. Especially with the rise of conversational AI technologies, we assume that engineers will be able to provide even more convincing explanations in the future. Our findings underscore the importance of exercising caution when generating explanations, particularly in cases involving tasks that are inherently challenging to assess. Providing explanations in such contexts may inadvertently promote overreliance and engender unwarranted trust in the system.

2 BACKGROUND AND RELATED WORK

More recently, there has been a growing integration and utilization of AI and other highly complex systems across various domains [2, 13, 28, 56, 63]. This urged to raise the question "Can machines be trusted?" [31, 51]. Amidst this context, the field of *Explainability* has emerged as a means to enhance the comprehensibility of AI and other complex systems for human users [11, 12, 21]. Notably, various application domains have already begun to incorporate forms of explainability within their complex systems [22, 29, 58, 64]. However, as discussed in section 1 we argue that the diversity of outcomes in the trust-explanation relationship stems from variations in multiple factors. In particular, our focus is to investigate the concept of trust in situations where users are confronted with potentially misleading explanations in the context of unreliable systems especially when performing a hardly assessable

task. In this direction, as shown in Figure 1, we can delineate four distinct quadrants based on the levels of Assessibility and Genuity of tasks and explanations. We posit that the diversity of outcomes observed in previous studies could potentially be attributed to the specific zone within which each study falls. Furthermore, we emphasize the significant importance of Quadrant 4 (Q4), which has largely remained overlooked in the existing literature, and is the main focus of this work.

Quadrant 1 and 3 (Q1,3): In situations where users can easily assess the accuracy of a machine's decision or prediction, trust predominantly hinges on the system's overall performance. This theory helps to elucidate the results of studies such as [54, 69] belonging to **Q1**, wherein the effect of explanations on trust is found to be negligible[1]. A plausible assumption is that individuals tend to bypass explanations when the judgment process is straightforward and they perceive the system to work effectively[59]. This rationale finds support in the study conducted by Bansal et al. [4] where participants claimed that they mostly ignored AI in easily assessable tasks (the sentiment analysis) with up to tripled ratio with respect to more hardly assessable tasks in that study (the LSAT question answering). Furthermore, this quadrant effectively accounts for the perplexing outcomes observed in studies that indicated that explanations could potentially diminish trust [4, 57, 61, 62, 65]. In fact, for tasks characterized by easy assessability, explanations could help users identify errors within the model, consequently mitigating general trust levels or excessive reliance. For example, when the explanation revealed to the participant that the correct classifications of the wolf pictures are only due to snow in the background [57], their trust in the system decreased (which, in turn, led to increased so-called calibrated trust) [25]. This behavior can also be expounded through the lens of cost-benefit frameworks from behavioral economics [38], as adopted by Vasconcelos et al. [65] to elucidate how individuals interact with AI predictions and explanations. They contend that overreliance materializes when the cognitive costs associated with processing an explanation and executing an AI task are nearly equal and both entail non-trivial costs. Quadrants Q1 and Q3 reside within a cost-related space where, due to the cognitive simplicity of the AI task, the processing task for explanations is either equally effortless or less demanding. In such circumstances, a prudent human strategy would likely involve disregarding the explanation [8, 65].

Currently, there are a limited number of studies within the literature that correspond to **Q3**. Nevertheless, the arguments discussed above remain applicable in this context, postulating that the impact of explanations on trust in Q3 would be akin to that of Q1, particularly when the system is performing well. However, Q3 introduces a different dynamic compared to Q1; a potential vulnerability surfaces when the system generates incorrect outcomes. In such cases, a misleading explanation might possess the capability to rationalize these erroneous behaviors. It may lead users to experience a diminished loss of trust than initially anticipated despite recognizing the system's inaccuracies. This phenomenon is exemplified in the findings of the study by Chu et al. [17], where they observed that faulty explanations did not significantly diminish trust. This suggests that users might be more forgiving of a system's perceptible failures if the explanations provided are skillfully persuasive.

Quadrant 2 and 4 (Q2,4): **Q2** introduces a realm of uncertainty, encompassing tasks carried out by intelligent systems that inherently possess a high degree of uncertainty, making it challenging for human users to attain definitive assurance regarding the outcomes. Numerous studies in the literature are located within this quadrant [9, 37, 39, 42, 45], examining the relationship between explanations and trust. These studies often involve experiments revolving around prediction and recommendation tasks, where users (and the machine itself) invariably grapple with a certain degree of uncertainty in the results. In this context, since users cannot directly assess system performance, their trust in the system becomes influenced by additional factors such as explanations. This phenomenon can elucidate the results of numerous studies that have detected a positive correlation between explainability and trust, or even a tendency toward overreliance [54]. For instance, Poursabzi-Sangdeh et al. [54] conducted a study wherein participants exposed to conditions involving transparent models exhibited, on average, less deviation from the models' predictions (composed of correct and wrong predictions) compared to participants in conditions involving black-box models. This unexpected finding contradicted their initial design assumptions, which presumed that participants exposed to a transparent model would be more adept at detecting and rectifying significant errors, compared to those exposed to a black-box model. Though not explicitly articulated in their paper, this finding points to overreliance and implies that participants in the transparent model condition followed the system's prediction regardless of whether such predictions were accurate or erroneous.

Notably, the impact of external variables on trust extends beyond explanations. This is evident in a multitude of studies demonstrating that even exposing the accuracy or confidence level of a machine's prediction can yield a substantial effect on trust formation [9, 42, 68]. In essence, when users are uncertain about system performance due to the hard assessability of a task, they are likely to employ alternative strategies and heuristics to attain a degree of certainty [8, 65]. Research has shown that individuals approach explanations and trust differently, leading to an explanation-trust relationship that is contingent on individual characteristics, expertise, cognitive biases, and automation bias or aversions, etc. [3, 18–20, 35, 44, 46, 60]. For example, Schaffer et al. [60] showcased that the explanation exerted an influence solely on individuals who reported very low familiarity with the task at hand. This underscores the complex interplay between explanation features, user characteristics, and the intricate process of trust formation.

