



Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making

Xinru Wang
xinruw@purdue.edu

Purdue University
West Lafayette, Indiana, USA

Ming Yin
mingyin@purdue.edu

Purdue University
West Lafayette, Indiana, USA

ABSTRACT

AI explanations have been increasingly used to help people better utilize AI recommendations in AI-assisted decision making. While AI explanations may change over time due to updates of the AI model, little is known about how these changes may affect people's perceptions and usage of the model. In this paper, we study how varying levels of similarity between the AI explanations before and after a model update affects people's trust in and satisfaction with the AI model. We conduct randomized human-subject experiments on two decision making contexts where people have different levels of domain knowledge. Our results show that changes in AI explanation during the model update do not affect people's tendency to adopt AI recommendations. However, they *may* change people's subjective trust in and satisfaction with the AI model via changing both their perceived model accuracy and perceived consistency of AI explanations with their prior knowledge.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Explainable AI; AI updates; Human-subject experiments

ACM Reference Format:

Xinru Wang and Ming Yin. 2023. Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3544548.3581366>

1 INTRODUCTION

AI-driven decision aids have been increasingly deployed to support human decision making in many activities ranging from making investment choices, to detecting harmful online content, to annotating biomedical images. To help people evaluate the trustworthiness of these decision aids and determine the best strategies to rely on their recommendations, it is critical to provide people with some insights into why the AI model underlying the decision aid makes a particular decision recommendation on a decision making task. To

this end, many explainable AI (XAI) methods have been designed to explain the reasoning processes underneath the black-box algorithmic decisions. For example, post-hoc techniques such as LIME [70] and SHAP [58] have been developed to illustrate the importance of different features to an AI model's final prediction.

In the real life, however, the AI model underlying the decision aid is not always static—it may get *updated* over time. The update of a model can come as a result of different reasons, such as the availability of additional or higher-quality training data, the incorporation of user feedback, the development of more advanced learning algorithms, and the needs to ensure fairness in the model. An increasing number of recent research has started to explore how end-users of an AI model perceive and react to the model as it changes over time. For example, it was found that a good first impression of the AI model is crucial for people to develop trust in the model [64, 84], while those with sufficient domain expertise are capable of dynamically adjusting their trust based on their observations of model performance over time [64]. On the other hand, novice users who have limited knowledge about AI or machine learning may expect the AI model to correct its errors and improve on its own, which reflects their misconceptions of AI models [79]. It was also shown that when the updated AI model has an error boundary that is “incompatible” with the old AI model (i.e., the updated model makes mistakes on cases where the old model used to be correct), users who make decisions with the help of this AI model can suffer from a significant decrease in decision making performance [6].

Beyond the changes in the model's decision recommendations and performance, updates in the AI model can also result in changes in the model's *explanations* for why it makes certain recommendations. For instance, recent studies have reported that when different learning algorithms are used to train a model, their explanations for the model's prediction can be quite different [49, 51]. This means that after an update, it is possible for the AI model's explanations to have a very low level of similarity with the explanations that would have been provided by the old model. While many empirical studies have been carried out to understand the effects of AI explanations on end-users' interactions with a *static* AI model in AI-assisted decision making [7, 19, 52, 56, 63, 87, 91, 95], a natural but currently under-explored question to ask is, how will *changes in the AI explanations* caused by a model update impact end-users' perceptions and usage of the AI model? Obtaining a solid understanding to this question can not only advance our empirical knowledge of people's interactions with an evolving AI model, but also inform the appropriate designs of AI explanations during model updates to ensure a smooth transition of people's mental models of AI and minimize the negative unintended consequences, if any.



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9421-5/23/04.
<https://doi.org/10.1145/3544548.3581366>

Therefore, in this paper, we conduct an experimental study to empirically examine that in AI-assisted decision making, how end-users of the AI-driven decision aid react to changes in AI explanations as the AI model gets updated. Specifically, we ask the following research questions:

- **RQ1:** Can end-users perceive changes in model explanations after a model update?
- **RQ2:** Will the level of similarity between the updated model's explanations and the old model's explanations change end-users' trust in and satisfaction with the AI model?
- **RQ3:** What are the potential mechanisms through which changes in model explanations affect end-users' trust in and satisfaction with an AI model?

Conjecturing the answer to any of these questions turns out to be quite challenging, and an important moderating factor here can be the level of *prior knowledge* the users have in the decision making domain. For example, one may rightfully conjecture that users will be able to perceive the AI explanation changes if the explanations of the updated model is sufficiently dissimilar from the explanations of the old model. However, when users have limited domain knowledge in the decision making tasks, they may have difficulty in making sense of the AI explanations [87] and thus can be less responsive to changes in them. Even if users successfully detect changes in model explanations, how they will impact users' trust in and satisfaction with the AI model is still unclear—competing hypotheses exist and many factors may play a role in mediating this impact. One plausible hypothesis is that human users may not desire a new model explanation that is significantly different from the old one, since that implies a substantial violation to their established mental model of the AI model, potentially leading to a degree of cognitive dissonance [30]. Following this line of thought, one may expect users to decrease their trust in and satisfaction with the AI model as its updated explanations become more dissimilar from the old ones. In contrast, if users generally expect an AI model to improve its performance after the update [79], it is also possible for them to use the similarity between the AI explanations before and after the update as a heuristic to gauge the magnitude of the improvement. In this case, it is reasonable to hypothesize that users may consider an updated AI model with more dissimilar explanations as a “better” model with more improvement, and therefore perceive it as more trustworthy and satisfactory.

To complicate things further, when users have some prior knowledge in the decision making domain, their perceived differences between the AI explanations before and after the model update may not only concern the similarity between the two explanations (i.e., the “size” of the change), but also whether the updated explanations become more or less consistent with their domain knowledge compared to the old ones (i.e., the “direction” of the change). Previous research has shown that for a static AI model, the more its explanations align with the human rationale, the more accurate users perceive the model to be [63]. However, whether similar observations can be made when the AI model gets updated is unknown. For instance, when the AI model's explanations become less aligned with users' domain knowledge after the update, users may consider the updated model as less “reasonable” and indeed decrease their trust in and satisfaction with the updated AI model. Yet, users may

also justify this misalignment simply as that due to the update, the already-trustworthy AI model (because its explanations largely align with human rationale before the update) further uncovers new hidden patterns in the data that they are not previously aware of [76], which may even lead to an increase in their trust in and satisfaction with the updated model.

To answer these questions, we designed and conducted a set of human-subject experiments where participants were recruited from Amazon Mechanical Turk (MTurk) and asked to complete a same sequence of decision making tasks with the help of an AI model. The tasks were divided into two phases and the model was updated between the two phases. All participants used the same AI model and saw the same AI explanations in Phase 1. However, in Phase 2, the AI model was updated in different ways for participants of different treatments, which led to varying levels of similarity between the updated model's explanations and the old model's explanations. To isolate the impacts of AI explanation changes before and after the model update on participants' trust in and satisfaction with the AI model, for participants across all treatments, the decision recommendations they received from the AI model were kept the same for both Phase 1 and Phase 2.

Furthermore, to account for various decision making domains where users may have different levels of prior knowledge in, we conducted two experiments on two different decision making contexts. Our Experiment 1 focuses on a decision making context where laypeople have little domain knowledge in, that is, determining if a mushroom is poisonous. In our Experiment 2, we look into a different decision making context in which laypeople have more domain knowledge—predicting the default risk of loans. In addition, to cover both the cases where the model update results in the explanations to be more or less consistent with users' prior knowledge in the domain¹, we conducted two sub-experiments in Experiment 2—in the first sub-experiment (Experiment 2.1), explanations of the old model presented in Phase 1 were largely *inconsistent* with users' prior knowledge, thus in Phase 2, explanations for updated models with lower similarity to those of the old model were *more* consistent with users' prior knowledge. In contrast, the second sub-experiment (Experiment 2.2) was the opposite—explanations of the old model presented in Phase 1 were largely *consistent* with users' prior knowledge, while in Phase 2, explanations for updated models with lower similarity to those of the old model were *less* consistent with users' prior knowledge.

Our experimental results show that in both experiments, participants can perceive the changes in model explanations after the AI model gets updated. This means that in general, users have some capability to detect explanation changes during the model update regardless of their level of prior knowledge in the decision making domain. In addition, in both experiments, we find no reliable evidence suggesting that the changes in AI explanations during the model update can affect users' *objective* trust in the AI model in terms of how frequently users are willing to adopt the AI model's decision recommendations. However, we find that when users have a degree of prior knowledge in the decision making

¹Here, by “users' prior knowledge”, we mean users' *general common knowledge* about the decision making domain rather than each *individual user's* own knowledge. We obtained users' general common knowledge about the decision making domain through a separate pilot study.

domain, as the AI model gets updated, their *subjective* trust in and satisfaction with the AI model will change with the increased or decreased level of consistency between the new AI explanations and their prior knowledge. This highlights the importance of taking the “compatibility” of human rationale and AI explanations into account when updating AI models to make the model update more “understandable” to end-users, or to help them understand why a “counter-intuitive” model update occurs. Finally, through path analyses, we confirm that the impacts of AI explanation changes on users’ trust in and satisfaction with the AI model during the model update are partially mediated by users’ perceived changes in the AI model’s accuracy, and their perceived changes in the consistency between the AI model’s explanations and their domain knowledge.

Taken together, our findings provide important implications on constructing and communicating AI explanations to human users after upgrading the AI model. Techniques for integrating humans’ domain knowledge into the explanation generation and updating processes, and for supporting people to make sense of the changes in explanations after a model update are both promising directions recommended to explore. We conclude with the discussions of our study implications and limitations (e.g., simplified AI explanations and simplified AI model updates resulting in explicit explanation changes). Despite these limitations, we hope this study can inspire more future work in empirically understanding the impacts of AI explanation updates, and in developing explainable AI methods that better support human-AI joint decision making in a fast-evolving AI development and deployment lifecycle.

2 RELATED WORK

2.1 Overview of AI explanation methods

While the widespread applicability of artificial intelligence (AI) technologies has opened up endless possibilities for real-world impacts, it also poses new challenges—for example, when AI models are used to support human decision making, the lack of explanations on the reasoning processes underlying the AI models can lead to biased and ill-informed decisions. Researchers, government bodies, and the media have advocated that data users should have the “right to explanation” of all decisions made or supported by AI and machine learning algorithms, as stated by the General Data Protection Regulation (GDPR) requirements [69]. To increase the interpretability of AI models, great progresses have been made on the development of a variety of techniques for explaining AI. For example, *global explanations* aim at explaining the behavior of the entire AI model, while *local explanations* provide rationales for specific model predictions [1, 25, 27]. Explanations can also be divided into *model-specific methods* and *model-agnostic methods* depending on whether it is designed for a specific type of model. Model-specific methods often include learning inherently interpretable models such as rule-based models, generalized additive models, decision trees and sets [17, 43, 53, 86], as well as visualizing pixels in images that are most relevant for the predictions given by a deep neural network (e.g., through saliency map) [46, 78, 81, 90]. On the other hand, examples of model-agnostic methods, which are often referred to as *post-hoc* explanations, include global-level feature importance [31], local feature contribution [59, 71], example-based explanation like

prototypes, influential training instances, and counterfactual examples [45, 48, 85], and model distillation [13, 38].