We propose that the effects of external factors are more pronounced within Q2 and Q4 due to the inherently challenging nature of task assessability, which amplifies the uncertainty experienced by users. Nonetheless, there are instances of studies within Q2 that have not reported a significant impact of explanations on trust. This phenomenon can be interpreted in two ways. Firstly, if the level of uncertainty surpasses a certain threshold, users might exhibit a behavior characterized by a refusal to engage in attempting to comprehend the system and its underlying model, consequently leading to the disregard of the provided explanation. Secondly, the

[1]Please note that [54] involved multiple experiments spanning various conditions. we considered the experiment with the condition of "apartment price prediction given only two features " as the one that is easily assessable.

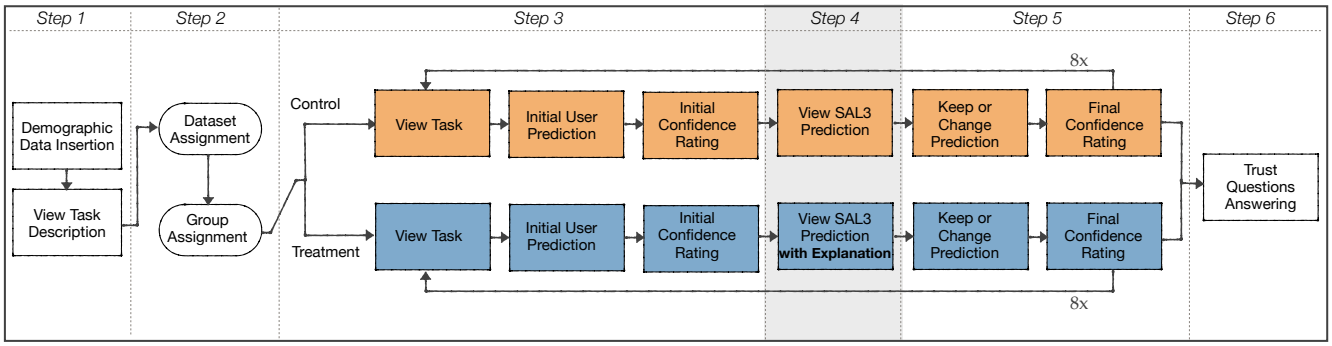

Figure 2: Experiment Procedure

absence of impact observed in certain studies may be attributed to the specific measure of trust employed. Specifically, some studies only report perceived trust without considering the demonstrated trust. Nevertheless, there is a consensus within the research community that perceived trust might not be the most accurate indicator of how users truly rely on explanations [40, 51, 60, 69].

Lastly, **Q4** represents a relatively underexplored area in the literature, where users encounter a scenario in which a task is difficult to assess while presented with a misleading explanation. The arguments established for Q2 remain valid in this quadrant, too: People are more likely to depend on external factors when deciding whether to trust a system due to the difficulty of evaluating the accuracy of the system. However, consequently, in Q4, the risk of encountering a persuasive explanation that inadvertently conceals errors and vulnerabilities of the artificial system is heightened. In this context, a pertinent question arises: How does such a misleading explanation impact the user's trust?

In this study, we undertake a novel exploration by designing an experiment that specifically examines the nuanced relationship between the hardly assessability of a task and the potentially misleading nature of an explanation in shaping the establishment of trust particularly when the system is unreliable.

3 STUDY DESIGN

In our experiment, we introduced an ML-based system named SAL3 (Smart Automatic Labeling System) that can solve binary classification tasks on various datasets. To mitigate the Hawthorne effect [49], we refrained from disclosing the primary focus of the study, i.e., investigating the effect of misleading explanations in the context of hardly-assessable tasks and unreliable systems. Our main goal was to replicate the typical use of ML systems in everyday life. As a result, participants were informed that their involvement was focused on contributing to the manual annotation of a large dataset, which is essential for the future development of SAL3. Furthermore, participants were informed that SAL3 assists them in this process by offering its classification suggestion; however, the participant must make the final decision for each classification task. We adopted a between-subjects design in which participants were randomly assigned to perform classification tasks for two of the three available datasets. This design choice was implemented to avoid the effects of fatigue and to reduce the likelihood of participants dropping out. The participants remained anonymous and our local university regulations did not require a formal ethics review for the study.

All participants completed a total of eight classification tasks, with each dataset consisting of four tasks. Participants were randomly assigned to either the treatment group, in which tasks were presented with the SAL3 prediction accompanied by a misleading explanation, or the control group, in which tasks were presented with only the prediction.

Participants. We recruited a total of 261 participants for this study by utilizing our professional networks as well as various recruitment channels such as a faculty-wide newsletter. We raffled off two 25€ Amazon gift cards among the participants who completed the survey[2]. Following a stringent screening process, we excluded 91 incomplete responses, 6 instances where participants failed an attention check, and 2 surveys completed in less time than the one percent percentile of all participants. Ultimately, our study was conducted with a final sample size of 162 participants.

SAL3. The primary objective of this study is to explore the effect of misleading explanations of ML-based systems in tasks that are difficult to evaluate, with a particular focus on the situation where the system is unreliable. Considering that the accuracy of a system that generates predictions for hardly assessable tasks is difficult to evaluate for users, the performance of the underlying system is inconsequential for this experiment, while the quality of the explanations remains paramount. To generate unreliable predictions and to create a balance between false and correct predictions, we use a random classifier instead of a full-featured ML system. We have further developed an algorithm that generates misleading but plausible explanations for these random predictions.

Procedure. Figure 2 provides an overview of the experimental procedure. First, participants answer demographic questions, review the task description, and are then randomly assigned to three of two datasets and either the *Control* or *Treatment* group. Each user then completes 4 tasks for each of the two datasets. Accordingly, each participant repeats Steps 3 to 5 (see Figure 2) eight times. In Step 3, participants need to solve a classification task based on the provided information and are asked to indicate how confident they

[2]Email addresses were collected in a separate survey with no way to associate email addresses with survey responses.

Table 1: Distribution of SAL3 predictions and outcomes for four cases in each dataset. Each dataset comprises four distinct cases, with the second column representing the predictions generated by the SAL3 model for each case. The selection of cases ensures an equal distribution of positive (low risk, rain, safe water) and negative (high risk, no rain, unsafe water) predictions. Similarly, the third column indicates whether these predictions align with the actual outcomes, maintaining an equal distribution of correct and incorrect predictions. The final column specifies the explanation provided to the participants for each case.

Dataset	Prediction	Correctness	Explanation Type
Stroke	low risk of having a stroke	true	feature-based
	high risk of having a stroke	false	counterfactual
	low risk of having a stroke	true	feature-based
	high risk of having a stroke	false	counterfactual
Rain	rain on the following day	true	counterfactual
	no rain on the following day	true	feature-based
	rain on the following day	false	counterfactual
	no rain on the following day	false	feature-based
Water	water is safe to drink	true	counterfactual
	water is not safe to drink	true	feature-based
	water is safe to drink	false	feature-based
	water is not safe to drink	false	counterfactual

are in their prediction (*initial confidence* measured on a 4-point Likert scale). In Step 4, the control group only sees the SAL3 prediction without explanation whereas the treatment group is also given an explanation for the prediction. In Step 5, participants are reminded of their initial prediction as well as the SAL3 prediction. Following the requirements for evaluating demonstrated trust formulated by Miller [51], users are explicitly asked whether they want to keep or change their initial prediction. Participants are then asked to rate their confidence in their final prediction (*final confidence*). After completing all eight tasks, participants are asked to answer a questionnaire designed to measure their perceived trust.

Datasets. We chose three datasets that allowed us to create difficult classification tasks and replicate a realistic use case for a modern ML system. The use cases are stroke prediction, rainfall prediction, and water quality prediction. We chose these problems to encourage engagement, as people have at least some, though very limited, understanding. We obtained the following datasets from kaggle.com:

- In the **Stroke dataset** each observation consists of ten health-related features and a label that indicates either high or low stroke risk.
- The **Rain dataset** contains about ten years of daily weather data from various locations in Australia. The target variable is "rain tomorrow", which indicates the probability of rain tomorrow based on the current day's data. To not overtax the people, we only considered 9 out of 23 features.
- The **Water dataset** consists of water quality metrics (e.g., the concentration of solids per liter) for 3000 different water sources that indicate potability.

To enhance participant understanding, short and general feature descriptions were included as drop-down buttons for all datasets. This approach allows participants, regardless of domain knowledge, to make more informed choices while maintaining the inherently challenging nature of the tasks.

Case Selection. All participants in the treatment and control groups experienced the tasks in the same order and were presented with the same model predictions (plus an explanation in the case of the treatment group). To ensure that the tasks were hardly assessable we selected them manually. To do so, we randomly sampled 100 cases per dataset and used SAL3 to generate random predictions and misleading explanations (as described in Section 3.1). Afterward, we manually selected four cases from each dataset according to the following criteria:

- We aimed to avoid cases that were too clear-cut, for example, cases in which extreme values gave clear indications for the correct prediction.
- As shown in the *Prediction* column of Table 1, we orchestrated the selection of cases to achieve an even distribution of instances where SAL3 suggests either a positive or negative classifier.
- As represented by the *Correctness* Column of Table 1, we structured the case selection to achieve an unreliable system (50% accuracy), but with a balanced distribution of instances where SAL3's predictions align with or deviate from the actual prediction in the dataset.

3.1 Misleading Explanation Generation

SAL3 can generate two types of explanations for a random prediction for a particular sample: a feature-based explanation and a counterfactual explanation. Both explanation types explain a prediction by relating it to "similar" cases from the training dataset.

As illustrated in Figure 3 (left), the feature-based explanation primarily presents the ratio of similar cases with the same prediction and the features with similar values compared to the sample. However, if the ratio of similar cases falls below a certain threshold, in

SAL3 predicts that this person has **a low risk** of having a stroke because our records show that **86.36%** of other people with similar values for the most influential features listed below are also at a low risk of having a stroke.

Gender: Female
Age: 28 years
Heart disease: No
Ever married: Yes
Residence type: Rural
Avg. glucose level: 83.66 mg/dL
Smoking status: never smoked

SAL3 predicts that it **will rain** on the following day because if the values for the critical features below were different, then the system would have predicted that it will not rain on the next day.

Wind Gust Direction: SSE
Wind Gust Speed: 26 km/h
Rain Today: Yes

Figure 3: An example of a feature-based explanation (left) and a counterfactual explanation (right)

our case, less than 60%, the feature-based explanation might not be very persuasive. In such cases, the algorithm resorts to generating a counterfactual explanation [66], shown in Figure 3 (right), which enumerates the features of the sample that deviate from those of the most similar cases, highlighting that altering the values of these features could potentially lead to a different prediction.