2.2 Changes in AI predictions and explanations after model update

AI models get updated quite often in the real life. This inspires a growing line of recent research investigating into properties of the AI model during updates. Earlier work on AI model update focuses on changes in the model’s *predictions*. For example, some researchers have looked into the problem of analyzing changing trends in continuously learned models [10, 47]. In addition, Bansal et al. [6] explored the changes in a model’s error boundary after the AI model’s update. On the other hand, changes in AI *explanations* after the model update can also be quite common. A few studies have been carried out on analyzing the level of disagreement among AI explanations. For example, Lai et al. [51] compared the agreement level between the feature importance explanations of different machine learning models and different explanation methods in text classification. They found that important features do not always resemble each other better when two models agree on their prediction labels. Another recent work [62] observed that most of the time none of their tested explanation methods agrees with each other by computing rank correlation. Research by Krishna et al. [49] further formalized and quantified how often explanations disagree with each other, and they also studied how such disagreements are being resolved by *practitioners* in machine learning. While all of these studies provide important insights into the magnitude of difference one may expect to see in AI explanations after a model update, how the changes in model explanations will affect the *end-users* of the model, that is, those people who are actually assisted by the AI model in their decision making, remains largely unclear.

2.3 Empirical studies on AI explanations and the dynamics of users’ interactions with AI

Empirical studies on AI explanations. A growing number of empirical studies have been conducted to evaluate how various AI explanations influence people’s perceptions and usage of AI models [7, 14, 16, 19, 52, 56, 67, 87, 88, 91, 95]. These studies look into different aspects of effects of AI explanations, including how they affect people’s understandings of the AI model [19, 67, 87, 88], awareness of AI uncertainty [87, 88, 95], trust in the AI model [14, 65, 95], degree of trust calibration in the AI model [87, 88, 91], and the decision making performance of the human-AI team [7, 16, 52, 56, 67]. Results reported in these studies suggest that effects of AI explanations on people may largely be moderated by factors like the explanation formats [91], the interactivity of the explanations [19, 56], and the meaningfulness of the explanations to human users [63]. Another common moderating factor of the effects of AI explanations is users’ *domain knowledge* in the decision making task [24, 55, 64, 82, 87, 88]. For example, while domain experts were found to be capable of dynamically adjusting their perceived trustworthiness of an AI model given its explanations [64], the provision of explanations may cause lay users who have little domain knowledge to over-rely the AI model [64, 77]. Wang and Yin [87] found that AI explanations are more effective in improving users’ understanding of the

AI model and increasing users' awareness of the uncertainty underlying the AI model's recommendations when people have a higher level of prior knowledge in the decision making domain. It was also suggested that the best explanation modality may differ between domain experts and lay users [82].

Empirical studies on the dynamics of users' interactions with AI models. Many recent empirical works have started to study the dynamics of people's interactions with the AI model over time in AI-assisted decision making. For example, Tolmeijer et al. [84] and Nourani et al. [64] explored how users' trust in an intelligent system evolves as they observe the changing trend of the system's performance over time, and they both highlighted the importance of a good first impression of the intelligent system for user trust to be developed. Other researchers studied how people's trust in an AI model changes as the distributions of the decision making cases shift and the model gets applied to the out-of-distribution data [20, 56]. Bansal et al. [6] explicitly considered the update of AI models over time, and examined how changes in an AI model's error boundary after the model's update affect the joint human-AI team performance in decision making. They showed that when the updates violate people's mental models in terms of their expectations of where the AI recommendations will be right and where they can go wrong, the team performance is significantly decreased.

Another line of empirical research on the updates of AI model over time falls under the umbrella of "human-in-the-loop machine learning" or "interactive machine learning," where users can provide feedback to the AI model to potentially improve its performance [29]—in other words, the update of the AI model is driven by the human users' inputs. These studies often focus on examining how the possibility to provide feedback to an AI model changes users' perceptions of the model. For example, Honeycutt et al. [40] found that the act of providing interactive feedback to improve an AI model may negatively impact user's trust in the model. More recently, researchers have started to study how AI explanations can be used to augment humans' capability in improving the AI model and influence user experience in interactive machine learning [34, 50, 54]. For instance, some explanatory frameworks were proposed to facilitate users' diagnose of model limitations using XAI methods [80, 83]. Smith-Renner et al. [79] further demonstrated that the granularity of user feedback solicited and the provision of AI explanations should be combined appropriately to create a positive user experience and maintain user trust in the AI model.

The remaining research gap. We note that most empirical studies on the effects of AI explanation take a static point of view—the AI explanations tested in these studies are produced for a single version of the AI model. However, in real world scenarios, the development and deployment of AI model is often an iterative process, resulting in frequent AI model updates. It is therefore imperative to take a more realistic, dynamic point of view to re-examine the effects of AI explanations on users' perceptions and usage of the AI model during model updates. Meanwhile, while there have been some recent research on empirically understanding users' interaction dynamics with the AI model over time, the focus of this research is usually on how users get affected by changes in the AI model's *performance* as the model keeps evolving, or how the *provision* of explanations may affect users' impression of the

AI model and ability to improve it. In contrast, knowledge on how *changes in AI explanations* itself during the model update may affect end-users' perceptions and usage of the AI model in AI-assisted decision making is largely lacking. Our study thus aims to fill this gap. As existing studies clearly suggest that users' prior knowledge in the decision making domain will influence the ways that users process the AI explanations, we conduct our study on two different decision making domains with different levels of domain expertise requirements, hoping to provide a more nuanced understanding.

3 EXPERIMENT 1: POISONOUS MUSHROOM PREDICTION

The goal of our study is to empirically understand whether, how, and why changes in AI model explanations due to an update affect end-users' perceptions and usage of the AI model in AI-assisted decision making. We begin our study with a first randomized human-subject experiment on a decision making domain in which people may have limited domain knowledge.

3.1 Experimental Task

In this experiment, we asked participants to complete a sequence of decision making tasks to predict whether a mushroom is poisonous or not, with the help of a decision aid powered by an AI model. Specifically, in each task, participants were asked to review the profile of a mushroom, which consisted of 5 categorical features that describe the mushroom's physical characteristics—the surface texture of the cap of the mushroom, the spacing between the mushroom gills, the shape of the mushroom stalk, the habitat that this mushroom species usually grows on, and the growth habit of a population of this mushroom species. In addition to the mushroom's profile, participants were also presented with a binary prediction given by our AI model in terms of whether the mushroom was predicted to be poisonous, along with the model's *explanations* for its prediction (in the form of the top two features in the mushroom's profile that contribute the most to the AI model's prediction; see more details in Section 3.2). After reviewing all this information, participants were asked to make a decision on whether they believed this mushroom was poisonous or not. The mushroom profiles that we presented to participants were selected from the UCI mushroom dataset [28], which includes 8,124 North American mushroom species described in terms of physical characteristics, with each species identified as either edible or poisonous. In the original dataset, each mushroom species contains 22 categorical features. To simplify the decision making task, we reduced the number of categorical features presented to participants in a profile to five. Figure 1 shows an example of the task interface.


We chose the poisonous mushroom prediction tasks in our Experiment 1 because we speculated that most participants may not have much domain knowledge in this task. As a result, when the AI model as well as its explanations gets updated, participants may only be able to tell whether the updated model explanations are consistent with the old ones (i.e., how similar the model explanations are before and after the update), without having strong feelings about whether the updated explanations become more or less aligned with their *prior knowledge*, or making further judgements on whether the explanation updates are *sensible* or not. Conducting

Task (1/30)

Please review the profile below and predict whether this mushroom is poisonous. If you don't remember the meaning of a feature, click on the red circle on that feature to view its meaning. Click [here](#) to view findings from a large mushroom database.

Profile of this mushroom:

Feature	Value
1. cap surface:	smooth
2. gill spacing:	close
3. stalk shape:	tapering
4. habitat:	grasses
5. population:	scattered



Make Your Prediction:

Do you think this mushroom is poisonous?

☐ Yes, I think this mushroom is poisonous.

☐ No, I think this mushroom is **not** poisonous.

Machine Learning Prediction:

Our machine learning model predicts that this mushroom is poisonous. The two features that contribute the most to the model's prediction is **cap surface (smooth) and gill spacing (close)**.

Make your final prediction:

Now, do you think this mushroom is poisonous?

☒ Yes, I think this mushroom is poisonous.

☐ No, I think this mushroom is **not** poisonous.

[Next](#)

Figure 1: An example of the task interface for the poisonous mushroom prediction task in Experiment 1.

our experiment on this task, thus, allows us to isolate the effects of model explanation updates on people in AI-assisted decision making that are caused directly by the similarity levels between the explanations before and after the model update.

3.2 Experimental Design

3.2.1 Overview of Experimental Treatments. We created three experimental treatments for Experiment 1. Specifically, all participants of Experiment 1 went through a sequence of 30 decision making tasks in the experiment. These 30 tasks were divided into two phases, each containing 15 tasks. In the first 15 tasks (i.e., Phase 1), participants in all treatments saw the same set of 15 mushroom profiles, and they were aided by the same AI model M_0 . Since all subjects were given the predictions produced by the same model M_0 in Phase 1, the model explanations they saw in Phase 1 (i.e., the top two most “important” features for the AI prediction in each task) were also the same. Details on how we developed M_0 and its explanations in Phase 1 are described in Section 3.2.2.

After Phase 1, we explicitly told the participants that the AI model was updated. In the next 15 tasks (i.e., Phase 2), participants in all treatments still saw the same set of mushroom profiles, but participants in different treatments used a different version of the updated AI model (i.e., M_1 , M_2 , or M_3). The tasks in Phase 2 were carefully selected such that different updated AI models still made the *same* binary predictions on each task. However, the explanations of the updated models were different on Phase 2 tasks across the three treatments, and they exhibited varying levels of similarity when compared to the model explanations that would have been provided by the AI model before the update (i.e., M_0). In particular, we had the following three experimental treatments:

- **High similarity (HS):** Participants in this treatment received an updated model M_1 in Phase 2, whose explanations on Phase 2 tasks had a *high* similarity with the explanations that would have been provided by M_0 (i.e., the AI model before the update).
- **Medium similarity (MS):** Participants in this treatment received an updated model M_2 in Phase 2, whose explanations on Phase 2 tasks had a *medium* similarity with the explanations that

would have been provided by M_0 (i.e., the AI model before the update).

- **Low similarity (LS):** Participants in this treatment received an updated model M_3 in Phase 2, whose explanations on Phase 2 tasks had a *low* similarity with the explanations that would have been provided by M_0 (i.e., the AI model before the update).

Details on how we operationalized these three treatments in Phase 2 are described in Section 3.2.3.

3.2.2 Operationalization of Phase 1. We randomly selected 50% of data samples in the original UCI mushroom dataset as the held-out test dataset, and the rest 50% as the training dataset. Using a random subset of the training dataset, we first trained a logistic regression model, which was used as the AI model M_0 in Phase 1. We further adopted the SHAP algorithm [58], which is a model-agnostic explanation method that can be applied to any supervised learning model, to compute the contribution that each of the five features in the mushroom’s profile made to the AI model’s prediction on that task. We then explained the model’s prediction to participants by highlighting on the mushroom’s profile the feature-value pairs for the *top two features* which had the highest contribution scores in the same direction as the AI model’s prediction².