To determine the most similar cases, the algorithm compares the number of similar features between each case in the training data and the features of the sample. Categorical features are considered similar if the values are equal. Numerical features are considered similar if the normalized difference between the values is below a certain threshold. Further details and metrics are provided in Sup. 1. The set of *most similar cases* is then just the set of cases with the largest number of similar features.

We note that neither of these explanations is based on the actual reasoning of SAL3, which is a random classifier that does not consider any feature when making decisions. Even though the information provided in the explanations is not erroneous or fabricated, the explanations are still misleading because they are incomplete to draw a comprehensive understanding. This is because they do not denote the number or proportion of excluded cases (i.e., the count of most similar cases in the entire dataset). Furthermore, the interpretations lack clarity. For instance, in the feature-based explanation, the listed features are attributed as the most influential, however, while a high ratio might imply a correlation between these features and the particular classification, it is not substantiated by conclusive evidence.

3.2 Hypotheses

Given our experiment, we formulate five hypotheses to operationalize and investigate our research questions. In the context of RQ1, our first hypothesis posits that when there is a high uncertainty and limited user knowledge, individuals are more likely to rely on external sources, such as explanations, to form a general perception of the system and determine their trust in it. Accordingly, the presence of a persuasive explanation is expected to be effective in convincing users to accept the system's output, regardless of its accuracy. Therefore, we formulated Hypothesis 1 (H1) as follows.

H1: Participants using a system that provides misleading explanations align their prediction with the system's prediction more often than participants using a system that provides no explanation.

Moreover, within the scope of this research question we are interested in examining how users' confidence in performing a task could potentially interact with the effect of the explanation. Our second hypothesis theorizes that, when users exhibit very low confidence in their own task performance, it indicates that the task appeared more challenging to them compared to users with higher confidence levels. Consequently, those with low confidence are more susceptible to being swayed by a misleading explanation and are more inclined to follow the system's decision. Hypothesis 2 (H2) is articulated as follows.

H2: The effect of misleading explanations on the probability that participants align their prediction with the system's prediction is greater when participants are less confident in their initial prediction.

In RQ2, we aim to examine how a misleading explanation impacts the final confidence of participants when performing tasks that are difficult to assess. We hypothesize that a misleading explanation can boost participants' confidence in their final decisions. We formulate our third hypothesis (H3) as follows.

H3: Participants using a system that provides misleading explanations are more confident in their final prediction than participants using a system that provides no explanation.

We further hypothesize that misleading explanations could amplify the influence of the confirmation bias, the tendency of decision-makers to give more weight to information that confirms a preconceived hypothesis[52, 67]. To answer this question we first need to evaluate if participants suffer from confirmation bias. Therefore, the next hypothesis (H4.1) is expressed as follows.

H4.1: Participants are more confident in their final prediction if the system prediction is equal to their initial prediction.

Accordingly, H4.2 reads:

H4.2: The effect of the confirmation bias is larger for participants using a system that provides misleading explanations than for participants using a system that provides no explanation.

Lastly, in RQ3, we hypothesize that when users face uncertainty due to the challenging nature of the task, the presence of explanations may create an impression of sophistication and proficiency even within a potentially unreliable system. This leads users to

perceive the system as intelligent, even if they cannot fully comprehend its predictions and the accompanying explanations. Consequently, users are more likely to place and express trust in the system. Specifically, Hypothesis 5 (H5) is stated as follows.

H5: Participants using a system that provides misleading explanations indicate that they trust the system more than participants using a system that provides no explanation.

3.3 Statistical Modeling

As stated in section 1, we investigate how misleading explanations in hardly assessable tasks affect users' perceived and demonstrated trust. We adopt the common definition of trust in HCI, which characterizes trust as the extent to which people follow the decisions made by the machine [32, 60, 65, 69]. However, perceived trust, as an attitudinal and subjective measure of trust, typically obtained through self-report measures, may not be the most accurate indicator of real-world reliance on a system [7, 50, 51, 60]. As a result, a growing number of studies have emphasized the importance of demonstrated trust, a behavioral and objective measure that reflects participants' tendency to delegate decision-making to the machine [32, 60, 61, 69]. To consider both perspectives and to answer our research questions as comprehensively as possible, we model both, **Perceived Trust** and **Demonstrated Trust**. To control for participant-level effects due to multiple measurements per participant and the unbalanced study design, we fit mixed-effects models for each measurable outcome of interest to answer RQ1 and RQ2. Mixed-effects models are well suited for testing our hypotheses because they can handle unbalanced designs and account for grouping hierarchies [5, 48].

Modeling Users' Demonstrated Trust. To investigate users' demonstrated trust and to answer **RQ1** and **RQ2**, we consider two different dependent variables, namely *alignment* and *final confidence*. The dichotomous variable *alignment* indicates whether the participant adjusted their initial prediction to match the system's prediction. The ordinal variable *final confidence* represents the final confidence as indicated by the participants after being presented with the system prediction and an explanation (Step 5 in Figure 2). We measure confidence on a 4-point Likert scale (not confident, fairly not confident, fairly confident, confident). To answer **RQ1**, we model the effect of explanations on participants' probability of aligning with the system prediction. To do so, we only consider measurements where the user's initial prediction differed from the system's prediction (n = 473). We then fit a logistic mixed effects model with random intercepts. To investigate **H1**, we include the binary variable *explanation* as a fixed effect and *participant id* and *dataset id* as random effects. To evaluate **H2** we add the ordinal variable *initial confidence* as a fixed effect and allow for interactions. The initial confidence indicates the confidence of a user in their prediction before they were shown the model prediction (Step 3 in Figure 2)and is also measured on a 4-point Likert scale. To answer **RQ2** and evaluate **H3** and **H4.1/4.2**, we model participants' final confidence by fitting a two-way repeated measures ordinal regression model (n = 1296). We include *explanation* and *initial alignment* as fixed effects and allow for interactions. The binary variable *initial alignment* indicates whether the system's prediction matched the user's initial prediction. By including *initial alignment* we measure and account for the potential effect of confirmation bias. We again include *participant id* and *dataset id* as random effects.