Moreover, the goal of Phase 1 was to help participants establish a mental model of how the AI model makes prediction. Since we used the top two most important features identified by the SHAP algorithm as the model’s explanation on each task, it is natural to expect that participants’ mental model of the AI model’s logic comes as patterns described by if-then rules, e.g., “if $X_1 = a$ and $X_2 = b$, then the model will predict $Y = y$.” Thus, the 15 task instances in Phase 1 were selected so that participants repeatedly observe three explanation patterns as follows:

- **Pattern 1.a:** When “cap surface=fibrous” and “gill spacing=crowded”, the AI model M_0 predicts “edible.”
- **Pattern 1.b:** When “cap surface=smooth” and “gill spacing=close”, the AI model M_0 predicts “poisonous.”
- **Pattern 1.c:** When “stalk shape=enlarging” and “gill spacing=close”, the AI model M_0 predicts “poisonous.”

In other words, we hope that after participants completed the 15 tasks in Phase 1, they could form their mental models of the AI model by memorizing these three explanation patterns. We note that a complete description of the AI model M_0 ’s global behavior on all kinds of task instances requires much more explanation patterns. Here, we selected the task instances to restrict participants’ attention to the above three patterns only and enable them to develop some mental models of the AI model’s local—instead of global—behavior.

3.2.3 Operationalization of Phase 2. The goal of Phase 2 was to have participants in the medium or low similarity treatments realize that their mental models were “broken down.” This means that given a task instance in Phase 2, participants in medium or low similarity treatments might find it to directly relate to their mental model. Thus, they retrieved an if-then rule from their memory and

²In practice, any explainable methods that can identify the most important features to the AI model’s predictions can be used. We decided to choose SHAP as our explanation method because by design, SHAP guarantees local fidelity with the AI model being explained (i.e., the explanation model’s prediction is always the same as the prediction of the AI model being explained) and has high level of internal consistency.

expected that the AI model would predict $Y = y$ on this instance because $X_1 = a$ and $X_2 = b$, but only to find out that while the updated AI model still predicted $Y = y$, the top two feature-value pairs it highlighted as its explanations (which was again computed by the SHAP algorithm) were changed.

To obtain different updated AI models M_1 , M_2 and M_3 , whose explanations on Phase 2 task instances would show different levels of similarity with those of M_0 , we re-sampled the training dataset and re-trained the logistic regression model. For instance, to train M_2 —whose explanations on Phase 2 tasks have a medium similarity with that of M_0 —we re-sampled the training dataset mostly within the set of data samples with the feature-value pair “*cap surface = smooth*” and then re-trained the logistic regression model. By doing so, the updated model M_2 would seldom highlight the feature “*cap surface*” in its explanations (because most data samples in its training dataset had the same value on this feature, making it not informative for the prediction). Thus, given a task instance for which the old model M_0 would use one of the either Pattern 1.a or Pattern 1.b to explain its predictions, the explanation of the updated model M_2 would likely differ on at least one highlighted feature-value pair. Note that this kind of model updates can be realistic in the real world, as the training dataset may constantly get updated [39], yet the additional training data obtained may be biased (e.g., due to sampling biases).

With the updated models prepared, we then move on to select task instances for Phase 2. Given a task instance, we can compute the similarity between two AI models’ explanations on this instance using the feature agreement metric introduced in [49] (i.e., the size of the intersection of the two sets of top- k features divided by k ; $k = 2$ in our study). We carefully selected the 15 tasks in Phase 2 such that on each task:

- (1) all the four AI models’ (i.e., the original model M_0 , and the three updated models M_1 , M_2 , M_3) binary prediction was the same;
- (2) the explanation that would have been provided by the model M_0 is one of the three patterns as shown above;
- (3) compared to the two most important feature-value pairs highlighted by M_0 as its explanations, the explanation given by M_1 in the high similarity treatment was the same (the average feature agreement score between M_0 and M_1 ’s explanations across the 15 tasks in Phase 2 was 1.0), the explanation given by M_2 in the medium similarity treatment usually had one feature-value pair in common (the average feature agreement score between M_0 and M_2 ’s explanations across the 15 tasks in Phase 2 was 0.6), while the explanation given by M_3 in the low similarity treatment usually had no feature-value pair in common (the average feature agreement score between M_0 and M_3 ’s explanations across the 15 tasks in Phase 2 was 0.1)³.

3.3 Experimental Procedure

We posted our experiment as a human intelligence task (HIT) on Amazon Mechanical Turk (MTurk). Upon arrival, participants were randomly assigned to one of the 3 experimental treatments as described in Section 3.2. They first completed a questionnaire on their background, including their demographics, technical literacy, and

expertise in AI and machine learning. Then, we presented participants with an interactive tutorial to explain the task to them and walk them through the interface. Since participants might have little prior knowledge on how to determine if a mushroom is poisonous, we added a training component in the tutorial to help participants get familiar with the mushroom prediction task. In particular, we provided participants with a list of assistive information extracted from the UCI mushroom dataset about how values on the five features of a mushroom’s profile may relate to the mushroom’s poisonous status (e.g., “in a large database, 10% of mushrooms whose gill spacing is crowded are poisonous”). This assistive information was also made available to participants during the actual 30 decision making tasks. Upon completion of the tutorial, participants were asked to answer a few qualification questions to show they understood all the information presented in the tutorial, and they could not proceed to the next part of the experiment unless they answered all the qualification questions correctly.

After passing the qualification, participants started to work on the same set of mushroom prediction tasks that were divided into two phases with 15 tasks each (the order of tasks was randomized within each phase). As discussed earlier, in Phase 1, participants in all three treatments saw exactly the same model prediction and explanations for each task. In contrast, in Phase 2, participants still saw the same model prediction for each task, but the model explanations were associated with different levels of similarity compared with the explanations provided by the old model used in Phase 1. In each task, participants followed a three-step procedure to complete the task. They were first asked to review the profile of the mushroom to make their own prediction. Then, we would present to them the AI model’s prediction along with its explanations. Lastly, the participants needed to make a final prediction. The AI models made correct predictions on 10 tasks in Phase 1 and on 12 tasks in Phase 2, although the participants were not given any accuracy feedback on either their prediction or the model’s prediction throughout the experiment.

Note that between Phase 1 and Phase 2, we explicitly told participants that the AI model was updating and asked them to complete a mid-point questionnaire while waiting for the model update to be completed. To see if the participant successfully formed a mental model of the AI model in Phase 1, we included in the questionnaire three multiple-choice *understanding questions*, each corresponding to one of the three explanation patterns appeared in Phase 1 (e.g., “If a mushroom’s cap surface is smooth and its gill spacing is close, what is our machine learning model’s prediction?”). In addition, the participant was also asked to self-report their subjective trust in and satisfaction with the AI model in Phase 1 on a 7-point Likert scale (1 is the lowest and 7 is the highest), and they also indicated their agreement with the following statement from 1 (“strongly disagree”) to 7 (“strongly agree”):

- **Perceived explanation consistency with prior knowledge:** “The machine learning model’s explanations in Phase 1 agrees with my own knowledge about how to predict poisonous mushroom.”

To make participants feel the update of the AI model was real, after participants completed the mid-point questionnaire, we had them

³See Table A1 in Appendix A for different models’ explanations on the selected 15 Phase 2 task instances in Experiment 1.

wait for 10 more seconds before telling them that the model update was completed and allowing them to proceed to Phase 2.

Finally, after the participant completed Phase 2, they needed to complete an exit questionnaire to again self-report their subjective trust in and satisfaction with the AI model in Phase 2, as well as their perceived consistency of the AI model's explanations in Phase 2 with their own prior knowledge on a 7-point Likert scale. They were also asked to express their agreement with two statements regarding their perceived changes of the AI model after the update, using a scale of 1 ("strongly disagree") to 7 ("strongly agree"):

- **Perceived explanation change:** "After the model update, the updated model in the last 15 tasks utilizes very different features to make predictions compared to the old model shown in the first 15 tasks."
- **Perceived accuracy change:** "The updated machine learning model in the last 15 tasks seems to be more accurate than the old machine learning model in the first 15 tasks."

We included three attention check questions at different places throughout the HIT (one each in Phase 2 prediction tasks, the mid-point questionnaire, and the exit questionnaire). In these questions, participants were instructed to select a pre-specified option as their prediction in the task or as their response to a 7-point Likert question in the questionnaire. These attention check questions later helped us to filter out the data from inattentive participants. Our experiment was open to U.S. workers only, and each worker was allowed to participate only once. The base payment of the experiment was \$1.80. To incentivize participants to carefully read about the model's explanation in each task and adjust their trust accordingly, we further provided them with additional performance-contingent bonuses—if the overall accuracy of a participant's final predictions on the 30 tasks was at least 55%, they could earn a bonus of \$0.04 for each of their correct final predictions. Thus, the maximum amount of bonus a participant could earn in this experiment was \$1.20.

3.4 Analysis Methods

3.4.1 Independent Variables. The main independent variable we used in our analysis is the experimental treatment that a participant was assigned to, i.e., the level of similarity between the explanations of the updated AI model that the participant received in Phase 2 and the explanations of the AI model M_0 used in Phase 1.

3.4.2 Dependent Variables. To quantify participants' perceived changes in the model explanations due to the model update, we use their self-reported scores in the exit questionnaire as our dependent variable; the higher the score, the more the participant finds the updated model explanations in Phase 2 to be different from what would have been provided by the old model in Phase 1.

Moreover, to measure the changes in participants' trust in the model due to the model update, we compute their *trust gain* from Phase 1 to Phase 2, for both *objective trust* and *subjective trust*. Participants' objective trust in the model in a phase is computed as the fraction of tasks of that phase in which the participant's final prediction was the same as the model's prediction. Meanwhile, participants' subjective trust in the model in a phase is obtained from their self-reports at the end of that phase. Given a participant's objective trust or subjective trust scores in both phases, their trust gain is then computed as the Phase 2 trust score minus the Phase

1 trust score; the larger the difference, the more the participant increased their trust in the model after the model update.

Finally, to measure the changes in participants' satisfaction of the model due to the model update, we compute their *satisfaction gain* from Phase 1 to Phase 2 as their self-reported satisfaction with the model in Phase 2 in the exit questionnaire minus that reported for Phase 1 in the mid-point questionnaire. Again, the higher the value, the more the participant increased their satisfaction with the model after the model update.

3.4.3 Statistical Methods. We start by examining that after a model update, whether participants can perceive the changes in model explanations (RQ1) and whether the perceived model explanation similarity before and after the update changes participants' trust in and satisfaction with the AI model (RQ2). To avoid multiple comparison problems and control false discovery, we conduct our analyses using the interval estimate method [26]. That is, we first visualize our data by plotting the mean values of the dependent variables of interest for each treatment along with the 95% bootstrap confidence intervals ($R = 5000$). Then, we construct OLS regression models to predict the dependent variables' values while controlling for covariates (e.g., participants' demographics), both for the entire set of participants and for the subset of participants who had different levels of understandings of how the AI model worked after Phase 1 (e.g., the subsets of participants who answered different numbers of understanding questions correctly in the mid-point questionnaire). Results of these models are interpreted via the estimated coefficient values for the independent variables as well as their 95% bootstrap confidence intervals⁴.

Next, to explore RQ3 (i.e., the mechanism underlying the effects of model explanation updates on end-users' trust in and satisfaction with an AI model), we posit three hypotheses and illustrate our hypothesized model in Figure 2:

- [H1.1] The similarity level of model explanations before and after the model update (i.e., between Phase 1 and 2) has a direct effect on participants' perceived change in the model explanations.
- [H1.2] Participants' perceived change in the model explanations has a direct effect on their perceived change in the AI model's accuracy after the model update.
- [H1.3] After the model update, participants' perceived change in the AI model's accuracy directly affects their objective and subjective trust in the AI model, and their satisfaction with the AI model.