Modeling Users' Perceived Trust. To investigate users' perceived trust (**RQ3**) and evaluate **H5** we analyze the answers participants gave to the questions *"I trust the system"* (Q1) and *"I trust the system to perform the rest of the labeling alone."* (Q2). We measure the perceived trust on a 4-point Likert scale (strongly disagree, disagree, agree, strongly agree) and exclude all participants who indicated that they do not know how to assess this question. Since we only measure the perceived trust once for each participant at the end of the experiment, we model *trust* by fitting a one-way ordinal regression model, with *explanation* as the independent variable.

4 RESULTS

We compare participant's behavior across the different conditions by deriving statistical tests from the models introduced in subsection 3.3. The raw and processed data used for modeling are available in Sup. 2 and the analysis code in Sup. 3. We performed all of our analyses using R statistical software (v4.3.1) [55]. We used the `lme4` package (v1.1.31) [5] and the `ordinal` package (v2022.11.16) [16] to build and fit the mixed-effects models. We obtained p-values using the `lmertest` package (v3.1.3) [41] and computed pairwise post hoc tests using the `emmeans` package (v1.8.7) [43]. The regression tables were generated using the `stargazer` package (v5.2.3) [30]. We define statistical significance at the level of $\alpha = 0.05$. Our results are as follows[3]:

H1. Final alignment probability. To answer **H1**, we only consider observations where the initial prediction of the participants differed from the system prediction (N = 473). In line with **H1**, our modeling results show that participants who used a system that

[3]Complete coefficient tables and additional details are provided in the Appendix A

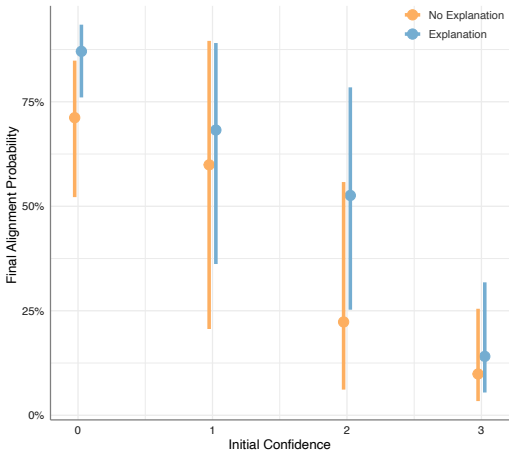

Figure 4: Forest plots of the marginal effects of both interaction terms *initial confidence* and *explanation* on probability for a participant to align their prediction with the system prediction.

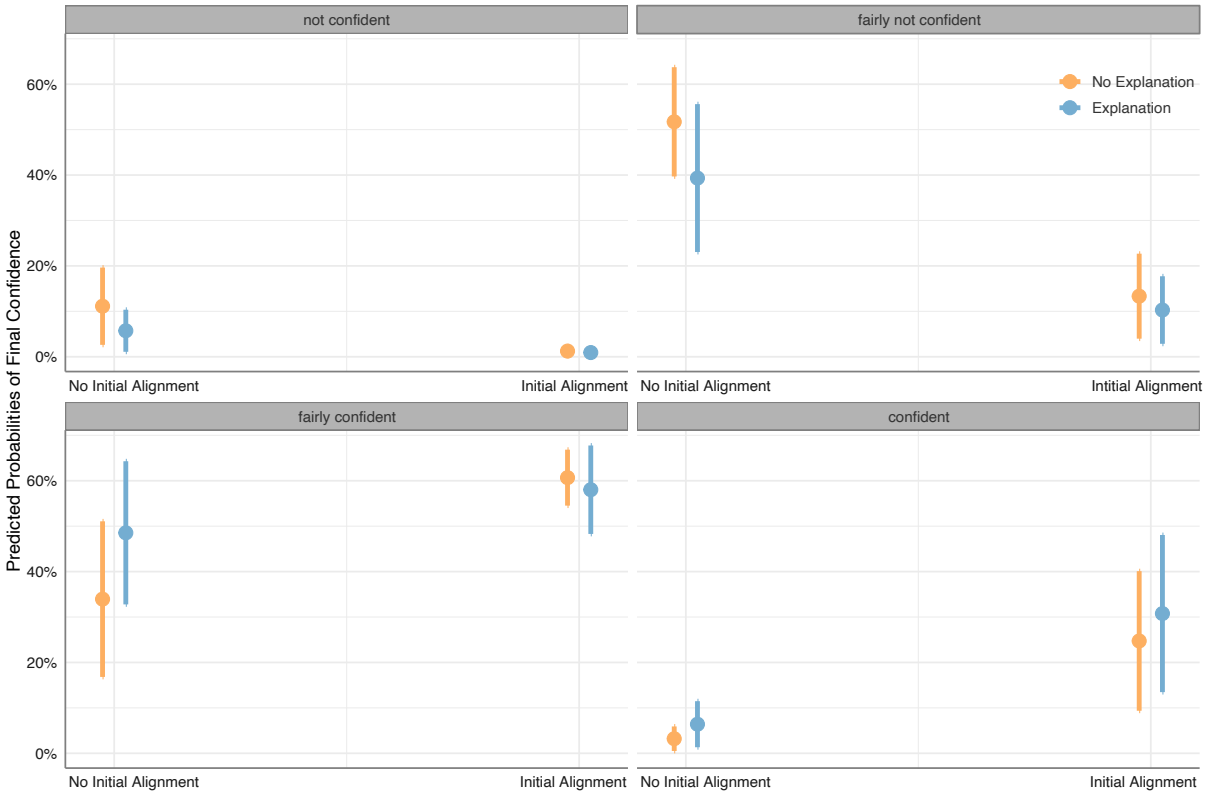

Figure 5: Forest plots that show the marginal effects of both interaction terms *alignment* and *explanation* on the predicted probabilities for four levels of *final confidence*.