In other words, we hypothesize that the effects of model explanation updates on end-users' trust in and satisfaction with an AI model are mediated by their perceived similarity between the explanations of the updated model and the old model, and their perceived change in the model's accuracy. Since participants are not likely to have much domain knowledge in the mushroom prediction task, in this experiment, we do not expect the model explanation updates will affect participants' trust in and satisfaction of the AI model through

⁴We applied standardization to the dependent variables and encoded independent variables (IV) using dummy coding, thus the estimated coefficient of an IV could be directly interpreted as the change in dependent variable (in terms of standard deviations) resulted from the corresponding treatment.

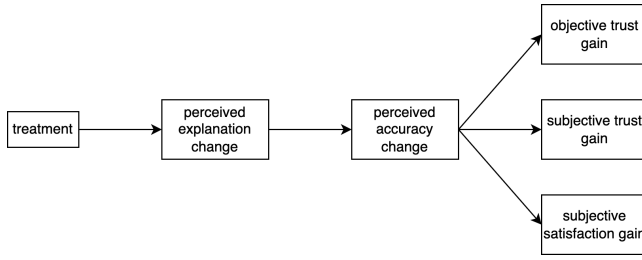


Figure 2: Our hypothesized model of how explanation updates of the AI model affect participants’ trust of and satisfaction with the AI model in Experiment 1, which involves a task domain that participants do not have much domain knowledge in.

influencing their perceived change in the consistency between the model explanations and their domain knowledge.

We perform path analysis [22], a type of structural equation modeling (SEM) [41, 60, 72] without latent variables, to test these hypotheses and explore the potential causal mechanisms underlying the effects of model explanation updates⁵. We use five indicators to evaluate the goodness of fit of the model: (1) the χ^2 test indicating absolute/predictive fit; (2) the Comparative Fit Index (CFI), (3) the Tucker–Lewis Index (TLI) indicating comparative fit, (4) the Root Mean Square Error of Approximation (RMSEA), and (5) the Standardized Root Mean Square Residual (SRMR). A model fits the data well when the p -value associated with the χ^2 test is non-significant, the CFI and TLI values are over 0.90, and the RMSEA and SRMR values are below 0.08 [9, 11, 21].

Since this set of path analysis is mostly meaningful for those people who actually had formed an accurate mental model of how the AI model worked, for **RQ3**, we restrict our analysis only on the data obtained from those participants who correctly answered all three understanding questions in the mid-point questionnaire.

3.5 Experimental Results

In total, 475 participants completed our experiment HIT. The median time participants spent on the experiment was 12.5 minutes, leading to a median hourly wage of \$11.00. After filtering the data from participants who did not pass the attention check, we were left with valid data from 361 participants for Experiment 1 (49.9% male, the average age is 38). We analyze these valid data to answer our research questions. As a sanity check, we first construct an OLS regression model to examine whether there are any differences across the three treatments regarding participants’ perceived changes in how consistent the model explanations are with their own prior knowledge, utilizing their self-reports at the end of Phase 1 and Phase 2. We do not find any reliable differences, which is consistent with our expectation.

3.5.1 RQ1: Effects on perceived explanation change. We start by examining participants’ perceived change of the model explanations between Phase 1 and Phase 2. Figure 3(a) compares across the three

treatments participants’ perceptions of the model explanation’s change (see the “overall” group). To explore whether participants’ understanding of the AI model (or their capability to form an accurate mental model of AI) has any moderating effect, we also present the same comparison separately for participants with relatively low levels of understanding (i.e., who answered no more than 2 understanding questions correctly in the mid-point questionnaire; the “understanding ≤ 2 ” group), and those with high levels of understanding (i.e., who answered all 3 understanding questions correctly; the “understanding=3” group). We find that participants’ perceived change of explanations increased as the explanations of the updated model in Phase 2 became more dissimilar from those of the old model used in Phase 1. In other words, participants in our experiment could perceive the change in model explanations brought up by a model update. Moreover, it appears that the better the participants could form an mental model of the AI model, the more they could perceive the change in model explanations.

We then construct OLS regression models to predict a participant’s perceived change in the model explanations between the two phases while controlling the participant’s demographic background (e.g., age, gender, education), as covariates. Our regression results are consistent with what we have observed in Figure 3(a). In particular, participants in both the medium and low similarity treatments reported higher levels of changes in the model explanations due to the model update (MS: estimated coefficient $\beta = 0.232$, 95% CI=[0.017, 0.461]; LS: $\beta = 0.202$, 95% CI=[-0.025, 0.432]). We further construct two separate regression models for participants who answered no more than 2 understanding questions correctly in the mid-point questionnaire and those who answered all 3 understanding questions correctly, respectively. For the former group of participants (the “understanding ≤ 2 ” group), we do not obtain coefficients that are reliably different from zero, while for the latter group (the “understanding=3” group), we find that they reported a slightly higher level of perceived model explanation change if they were in the low similarity treatment ($\beta = 0.429$, 95% CI=[-0.022, 0.869]).

3.5.2 RQ2: Effects on trust and satisfaction change. We next analyze our data to examine whether people’s trust in and satisfaction with the AI model is influenced by the model explanation updates. Figures 3(b) and 3(c) show participants’ objective trust gain and subjective trust gain in the AI model from Phase 1 to Phase 2, both across all participants and within subgroups of participants with different levels of understanding of the AI model. However, we find that neither participants’ objective trust nor their subjective trust seems to be affected by the similarity level of model explanations between Phase 1 and Phase 2. Figure 3(d) further shows participants’ subjective satisfaction gain from Phase 1 to Phase 2, conditioned on their understanding score. Still, participants did not seem to significantly change their satisfaction with the AI model as the similarity of model explanations before and after the update varied. Our regression models also don’t show any reliable treatment effects either for all participants or for any subsets of participants.

3.5.3 RQ3: Mechanisms underlying the effects of model explanation updates. As discussed earlier, we restrict our attention to the 98 participants who correctly answered all three understanding questions in the mid-point questionnaire, and we test the hypothesized path model on the data obtained from them. We start by

⁵The R package Lavaan [74] is used to estimate the paths in the hypothesized model, which allows simultaneous testing of magnitude as well as significance of the complex predictive relationships between a set of observed variables, and the maximum likelihood estimation (MLE) method is used.

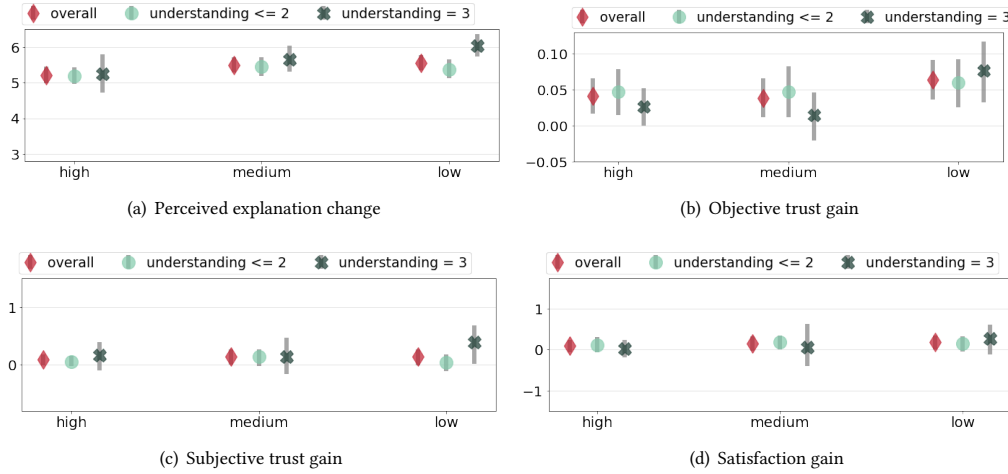


Figure 3: Comparing how the similarity level between the model explanations before and after the update affects participants' perceived change of model explanations, the objective and subjective trust gain in the AI model, as well as the satisfaction gain with the AI model in Experiment 1. Error bars represent 95% bootstrap confidence intervals.

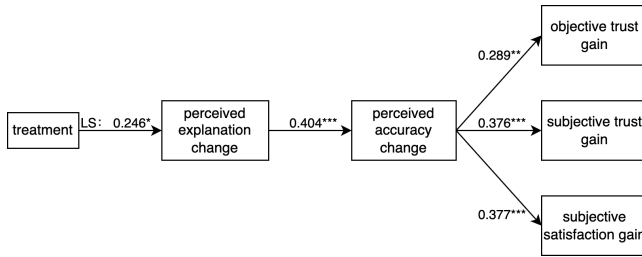


Figure 4: Path analysis results of the proposed model in Experiment 1. Standardized path coefficients are reported, and *, **, * represent significance level of $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively.**

adding all covariates (i.e., the participant's age, gender, education, task familiarity, technical literacy and expertise in AI and machine learning) to the regression models for all paths, and we refine the regression models by pruning covariates with insignificant contributions to achieve better model fit. As a result, the fit statistics for the final model we get are $p(\chi^2) = 0.240$, $CFI = 0.971$, $TLI = 0.932$, $RMSEA = 0.047$, $SRMR = 0.051$, which indicate a good fit. Estimates of the path coefficients and the results of significance testing of the path model are presented in Figure 4.

Our path analysis results validate all of our hypotheses **H1.1–H1.3**. It's shown that the first mediation step of the treatment effects is whether people can perceive the change in model explanations after a model update, and in Experiment 1, we detect that those participants for whom the updated model explanations in Phase 2 had a low similarity with the old model in Phase 1 perceived a significantly larger change in the model explanations. Interestingly, the more people perceive the model explanations have changed, the more likely they feel the updated model's accuracy is increased. Finally, the change in people's trust in and satisfaction with the AI model after the update are all positively affected by their perceived increase in the updated AI model's accuracy. Notably, while all

mediation paths in our path analysis are significant, we do not observe a total effect of the treatment on participants' trust gain or satisfaction gain in Section 3.5.2, which seems to be contradictory. We conjecture that one possible explanation for this observation is that there may exist other competing effects that suppress the path that we have tested in our path analysis, such that multiple direct and indirect effects of opposing direction can result in a near-zero total effect [3, 37, 75]. Identifying the additional mediation paths for the effects of model explanation updates on changes in people's trust in and satisfaction with the AI model will be an interesting future work.

4 EXPERIMENT 2: LOAN DEFAULT PREDICTION

In the second experiment of our study, we move on to examine whether the effects of AI model explanations updates on end-users' perceptions and usage of the AI model in AI-assisted decision making will be different if users have more prior knowledge in the decision making domain. Therefore, in this experiment, we replicate Experiment 1 on a different decision making domain in which people have some domain knowledge.

4.1 Experimental Task

In this experiment, we asked participants to complete a sequence of decision making tasks to predict loan default risks with the help of a decision aid powered by an AI model. We chose the loan default risk prediction task for our second experiment because we conjectured that people might perceive themselves as having a degree of domain knowledge in solving this type of task, because they could apply their day-to-day, common sense knowledge to make their predictions. Specifically, in each task, the participant was presented with the profile of a loan application consisting of six features—the amount and the issued month of the loan, as well as the applicant's annual income, state of living address, credit score, and the month

when the applicant's earliest credit account was open (see Section 4.2.1 for details on how we decided to include these features in the profile of each task). Moreover, we also showed to participants the binary prediction given by an AI model in terms of whether this loan applicant would default on their loan. After reviewing all this information, participants were asked to make a decision on whether they believed this loan applicant will default on the loan or not. Loan applicant profiles that we showed to participants in the experiment were taken from a public dataset that records the loan information of a peer-to-peer lending platform, LendingClub [92]. To simplify the problem as a binary prediction, we restricted our attention only to those cases where the loan applicant either fully paid back the loan or defaulted on the loan. To simplify the task, we also discretized all features with continuous values (e.g., the applicant's annual income) into categories.

4.2 Experimental Design and Procedure

4.2.1 A pilot study to collect people's general knowledge about loan default prediction. In a task domain that people have some domain knowledge in, people's perceptions of and reactions to an AI model's explanation updates may be influenced by their judgements of how "sensible" the explanation updates are. One possible way for them to make these judgements is to compare the model's explanations with their prior knowledge about the decision making task, before and after the model update. So, before we start the design of our Experiment 2, it is critical for us to first obtain an understanding of people's general knowledge about making loan default risk predictions (e.g., what features do people usually consider as informative for making loan default predictions?).

Therefore, we conducted a pilot study to understand in loan default risk prediction tasks, how relevant people considered different pieces of information was for predicting the default risk of a loan applicant. In this pilot study, each participant was asked to complete a sequence of 10 loan default risk prediction tasks, and in each task they reviewed a loan application profile consisted of 13 features selected from the original LendingClub dataset—the loan's amount, issued month, monthly installment, interest rate, purpose, and the number of months to pay off the loan, as well as the applicant's state of living address, annual income, credit score, home ownership status, total number of credit accounts, the number of years employed, and the month when their earliest credit account was open. We then asked the participant to indicate how relevant they thought each feature was for determining a loan applicant's likelihood of defaulting on a loan, in three different ways:

- (1) **Multiple-choice:** Assign each feature into one of the three categories: 1 (irrelevant), 2 (not sure), or 3 (relevant)
- (2) **Ranking:** Rank the relevance of all features from most relevant to least relevant
- (3) **Likert-scale:** Rate each feature on a 10-point scale from 1 (not relevant at all) to 10 (extremely relevant)

In total, we collected survey responses from 184 MTurk workers and then aggregated their responses. For Questions (1) and (3), we ranked all features based on the mean ratings participants reported for them. For Question (2), we used the majority aggregator [73] and Kemeny-Young aggregator [8, 44, 94] to aggregate all the rankings. Based on the 4 aggregated rankings of features that we

obtained from different questions or different aggregation methods, we identified the sets of features that were *consistently* considered by our participants as most or least relevant for predicting the loan default risk (i.e., consistently appear at the top or bottom of the 4 aggregated rankings)—the issued month of the loan, the applicant's state of address, and the month of the applicant's earliest credit account were consistently considered as *least* relevant for predicting loan default risk, while the loan amount, the applicant's annual income and credit score were consistently deemed as *most* relevant for the prediction. We thus included these 6 features in the final loan application profile of each task in Experiment 2, and we leveraged the differences in people's perceived relevance of different features to design our experiment.

4.2.2 Experimental Treatments. Similar as that in Experiment 1, we again created experimental treatments by varying the level of similarity between the AI model's explanations before and after the model update. However, for decision making tasks that people have some domain knowledge in, depending on how much people consider the explanations of the AI model *before* the update align with their prior knowledge in the domain (i.e., how "sensible" the explanations before the update are), a more dissimilar model explanation after the model update could imply either increased or decreased level of consistency between the AI explanations and people's knowledge.

Therefore, we conducted two sub-experiments in our Experiment 2. In both sub-experiments, we again created three experimental treatments—high similarity (HS), medium similarity (MS), and low similarity (LS). Across the three treatments in the same sub-experiment, participants completed the same set of 30 prediction tasks divided into two phases of 15 tasks each—in Phase 1, participants in all three treatments used the same AI model M_0 (i.e., a logistic regression model); in Phase 2, participants in different treatments used different updated versions of the model M_0 , which made the same binary predictions but provided different explanations. Importantly, in the first sub-experiment (i.e., **Experiment 2.1**), explanations of the AI model shown in Phase 1 largely contradicted with people's general knowledge about loan default risk predictions such that in Phase 2, the lower explanation similarity implied the updated explanations to be more consistent with people's domain knowledge. The second sub-experiment (i.e., **Experiment 2.2**) was exactly the opposite—the AI model's explanations shown in Phase 1 were highly consistent with people's domain knowledge, and in Phase 2, the less similar the updated model's explanations compared to M_0 , the more inconsistent they were with people's domain knowledge.

More specifically, in Phase 1 of Experiment 2.1, the 15 task instances in Phase 1 were carefully selected so that participants repeatedly observed the following three explanation patterns:

- **Pattern 2.1.a:** When "state of address=California" and "month of earliest credit account=August", the AI model M_0 predicts "will not default."
- **Pattern 2.1.b:** When "state of address=California" and "issued month=March", the AI model M_0 predicts "will not default."
- **Pattern 2.1.c:** When "state of address=Alabama" and "issued month=June", the AI model M_0 predicts "will default."

In all these patterns, the features being selected as contributing the most to the AI model M_0 's predictions were all considered as *irrelevant* for the predictions by participants in our pilot study. So, we expected participants of Experiment 2.1 to perceive the explanations of the AI model in Phase 1 to be highly inconsistent with their domain knowledge. Then, for the 15 task instances selected for Phase 2 of Experiment 2.1, the updated AI model's explanations in the high similarity treatment were exactly the same as what would have been provided by M_0 . In contrast, the updated AI model's explanations in the medium (low) similarity treatment often shared one (no) feature in common with what would have been provided by M_0 , while the removed features in the explanations were replaced by other features that people considered to be relevant for the predictions. As a result, the feature agreement scores of the updated explanations and the old explanations were 1.0, 0.5, and 0.07, for high, medium, and low similarity treatments, respectively⁶.

For Experiment 2.2, the explanation patterns participants kept observing in Phase 1 were:

- **Pattern 2.2.a:** When “*credit score=good*” and “*annual income = \$40,000–\$60,000*”, the AI model M_0 predicts “*will not default*.”
- **Pattern 2.2.b:** When “*credit score=good*” and “*loan amount < \$5,000*”, the AI model M_0 predicts “*will not default*.”
- **Pattern 2.2.c:** When “*credit score=fair*” and “*annual income < \$40,000*”, the AI model M_0 predicts “*will default*.”

Here, the features being selected in the explanations were all relevant for the predictions based on our pilot study, so we expected participants of Experiment 2.2 to consider the explanations of the AI model in Phase 1 to be highly consistent with their domain knowledge. Again, on each of the Phase 2 tasks, the number of features in the updated AI model's explanations that were in common with what would have been provided by M_0 was two, roughly one, and roughly zero for the high, medium, and low similarity treatments, respectively—the differences were caused by some or all relevant features included in the old explanations being replaced by the irrelevant ones. Thus, for high, medium, and low similarity treatments, the feature agreement scores of the updated explanations and the old explanations were 1.0, 0.5, and 0.03, respectively⁷.

4.2.3 Experimental Procedure. In Experiment 2, the procedure for both sub-experiments was the same as that for Experiment 1, except for that participants of Experiment 1 were not allowed to take part in Experiment 2 again.

4.3 Analysis Methods

We used the same independent variables and dependent variables as those used in Experiment 1, as well as the statistical methods for **RQ1** and **RQ2** (see Section 3.4). For **RQ3**, we posit a few new hypotheses on how model explanation updates may affect end-users' trust in and satisfaction with an AI model, when they have some prior knowledge in the decision making domain:

- **[H2.1]** The similarity level of model explanations before and after the model update (i.e., between Phase 1 and Phase 2) has

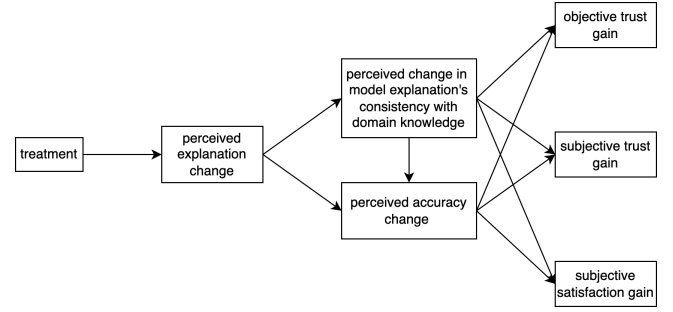


Figure 5: Our hypothesized model of how explanation updates of the AI model affect participants' trust in and satisfaction with the AI model in Experiment 2, which involves a task domain that participants have some domain knowledge in.

a direct effect on participants' perceived change in the model explanations.

- **[H2.2]** Participants' perceived change in the model explanations has a direct effect on their perceived change in how consistent the model explanations are compared to their domain knowledge.
- **[H2.3]** Participants' perceived change in the AI model's accuracy after the model update is affected by both their perceived change in the model explanations, and their perceived change in the model explanation's consistency with their domain knowledge.
- **[H2.4]** After the model update, participants' perceived changes in both the model explanation's consistency with their domain knowledge and the AI model's accuracy jointly affect their objective trust and subjective trust in the AI model, and their satisfaction with the AI model.

The hypothesized model is shown in Figure 5. Compared to the hypothesized model in Experiment 1, here, we conjecture that the effects of model explanation updates is also mediated by participants' perceived change of the model explanation's consistency with their domain knowledge. This mediator is computed as participants' self-reported explanation consistency with their prior knowledge in Phase 2 minus their self-reported rating in Phase 1.

Since we have two sub-experiments in Experiment 2 (i.e., two “groups” of experimental data), we use *multigroup path analysis* [5, 35] to compare the results of them and to test how different types of explanation updates in the two sub-experiments moderate the associations between explanation updates and users' trust in and satisfaction with an AI model⁸. In particular, multigroup path analysis begins with the estimation of two models: A fully unconstrained model in which all parameters are allowed to differ between groups, and a fully constrained model in which the value of each parameter is held the same across groups. If the two estimated models are not significantly different, and the latter fits the data well, it implies that there is no variation in the path coefficients by group. In this case, we will report the output from the fully constrained model. However, if these two estimated models are significantly different, we will go through a series of steps to test which parameters need to be unconstrained across groups by relaxing

⁶See Table A2 in Appendix A for different models' explanations on the selected 15 Phase 2 task instances in Experiment 2.1.

⁷See Table A3 in Appendix A for different models' explanations on the selected 15 Phase 2 task instances in Experiment 2.2.

⁸By specifying the group argument in the R package Lavaan [74], we are able to control whether the estimated path coefficients can vary across multiple groups (i.e., data collected from the two sub-experiments).