provided misleading explanations aligned their prediction with the system's prediction significantly more often than participants who used a system that presented the prediction without an explanation (*Odds Ratio* $(OR) = 2.18$, *Confidence Interval* $(CI) = 1.04{-}4.56$, $p = 0.039$, compare Figure 7 in Appendix A). In Figure 4 we see that the final alignment probability (y-axis) is higher in cases in which user received an explanation (blue) compared to cases in which users did not receive an explanation (orange).

H2. Interaction between explanations and initial confidence. Our results show that participants who are more confident in their initial prediction are less likely to align the final prediction with the system prediction ($OR = 0.09, CI = 0.02 - 0.31, p < 0.001$). For example, participants who were confident (3) in their initial prediction, aligned their final prediction only in ~ 10%/~ 14% (No Explanation/Explanation) of the cases while participants who were not confident (0) aligned their final prediction in ~ 71%/~ 87% of the cases. However, against our hypotheses **H2**, we do not see a significant interaction effect (compare Figure 4) between the explanation and the initial confidence of the user prediction. Thus, there is no evidence that the effect of misleading explanations on participants' alignment behavior is greater when participants are not confident. Thus, we have to reject H2.

H3. Final confidence. In line with our hypothesis, our results (see Figure 6 in Appendix A) show that participants who receive a misleading explanation alongside the system's prediction are more confident in their final prediction compared to participants who receive no explanations ($OR = 2.06, CI = 1.04 - 4.07, p = 0.037$). This overall effect can be observed in Figure 5. Here, each quadrant corresponds to one of the four confidence levels. The forest plots within the quadrants indicate the probability that this final confidence level is predicted by the statistical model, depending on whether the initial alignment and explanation were given or not. Therefore, we see that the probability of predicting high confidence levels (lower left and right quadrants) is greater when an explanation is given (blue forest plots) than when no explanation is given (orange forest plots). This effect is independent of the initial alignment (left and right parts within the quadrants). For low confidence levels (upper left and right quadrants), the effect is reversed.

H4.1/H4.2. Confirmation Bias. Our modeling results (see Figure 6 in Appendix A confirm **H4.1** and show that participants are more confident in their final prediction if the system prediction is equal to their initial prediction ($OR = 9.91, CI = 6.72 - 14.59, p = 0.0017$). The odds ratios and confidence intervals shows that the

effect of the confirmation bias is larger than the effect of the misleading explanations. This can also be observed in Figure 5. In each quadrant, the difference between the left and right (*No Initial Alignment* vs. *Initial Alignment*) forest plots is larger than the difference between the orange and the blue plots (*No Explanation* vs. *Explanation*). However, we found no statistically significant interaction effect between *initial alignment* and *treatment* ($OR = 0.66, CI = 0.40 - 1.07, p = 0.093$), and thus reject **H4.2**.

H5. Perceived Trust. To evaluate **H5**, we excluded all participants who selected the *"Don't know"* option for the questions of how they trust the system. Running a two-sample ordinal regression test on the remaining data from 140 (*Q1*) and 148 (Q2) participants, we found that participants who use a system that provides misleading explanations indicate a higher perceived trust than participants who use a system that does not explain its predictions (Q1: $OR = 2.68, CI = 1.36 - 5.31, p = 0.005$, Q2: $OR = 1.98, CI = 1.05 - 3.72, p = 0.034$). Therefore, the results shown in see Figure 8 in Appendix A support **H5**.

5 DISCUSSION

Our experiment was designed to complement existing literature, focusing on an underexplored context: explaining unreliable predictions for hardly assessable tasks. Overall, our study provides new evidence that supports the explainability-trust hypothesis [36] revealing a significant impact of explanations on both perceived and demonstrated trust. Furthermore, it sheds light on the somewhat mixed findings from previous empirical studies regarding the relationship between trust and explanations. Notably, our results emphasize the critical role of task assessability in determining the likelihood that individuals rely on explanations to build trust, a factor that has been overlooked in prior research.

Concerning demonstrated trust, we show that misleading explanations can be used to convince participants to revise their predictions to match the machine's predictions, even when the system is highly unreliable (50% accuracy in our case). As expected, individuals with lower initial confidence were more likely to revise their predictions to match the machine's prediction. Furthermore, the treatment group, exposed to misleading explanations, consistently displayed a higher likelihood of aligning their predictions irrespective of the initial confidence. This finding shows how vulnerable users are to misleading explanations of ML systems. This overtrust in the system's capabilities can potentially be exploited to manipulate users.

Among individuals with no or low initial confidence (levels 0 or 1), a substantial degree of prediction alignment occurred in both the control and treatment groups. Highly confident participants (level 3) aligned their predictions only in a few cases in both groups (< 15%). However, somewhat unexpectedly, participants with moderately high initial confidence (level 2) were most affected by the explanations when it came to aligning with the system's prediction. In the absence of any explanations, less than 25% of such participants aligned their predictions with the system's prediction. However, when provided with the explanation, this alignment ratio increased to over 50% (see Figure 3). This suggests that misleading explanations have a particularly large effect on people who initially feel quite but not entirely confident.