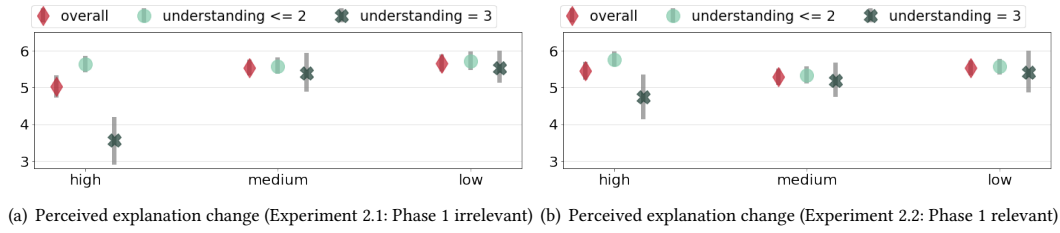


Figure 6: Participants’ perceived change of model explanations between Phases 1 and 2 in the two sub-experiments of Experiment 2. Error bars represent 95% bootstrap confidence intervals.

the constraint on one parameter at a time and testing the difference in fit between the fully constrained model and the partially constrained model (where a single parameter is unconstrained). Generally, the more constrained model will fit the data worse than a partially constrained model. Thus, for each parameter, we use a Chi-squared test to examine if the model fit improves significantly as a result of relaxing the constraint on that parameter—If yes, we assume that parameter is non-invariant (i.e., unequal) across groups; otherwise, we assume that parameter is invariant (i.e., equal) across groups. After identifying non-invariant parameters, we can test a final model where invariant parameters are constrained to be equal across groups while non-invariant parameters are freely estimated.

4.4 Experimental Results

494 participants completed Experiment 2.1, and 507 completed Experiment 2.2⁹. After filtering out inattentive participants, in total, we obtained valid data from 394 (58.4% male, the average age is 37) and 412 participants (55.1% male, the average age is 36) for Experiment 2.1 and 2.2, respectively. We analyze these data to answer our research questions.

To first confirm the validity of our experimental design, we conduct OLS regressions to understand how participants’ perceived changes in the model explanation’s consistency with their prior knowledge vary across treatments within each sub-experiment. Results of these regressions indicate that in both sub-experiments, the direction of participants’ perceived change in the model explanation’s consistency with their prior knowledge aligns with our expectations. For example, we find that in Experiment 2.1, participants with highest understanding score in the low similarity treatment noticed that the updated model explanations were more consistent with their domain knowledge ($\beta = 0.751$, 95% CI=[0.006, 1.484]), while in Experiment 2.2, participants with highest understanding score in the low similarity treatment considered the updated model explanations as less consistent with their domain knowledge ($\beta = -0.700$, 95% CI=[-1.332, -0.043]).

4.4.1 RQ1: Effects on perceived explanation change. Figure 6(a) and Figure 6(b) show participants’ perceived change in the model explanations between Phase 1 and Phase 2 for the two sub-experiments, respectively. Again, we find that in both sub-experiments, participants could sense the differences in the model explanations between Phase 1 and Phase 2, especially those who could develop an accurate

mental model of the AI model’s logic. The regression models reveal similar results: In Experiment 2.1, participants in both medium and low similarity treatments reported significantly higher levels of perceived changes in model explanations (MS: $\beta = 0.285$, 95% CI=[0.057, 0.510]; LS: $\beta = 0.422$, 95% CI=[0.174, 0.688]). Conducting separate regressions on the group of participants who answered no more than 2 understanding questions correctly and the group of participants who answered all 3 understanding questions correctly, we find that the higher levels of perceived changes in the model explanation were only reported by participants in the latter group (MS: $\beta = 0.904$, 95% CI=[0.475, 1.344]; LS: $\beta = 0.980$, 95% CI=[0.613, 1.383]). For participants in Experiment 2.2, we don’t find any reliable main effects of the treatments on people’s perceived change in model explanations across all participants. However, by restricting our attention only to participants who answered all 3 understanding questions correctly, we find those in the low similarity treatment clearly detected higher levels of changes in the model’s explanation after the update ($\beta = 0.483$, 95% CI=[0.048, 0.929]).

4.4.2 RQ2: Effects on trust and satisfaction change. Next, we examine whether participants’ trust in the AI model, as well as their satisfaction with the model, changes with the similarity level between the model explanations before and after the update. Figure 7 shows the comparison results on participants’ objective and subjective trust gain, while Figure 8 compares participants’ satisfaction gain across treatments. Visually, it appears that in both sub-experiments, participants did not change their objective trust in the AI model, regardless of how similar or different the updated model explanations were compared to the old model (Figure 7(a), 7(c)). In contrast, people’s subjective trust in the AI model or subjective satisfaction with the AI model is largely affected by the explanation similarity between the two phases—in Experiment 2.1, when the more dissimilar explanations involve more relevant features and become more consistent with participants’ domain knowledge after the update, both participants’ subjective trust and satisfaction *increased* as the similarity level of model explanations between Phase 1 and Phase 2 decreased (Figure 7(b), 8(a)); while in Experiment 2.2, when the more dissimilar explanations involve more irrelevant features and become less consistent with participants’ domain knowledge after the update, participants’ subjective trust and satisfaction *decreased* as the similarity level of model explanations between Phase 1 and Phase 2 decreased (Figure 7(d), 8(b)). Furthermore, it appears that the treatment effects on participants’ subjective trust and satisfaction changes between the two phases mainly come from participants who answered all 3 understanding questions correctly in the mid-point questionnaire.

⁹In Experiment 2.1, the median time participants spent on the experiment was 12.8 minutes, and the median hourly wage participants earned was \$10.3. In Experiment 2.2, the median completion time and median hourly wage were 11.3 minutes and \$11.9, respectively.

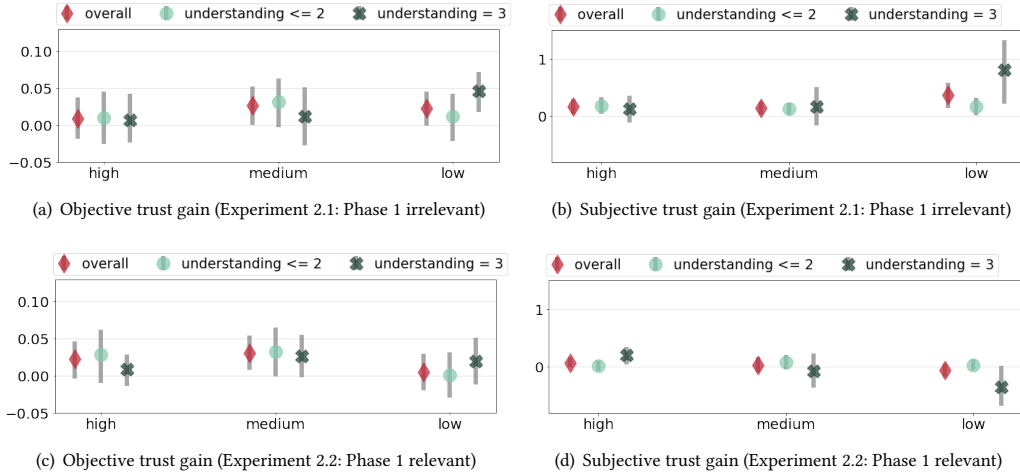


Figure 7: Comparing how the similarity level between the model explanations before and after the update affects participants' objective and subjective trust gain in the AI model, for the two sub-experiments of Experiment 2. Error bars represent 95% bootstrap confidence intervals.

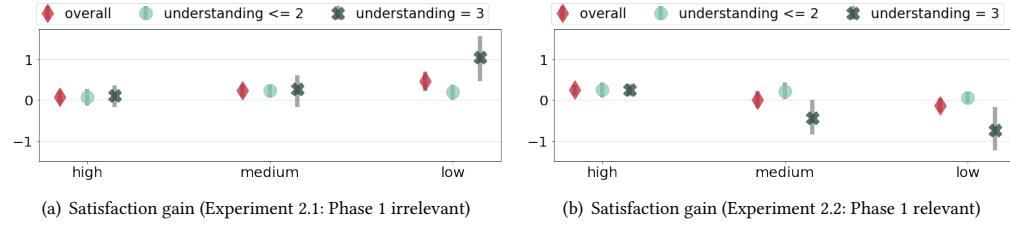


Figure 8: Comparing how the similarity level between the model explanations before and after the update affects participants' subjective satisfaction gain with the AI model, for the two-experiments of Experiment 2. Error bars represent 95% bootstrap confidence intervals.

The OLS regression models reveal consistent results. We find no reliable effects of the explanation similarity on participants' objective trust gain after the model update, even for participants with the highest understanding scores. On the other hand, in Experiment 2.1, participants' subjective trust in the model significantly increased in the LS treatment ($\beta = 0.266$, 95% CI=[-0.004,0.530]), and this effect mainly comes from participants who answered all understanding questions correctly (LS: $\beta = 0.523$, 95% CI=[0.066,0.965]). Similarly, participants' satisfaction with the AI model significantly increased in the LS treatment ($\beta = 0.378$, 95% CI=[0.124,0.631]), and for participants with the highest understanding scores the effect size is even larger ($\beta = 0.685$, 95% CI=[0.215,1.140]). For Experiment 2.2, the signs for all the estimated coefficients of the treatment's effects are reversed. For example, participants' subjective trust and satisfaction in the model both decreased in lower similarity treatments (subjective trust—LS: $\beta = -0.203$, 95% CI=[-0.412,0.015]; subjective satisfaction—MS: $\beta = -0.248$, 95% CI=[-0.474,-0.022]; subjective satisfaction—LS: $\beta = -0.385$, 95% CI=[-0.598,-0.169]). Again, similar but larger effects are found from the subset of participants with the highest understanding scores (subjective trust—LS: $\beta = -0.521$, 95% CI=[-0.901,-0.103]; subjective satisfaction—MS: $\beta = -0.474$, 95% CI=[-0.818,-0.125]; subjective satisfaction—LS: $\beta = -0.608$, 95% CI=[-0.978,-0.200]).

4.4.3 RQ3: Mechanisms underlying the effects of model explanation updates. Lastly, we explore the mechanisms underlying the effects of model explanation updates on end-users' trust in and satisfaction with the AI model when they have some prior knowledge in the decision making domain. Based on our hypothesized path model, we conduct multigroup path analyses on the data obtained from those participants who correctly answered all three understanding questions in the mid-point questionnaire (112 participants and 120 participants in Experiment 2.1 and Experiment 2.2, respectively).