Another interesting reflection is that creating misleading explanations is not very complicated. Our explanation generation system was a fairly naive algorithm that just explains decisions based on loose references to training data. However, we observed that the uncertainty of participants and their lack of expertise when dealing with hardly assessable tasks made them vulnerable to being influenced by such simple misleading explanations. Given the current progress in AI, we expect even more persuasive explanations in the future that may even increase the explanation bias [6, 10, 24, 26, 27, 33, 34]. This highlights the need for carefully evaluating the effects of explanations.

Besides evidence for the explainability-trust hypothesis, we also observed confirmation bias for hardly assessable tasks. Our results show that the effect of the confirmation bias on users' trust is even larger than that of the explanation bias. Although not statistically significant, we saw a stronger confirmation bias in the treatment group, which may indicate that explanations can increase confirmation bias.

6 CONCLUSIONS

In this paper, we have analyzed the impact of misleading explanations on trust in a system that performs hardly assessable tasks. We argue that this scenario is interesting to study because it reflects the realistic scenario that an ML system is unreliable (i.e., it may produce incorrect predictions) but it generates an explanation that justifies the prediction convincingly. Such misleading explanations are not necessarily the result of unethical behavior of engineers trying to manipulate users. Instead, they are a consequence of using simpler models or reasoning to explain the predictions of complex and opaque models. Our work contributes to the existing knowledge about the impact of explanations on user trust in ML systems. Our study is the first to analyze this impact in the context of an unreliable system providing misleading explanations for a hardly assessable task. Our results indicate that misleading explanations significantly impact demonstrated and perceived trust in a system that is wrong in 50% of the cases. We conclude from these findings that engineers should pay extra attention to when and under which conditions the system provides explanations. For example, explanations may be calibrated according to the system's confidence.

7 OPEN ACCESS

All research artifacts (an overview of the online survey, data processing and analysis code, raw and processed data) and a more detailed description of the explanation generation and survey are available at the following link: https://osf.io/azu72/

REFERENCES

[1] Imran Ahmed, Gwanggil Jeon, and Francesco Piccialli. 2022. From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where. *IEEE Transactions on Industrial Informatics* 18, 8 (Aug. 2022), 5031–5042. https://doi.org/10.1109/tii.2022.3146552

[2] Greg Allen and Taniel Chan. 2017. *Artificial intelligence and national security.* Belfer Center for Science and International Affairs Cambridge, MA.

[3] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. 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. 1–16.

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using **Lme4**. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[6] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[7] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[8] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[9] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[10] Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. *arXiv preprint arXiv:2106.09231* (2021).

[11] Larissa Chazette, Wasja Brunotte, and Timo Speith. 2021. Exploring explainability: a definition, a model, and a knowledge catalogue. In *2021 IEEE 29th international requirements engineering conference (RE)*. IEEE, 197–208.

[12] Larissa Chazette and Kurt Schneider. 2020. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering* 25, 4 (2020), 493–514.

[13] Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial intelligence in education: A review. *Ieee Access* 8 (2020), 75264–75278.

[14] Li Chen, Dongning Yan, and Feng Wang. 2019. User Evaluations on Sentiment-Based Recommendation Explanations. *ACM Trans. Interact. Intell. Syst.* 9, 4 (2019). https://doi.org/10.1145/3282878

[15] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[16] R. H. B. Christensen. 2022. ordinal—Regression Models for Ordinal Data. R package version 2022.11-16. https://CRAN.R-project.org/package=ordinal.

[17] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? A case study in model-in-theloop prediction. *arXiv preprint arXiv:2007.12248* (2020).

[18] Mary Cummings. 2004. Automation bias in intelligent time critical decision support systems. In *AIAA 1st intelligent systems technical conference*.

[19] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[20] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.

[21] Jakob Droste, Verena Klös, Mersedeh Sadeghi, Maike Schwarnmberger, and Timo Speith. 2023. Welcome to the Third International Workshop on Requirements Engineering for Explainable Systems (RE4ES). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*. IEEE, 307–308.

[22] Patrick Ebel, Christoph Lingenfelder, and Andreas Vogelsang. 2023. On the forces of driver distraction: Explainable predictions for the visual demand of in-vehicle touchscreen interactions. *Accident Analysis & Prevention* 183 (April 2023), 106956. https://doi.org/10.1016/j.aap.2023.106956

[23] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.

[24] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462* (2020).

[25] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[26] Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 166–175.

[27] Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. 2023. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clinic Proceedings: Digital Health* 1, 3 (2023), 226–234.

[28] Pavel Hamet and Johanne Tremblay. 2017. Artificial intelligence in medicine. *Metabolism* 69 (2017), S36–S40.

[29] Lars Herbold, Mersedeh Sadeghi, and Andreas Vogelsang. 2024. Generating Context-Aware Contrastive Explanations in Rule-based Systems. *arXiv preprint arXiv:2402.13000* (2024).

[30] Marek Hlavac. 2022. *Stargazer: Well-formatted Regression and Summary Statistics Tables*. Social Policy Institute, Bratislava, Slovakia.

[31] Robert R Hoffman. 2017. A taxonomy of emergent trusting in the human–machine relationship. *Cognitive systems engineering: The future for a changing world* (2017), 137–164.

[32] Aya Hussein, Sondoss Elsawah, and Hussein A Abbass. 2020. Trust mediating reliability–reliance relationship in supervisory control of human–swarm interactions. *Human Factors* 62, 8 (2020), 1237–1248.

[33] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813* (2020).

[34] Samia Kabir, David N Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2023. Who Answers It Better? An In-Depth Analysis of ChatGPT and Stack Overflow Answers to Software Engineering Questions. *arXiv preprint arXiv:2308.02312* (2023).

[35] John Kagel and Peter McGee. 2014. Personality and cooperation in finitely repeated prisoner's dilemma games. *Economics Letters* 124, 2 (2014).