Our multigroup path analyses first suggest that across the two sub-experiments, four effects are detected to be the same, that is, the coefficients on 4 paths are invariant across the two sub-experiments. This includes (1) the effect of the treatment on participants' perceived change in model explanations, (2) the effect of participants' perceived change in the explanation's consistency with their domain knowledge on their objective trust gain, (3) the effect of participants' perceived change in the explanation's consistency with their domain knowledge on their subjective satisfaction gain, and (4) the effect of participants' perceived change in model accuracy on their objective trust gain. As a result, cross-group equality constraints are only imposed across groups for the parameter of each of these 4 paths. The fit statistics we have obtained for the final model are $p(\chi^2) = 0.017$, $CFI = 0.960$, $TLI = 0.930$, $RMSEA = 0.057$, $SRMR = 0.055$. Although the χ^2 value is statistically significant, which may

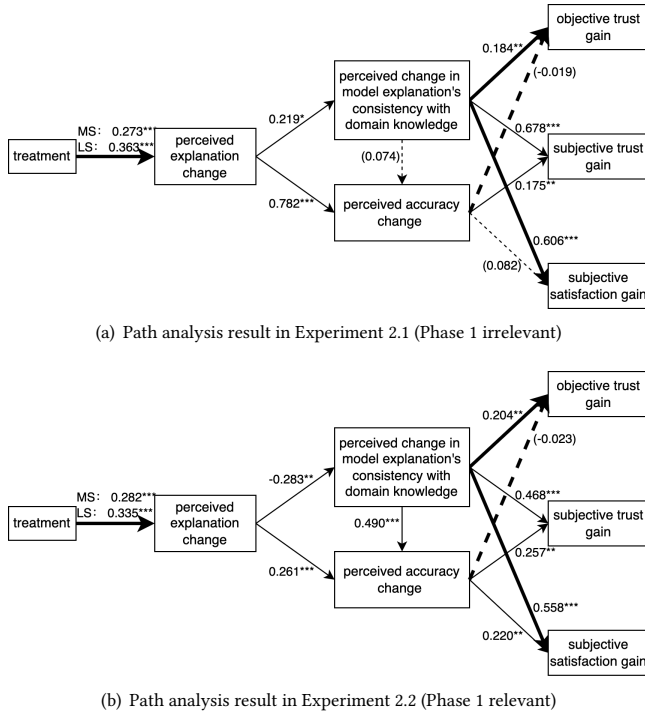


Figure 9: Path analysis results of the proposed model in Experiment 2. Standardized path coefficients are reported, and *, **, * represent significance level of $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively. Dashed lines represent insignificant paths. Bold lines highlight paths with cross-group equality constraints on the coefficients before standardization.**

indicate an inadequate model fit, many existing literature on structural equation modeling has argued that the Chi-squared test of model fit is strongly influenced by sample size, and the null hypothesis of perfect fit in this test may be unrealistic and implausible in most practical work [12, 42]. Thus, our model can still be considered as fitting the data reasonably well as all other tests satisfy the empirical standards. Figure 9 presents the path coefficient estimates and the results of significance testing of the final path model¹⁰.

As shown in Figure 9, our hypothesis that the first mediation step of the treatment effects is whether people can perceive the change in model explanations after an update (i.e., **H2.1**) has been confirmed in both sub-experiments. Then, aligning with our hypotheses **H2.2**, we indeed find that participants' perceived change in model explanations significantly affects their perceived change in the model explanation's consistency with their domain knowledge, and the direction of this impact depends on which sub-experiment they took part in. For our hypothesis **H2.3**, it is partially supported—On the one hand, we find people's perceived change in the model explanations significantly *increases* their perceived accuracy of the AI model after the model update; this is consistent with what we have observed in Experiment 1. On the other hand, for the hypothesized

direct effect of people's perceived change in the model explanation's consistency with their domain knowledge on their perceptions of the AI model's accuracy, we only find it to be significant in Experiment 2.2. Finally, for **H2.4**, we find that people's objective trust in the AI model is only impacted by people's perceived change in the model explanation's consistency with their own knowledge after the model update. However, changes in their subjective trust in and satisfaction with the AI model are positively affected by both people's perceived change in the model explanation's consistency with their domain knowledge and their perceived change in the model's accuracy after the model update, although the magnitude of effects for the former is consistently larger. Putting all intermediate effects together, the opposite directions of the subjective trust and satisfaction change that we observe in the two sub-experiments are mainly caused by the opposite sign of the coefficient of the causal path from people's perceived change in explanation to their perceived change in how consistent the explanations are compared to their prior knowledge.

5 DISCUSSIONS

In this section, we provide further discussions of our results as well as their implications, and discuss the limitations and future work.

5.1 The role of domain knowledge

Comparing the results we have obtained from the two experiments, we indeed find that the effects of model explanation updates on end-users' trust in and satisfaction with the AI model are moderated by the level of domain knowledge people have in the decision making domain. While how much users are willing to accept the AI model's recommendations (i.e., people's objective trust in the AI model) is not significantly affected by the AI explanation updates regardless of their prior knowledge level, their subjective feelings of the AI model (e.g., subjective trust and satisfaction) are affected by the AI explanation updates when they have some prior knowledge in the task domain. In fact, as shown in Figure 9, people's perceived change in the explanation's consistency with their domain knowledge largely dominates their perceived change in the model accuracy in influencing their trust in and satisfaction with the AI model after the update. Similarly, if we compare the standardized path coefficients estimated for the effects of people's perceived change in the model accuracy on the changes in their trust and satisfaction between Figure 4 and Figure 9, we can also see those in Figure 9 are consistently smaller, indicating decreased impacts for people's perceived change in the model accuracy when people have domain knowledge in the tasks. All of these highlight the key role that users' prior knowledge in a domain plays when they observe explanation updates in an AI model.

One possible explanation for the different results that we see in the two experiments is that without additional information, people may only be able to *make sense of* the feature contribution explanations if they have some domain knowledge about the task. For example, for participants working on the poisonous mushroom prediction task, while they might notice the change in model explanations after a model update, they might not be able to judge whether the new patterns utilized by the model were more or less meaningful; so, they simply reacted to different AI explanations

¹⁰Since variances are likely unequal among groups, even if the raw value of a path coefficient is constrained to be a same value across the two sub-experiments, the standardized coefficients are computed on a per group basis and can be unequal.

similarly. On the contrary, participants performing the loan default prediction task might find it rather straightforward to apply their prior knowledge (i.e., a proxy/heuristic of what is meaningful) and focus more on analyzing how consistent the AI explanation was with their knowledge, when evaluating the quality of the updated explanations [63]. This highlights the importance of helping end-users to make sense of explanations when they have limited prior knowledge in the task domain. To this end, one promising direction is to supplement explanations of AI models with explanations of the underlying *data* [4], which in effect may help people establish some “knowledge” or data-driven insights about the domain. Moreover, our findings on how users adjust their subjective trust in and satisfaction with the updated AI model when they have some knowledge in the task domain is largely consistent with what we would expect from users’ reactions to explanations of a static AI model. This implies that without additional information, users are unlikely to interpret “human-meaningless” explanations as revealing novel insights even in the context of AI models getting updated.

5.2 On people’s perceived change in model accuracy after a model update

An interesting finding we consistently see from both experiments is that the dissimilarity level between the model explanation before and after the update positively affect people’s perceived accuracy increase of the model. As discussed earlier, we conjecture that this may be resulted from a combination of two factors. First, people may use the similarity level between the model explanations as a heuristic to gauge how different the two models’ accuracy is, and they associate less similar model explanations with larger differences in model accuracy. Second, people may have a biased belief/misconception that a model update will always result in a “better” model, due to their day-to-day experience (e.g., the newer generation of a product is always advertised as having improved performance). Thus, people may consider updated models with less similar explanations as having a larger accuracy improvement.

Another interesting observation is that in those cases where people have some domain knowledge (i.e., Experiment 2), while we hypothesize that the similarity level between the model explanations before and after the update will indirectly affect people’s perceptions of the model accuracy through their perceptions of the explanation’s consistency with their prior knowledge, our results show that this is not always the case—we only observe this indirect effect in Experiment 2.2 when the update results in a decrease of consistency between the model’s explanations and people’s prior knowledge. In fact, in Experiment 2.1, the correlation between people’s perceived change in model accuracy and their perceived change in the model explanation’s consistency with their domain knowledge is quite weak (Pearson’s $r = 0.135$). We speculate that this asymmetric effect is observed because the model explanation update in Experiment 2.1 naturally aligned with people’s expectations, while the explanation update in Experiment 2.2 did not. In other words, most people might believe that the updated model should utilize more information that they (i.e., humans) consider as “predictive” to make decisions. Therefore, participants might get “shocked” by the insensible updates that they saw in Experiment 2.2, so that such violation of expectation became a key driver of

the decrease in their perceived model accuracy. On the other hand, participants in Experiment 2.1 might perceive the updated model explanation as simply meeting their expectation without giving extra credit to the updated model’s performance.

5.3 Implications for designing AI explanations during updates

Our findings imply a few important implications for designing effective AI explanations during the model update. First, as we find that people’s subjective trust in and satisfaction with the AI model during the model update can largely be influenced by the consistency of the AI explanations with their domain knowledge, novel methods should be developed for incorporating human expertise into the model development/updating process or the explanation generation process. This is closely connected to the line of research on human-in-the-loop machine learning [29], in which feedback is solicited from humans to improve and update the AI model. Indeed, as shown by many previous studies [18, 32, 33, 66], integrating expert knowledge into AI models may not only enhance the robustness and trustworthiness of the models, but also satisfy the expectations of users for expert-informed and user-centric explanations.

However, it is also possible that people may inappropriately decrease their trust in and satisfaction with an AI model because the updated AI explanations contain some novel and truly meaningful patterns which people are not aware of themselves. Indeed, one of the greatest promises of AI technologies is their strong capabilities in processing huge amounts of data to automatically identify hidden patterns and to generate data-driven insights. To avoid these undesirable scenarios, after a model update, instead of simply presenting the updated model explanations, it may be helpful to put more emphasis on the components of the explanations that have been changed, and provide more insights into *why* these changes occur. Compared to plainly explaining the updated model’s prediction, highlighting the changes in the explanation may attract user’s attention to the updated part of the explanation. Additional information on why explanation changes occur may enable people to go beyond their potentially limited domain knowledge in evaluating the “utility” of the changes, supporting them to better calibrate their perceptions of the updated model’s trustworthiness.

5.4 Limitations and future work

Our study have a few limitations. First, we adopted a relatively simplified setting in our experiment to study how changes in AI explanations during the model update affect users’ perceptions and usage of the AI model—the explanation used is simple (i.e., the top-2 important features), the task instances are selected to have participants repeatedly observe the AI model’s behavior in the same local area, and the experimental treatments are designed with rather salient changes in model explanations after the AI model gets updated. We acknowledge that in the real world, the explanations of an AI model can be much more complex—especially when trying to explain an AI model’s *global* behavior—and the model updates may have low chance of resulting in fundamentally different explanation patterns. However, we believe the study we conducted on the simplified setting had two important advantages and provided a starting point for more future research along this line. First, by

using simple explanations and restricting participants' attention to the AI model's local behavior, we maximized the possibility for participants to successfully form a mental model of the AI model before the update. This is critical because it allowed us to rule out the possibility that any null result of our study is simply caused by participants' inability to understand how the AI model works before the update. Second, by having participants in some treatment (e.g., the low similarity treatment) observe very distinct explanations after the model update, we pushed our experimental manipulations to the extreme to maximize their possible effects, if any. In this sense, one can argue that the empirical effects of model explanation changes that we found in this study are likely the *upper-bound estimates*. These upper-bound estimate results can still be quite informative. For example, even in our setting where the changes of model explanations are very salient, we did not find users' objective trust in the AI model is reliably affected by explanation changes during the model update. This may imply that in a more practical setting where the explanation changes during the model update are much more subtle, users' objective trust in the AI model is also unlikely to be influenced.

Another limitation of our study is the choice of some measurements. For example, we used "agreement fraction" (i.e., the chance for participants' final prediction in a task to agree with AI) to quantify participants' objective trust in the AI model. Although widely used in the literature [7, 15, 23, 52, 55, 57, 68, 95], we acknowledge that this metric may reflect the natural agreement between people's independent decisions and the AI recommendations to some extent. In practice, agreement fraction is often the only metric that can be adopted to objectively quantify people's trusting behavior when no information about people's own independent decision is available. In our study, however, we collected participants' initial prediction in each task, which allowed us to quantify participants' objective trust in the AI model using "switch fraction" (i.e., the fraction of tasks for which the participant's final prediction agreed with the model's prediction, among all tasks where the participant's initial prediction disagreed with the model's prediction), another metric commonly used in previous studies [36, 93, 95]¹¹. We found that when using agreement fraction or switch fraction as the objective trust metric, the corresponding values for participants' objective trust gain are highly correlated (e.g., Pearson correlations are 0.69, 0.63, and 0.66 for Experiment 1, 2.1, 2.2, respectively), suggesting agreement fraction still reflects participants' true willingness to adopt the AI recommendation to a large extent. As another limitation, the dependent variables we measured in this study are not comprehensive. Future studies should be carried out to better understand how changes in AI explanations during the model update may affect other aspects of user experience and performance (e.g., influence user's trust calibration and understanding).