[36] Lena Kästner, Markus Langer, Veronika Lazar, Astrid Schomäcker, Timo Speith, and Sarah Sterz. 2021. On the relation of trust and explainability: Why to engineer for trustworthiness. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE, 169–175.

[37] René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2390–2395. https://doi.org/10.1145/2858036.2858402

[38] Wouter Kool and Matthew Botvinick. 2018. Mental labour. *Nature human behaviour* 2, 12 (2018), 899–908.

[39] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[40] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. 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*. 1–12.

[41] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82, 13 (2017), 1–26. https://doi.org/10.18637/jss.v082.i13

[42] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[43] Russell V. Lenth. 2022. Emmeans: Estimated Marginal Means, Aka Least-Squares Means.

[44] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[45] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.

[46] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.

[47] Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2, 10 (Oct. 2018), 749–760. https://doi.org/10.1038/s41551-018-0304-0

[48] David A. Magezi. 2015. Linear Mixed-Effects Models for within-Participant Psychology Experiments: An Introductory Tutorial and Free, Graphical User Interface (LMMgui). *Frontiers in Psychology* 6 (Jan. 2015), 2. https://doi.org/10.3389/fpsyg.2015.00002

[49] Rob McCarney, James Warner, Steve Iliffe, Robbert van Haselen, Mark Griffin, and Peter Fisher. 2007. The Hawthorne Effect: a randomised, controlled trial. *BMC Medical Research Methodology* 7, 1 (2007). https://doi.org/10.1186/1471-2288-7-30

[50] David Miller, Mishel Johns, Brian Mok, Nikhil Gowda, David Sirkin, Key Lee, and Wendy Ju. 2016. Behavioral measurement of trust in automation: the trust fall. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 1849–1853.

[51] Tim Miller. 2022. Are we measuring trust correctly in explainability, interpretability, and transparency research? *arXiv preprint arXiv:2209.00651* (2022).

[52] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[53] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. arXiv:1907.12652 [cs] http://arxiv.org/abs/1907.12652 Number: arXiv:1907.12652.

[54] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[55] R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[56] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. AI in health and medicine. *Nature medicine* 28, 1 (2022), 31–38.

[57] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.

[58] Mersedeh Sadeghi, Lars Herbold, Max Unterbusch, and Andreas Vogelsang. 2024. SmartEx: A Framework for Generating User-Centric Explanations in Smart Environments. *arXiv preprint arXiv:2402.13024* (2024).

[59] Mersedeh Sadeghi, Verena Klös, and Andreas Vogelsang. 2021. Cases for explainable software systems: Characteristics and examples. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE, 181–187.

[60] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.

[61] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2022. On explanations, fairness, and appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2209.11812* (2022).

[62] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[63] Harry Surden. 2019. Artificial intelligence and law: An overview. *Georgia State University Law Review* 35 (2019), 19–22.

[64] Max Unterbusch, Mersedeh Sadeghi, Jannik Fischbach, Martin Obaidi, and Andreas Vogelsang. 2023. Explanation Needs in App Reviews: Taxonomy and Automated Detection. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*. IEEE, 102–111.

[65] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[66] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[67] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[68] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[69] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. 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*. 295–305.

	final confidence		
Predictors	*Odds Ratios*	*CI*	*p*
not confident | fairly not confident	0.13	0.05 – 0.30	**2.199e-06**
fairly not confident | fairly confident	1.69	0.73 – 3.94	2.221e-01
fairly confident | confident	30.15	12.64 – 71.88	**1.551e-14**
initial alignment [Alignment]	9.91	6.72 – 14.59	**4.409e-31**
treatment [Explanation]	2.06	1.04 – 4.07	**3.695e-02**
initial alignment [Alignment] : treatment [Explanation]	0.66	0.40 – 1.07	9.286e-02
N_{id}	162		
$N_{dataset}$	3		
Observations	1296		
Marginal R^2 / Conditional R^2	0.132 / 0.591		

Figure 6: Two-way repeated measures ordinal regression model for the final confidence.

	final alignment		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.58	0.26 – 1.32	1.974e-01
initial confidence [linear]	0.09	0.02 – 0.31	**1.794e-04**
initial confidence [quadratic]	0.79	0.30 – 2.11	6.437e-01
initial confidence [cubic]	1.51	0.76 – 2.98	2.405e-01
treatment [Explanation]	2.18	1.04 – 4.56	**3.867e-02**
initial confidence [linear] : treatment [Explanation]	0.83	0.18 – 3.88	8.174e-01
initial confidence [quadratic] : treatment [Explanation]	0.86	0.24 – 3.00	8.090e-01
initial confidence [cubic] : treatment [Explanation]	0.45	0.17 – 1.18	1.060e-01
N_{id}	160		
$N_{dataset}$	3		
Observations	473		
Marginal R^2 / Conditional R^2	0.203 / 0.508		

Figure 7: Mixed effects model for the alignment probability.

	perceived trust (Q1)			**perceived trust (Q2)**		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
0|1	0.06	0.02 – 0.14	**9.049e-10**	0.07	0.03 – 0.15	**6.405e-11**
1|2	0.86	0.54 – 1.38	5.355e-01	1.57	0.98 – 2.54	6.213e-02
2|3	26.72	11.80 – 60.50	**3.345e-15**	26.57	11.68 – 60.43	**5.207e-15**
treatment [Explanation]	2.68	1.36 – 5.31	**4.557e-03**	1.98	1.05 – 3.72	**3.386e-02**
Observations	140			148		
R^2 Nagelkerke	0.068			0.035		

Figure 8: One-way Repeated Ordinal Regression for the perceived trust

A APPENDIX