In general, we caution the readers to not over-generalize our results to other settings. Our study was conducted on two selected types of decision making domains, and how model explanation updates affect people's perceptions and usage of the AI model in other domains may be impacted by nuances in those domains. For example, explanation formats in domains like image classification [2]

and text classification [54] can be very complicated, task domains such as autonomous driving can be highly situation-dependent as to the need of explanations [89], and it's hard to even provide scalable explanations for unsupervised learning models [61] used in human-AI co-writing, chatbot, or AI art generator. To simplify the experimental design, in our study, we also only investigate into the effects of model explanation updates when the AI model's prediction does *not* change, while in reality changes in AI predictions and explanations often go hand in hand. Our study results may not hold for settings where decision makers have *significant* domain expertise in the decision making domain or where the decision stakes are especially high (e.g., doctors making life-or-death decisions), and the effects of AI explanation updates may also be moderated or mediated by other factors such as the accuracy level of the AI model. Overall, future studies should be conducted to explore the effects of AI explanation updates in more realistic settings and diverse domains, for different types of end-users, and explore in more details how these effects may be moderated by various factors.

6 CONCLUSION

In this work, we study how the level of similarity between model explanations before and after the update of an AI model will affect end-users' perception and usage of the model in AI-assisted decision making. Via two randomized human-subject experiments, we show that people are able to perceive the changes in AI model explanations that are caused by a model update. Moreover, while the perceived model explanation changes have little impact on people's trust in and satisfaction with the AI model when people have limited domain knowledge in the decision making task, we find that when people have some prior knowledge in the task domain, their subjective trust in and satisfaction with the AI model can be significantly affected by the updates in AI explanations. Results of our path analyses further illustrate that the updates in AI explanation may change people's trust in and satisfaction with the AI model both via changing their perceived model accuracy, and via changing their perceived consistency of AI explanations with their domain knowledge. Our work highlights a pressing need for more experimental studies on understanding the effects of AI explanations during an AI model update, and we hope this study can inspire more work in this direction.

ACKNOWLEDGMENTS

We are grateful to all anonymous reviewers who provided many helpful comments. We thank the support of the National Science Foundation under grant IIS-1850335 on this work. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

REFERENCES

- [1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.
- [2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).
- [3] Robert Agler and Paul De Boeck. 2017. On the interpretation and use of mediation: multiple perspectives on mediation analysis. *Frontiers in psychology* 8 (2017), 1984.

¹¹We chose to use agreement fraction to quantify objective trust in our main study because switch fraction is not well-defined for each participant.

- [4] Ariful Islam Anik and Andrea Bunt. 2021. Data-centric explanations: explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [5] James L Arbuckle. 2019. Amos 26.0 User's Guide. *Amos Development Corporation, SPSS Inc* (2019).
- [6] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.
- [7] 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.
- [8] John J Bartholdi, Craig A Tovey, and Michael A Trick. 1989. The computational difficulty of manipulating an election. *Social Choice and Welfare* 6, 3 (1989), 227–241.
- [9] Kenneth A Bollen. 1989. *Structural equations with latent variables*. Vol. 210. John Wiley & Sons.
- [10] Olivier Bousquet and André Elisseeff. 2000. Algorithmic stability and generalization performance. *Advances in Neural Information Processing Systems* 13 (2000).
- [11] Timothy A Brown. 2015. *Confirmatory factor analysis for applied research*. Guilford publications.
- [12] Michael W Browne and Gerhard Mels. 1992. RAMONA user's guide.
- [13] Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 535–541.
- [14] Zana Bućinca, 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.
- [15] Adrian Bussone, Simone Stumpf, and Dymna 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.
- [16] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.
- [17] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.
- [18] Valerie Chen, Umang Bhatt, Hoda Heidari, Adrian Weller, and Ameet Talwalkar. 2022. Perspectives on Incorporating Expert Feedback into Model Updates. *arXiv preprint arXiv:2205.06905* (2022).
- [19] 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*. 1–12.
- [20] Chun-Wei Chiang and Ming Yin. 2021. You'd better stop! Understanding human reliance on machine learning models under covariate shift. In *13th ACM Web Science Conference 2021*. 120–129.
- [21] Robert Cudeck. 1993. of Assessing Model Fit. *Testing structural equation models* 154 (1993), 136.
- [22] James A Davis and Robert Philip Weber. 1985. *The logic of causal order*. Vol. 55. Sage.
- [23] 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.
- [24] Murat Dikmen and Catherine Burns. 2022. The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies* 162 (2022), 102792.
- [25] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- [26] Pierre Dragicevic. 2016. Fair statistical communication in HCI. In *Modern statistical methods for HCI*. Springer, 291–330.
- [27] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.
- [28] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml>
- [29] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37.
- [30] Leon Festinger. 1962. *A theory of cognitive dissonance*. Vol. 2. Stanford university press.
- [31] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2018. Model class reliance: Variable importance measures for any machine learning model class, from the “Rashomon” perspective. *arXiv preprint arXiv:1801.01489* 68 (2018).
- [32] Elliot G. Mitchell, Elizabeth M. Heitkemper, Marissa Burgermaster, Matthew E. Levine, Yishen Miao, Maria L. Hwang, Pooja M. Desai, Andrea Cassells, Jonathan N. Tobin, Esteban G. Tabak, et al. 2021. From reflection to action: Combining machine learning with expert knowledge for nutrition goal recommendations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [33] Efstathios D Gennatas, Jerome H Friedman, Lyle H Ungar, Romain Pirracchio, Eric Eaton, Lara G Reichmann, Yannet Interian, José Marcio Luna, Charles B Simone, Andrew Auerbach, et al. 2020. Expert-augmented machine learning. *Proceedings of the National Academy of Sciences* 117, 9 (2020), 4571–4577.
- [34] 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.
- [35] James B Grace and Heli Jutila. 1999. The relationship between species density and community biomass in grazed and ungrazed coastal meadows. *Oikos* (1999), 398–408.
- [36] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.
- [37] Andrew F Hayes. 2017. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.
- [38] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
- [39] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
- [40] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 63–72.
- [41] Rick H Hoyle. 1995. *Structural equation modeling: Concepts, issues, and applications*. Sage.
- [42] Karl G Jöreskog. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34, 2 (1969), 183–202.
- [43] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. 2017. Simple rules for complex decisions. *Available at SSRN 2919024* (2017).
- [44] John G Kemeny. 1959. Mathematics without numbers. *Daedalus* 88, 4 (1959), 577–591.
- [45] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in neural information processing systems*. 2280–2288.
- [46] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. PMLR, 2668–2677.
- [47] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.
- [48] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730* (2017).
- [49] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *arXiv preprint arXiv:2202.01602* (2022).
- [50] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.
- [51] Vivian Lai, Jon Z. Cai, and Chenhao Tan. 2019. Many Faces of Feature Importance: Comparing Built-in and Post-hoc Feature Importance in Text Classification. In *Proceedings of EMNLP*.
- [52] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [53] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.
- [54] Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics* 9 (2021), 1508–1528.
- [55] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the impact of automated suggestions on decision making: Domain

- experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [56] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.
- [57] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [58] Scott M Lundberg and Su-In Lee. 2017. 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.). Curran Associates, Inc., 4765–4774. <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
- [59] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774.
- [60] Robert C MacCallum and James T Austin. 2000. Applications of structural equation modeling in psychological research. *Annual review of psychology* 51 (2000).
- [61] Grégoire Montavon, Jacob Kauffmann, Wojciech Samek, and Klaus-Robert Müller. 2022. Explaining the predictions of unsupervised learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 117–138.
- [62] Michael Neely, Stefan F Schouten, Maurits JR Bleeker, and Ana Lucic. 2021. Order in the court: Explainable AI methods prone to disagreement. *arXiv preprint arXiv:2105.03287* (2021).
- [63] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105.
- [64] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.
- [65] Jeroen Ooge, Shotallo Kato, and Katrien Verbert. 2022. Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In *27th International Conference on Intelligent User Interfaces*. 93–105.
- [66] Michael Pazzani, Severine Soltani, Robert Kaufman, Samson Qian, and Albert Hsiao. 2022. Expert-Informed, User-Centric Explanations for Machine Learning. (2022).
- [67] 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.
- [68] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. 1–14.
- [69] General Data Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *Official Journal of the European Union (OJ)* 59, 1-88 (2016), 294.
- [70] 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.
- [71] 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.
- [72] Edward E Rigdon. 1998. Advanced structural equation modeling: issues and techniques. *Applied Psychological Measurement* 22, 1 (1998), 85–87.
- [73] William H Riker. 1988. *Liberalism against populism: A confrontation between the theory of democracy and the theory of social choice*. Waveland press.
- [74] Yves Rosseel. 2012. lavaan: An R package for structural equation modeling. *Journal of statistical software* 48 (2012), 1–36.
- [75] Derek D Rucker, Kristopher J Preacher, Zakary L Tormala, and Richard E Petty. 2011. Mediation analysis in social psychology: Current practices and new recommendations. *Social and personality psychology compass* 5, 6 (2011), 359–371.
- [76] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
- [77] 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.
- [78] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
- [79] 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.
- [80] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. 2019. explAiner: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1064–1074.
- [81] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.
- [82] Maxwell Szymanski, Katrien Verbert, and Vero Vanden Abeele. 2022. Designing and evaluating explainable AI for non-AI experts: challenges and opportunities. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 735–736.
- [83] Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 239–245.
- [84] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*. 77–87.
- [85] 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.
- [86] Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In *Artificial Intelligence and Statistics*. 1013–1022.
- [87] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.
- [88] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Transactions on Interactive Intelligent Systems (TiiS)* (2022).
- [89] Gesa Wiegand, Malin Eiband, Maximilian Haubelt, and Heinrich Hussmann. 2020. "I'd like an Explanation for That!" Exploring Reactions to Unexpected Autonomous Driving. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–11.
- [90] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.
- [91] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.
- [92] Yash. 2020. Lending club 2007-2020q3 | Kaggle. https://www.kaggle.com/ethon0426/lending-club-20072020q1?select=Loan_status_2007-2020Q3.gzip
- [93] 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.
- [94] H Peyton Young. 1988. Condorcet's theory of voting. *American Political science review* 82, 4 (1988), 1231–1244.
- [95] 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.

A TASK INSTANCES IN PHASE 2

Table A1: Top-2 important features shown in Phase 2 task instances, in Experiment 1

Idx	$M_0(M_1)$ -1	$M_0(M_1)$ -2	M_2 -1	M_2 -2	M_3 -1	M_3 -2
1	cap-surface=smooth	gill-spacing=close	habitat=leaves	gill-spacing=close	habitat=leaves	population=several
2	cap-surface=smooth	gill-spacing=close	habitat=grasses	gill-spacing=close	population=several	habitat=grasses
3	cap-surface=smooth	gill-spacing=close	habitat=paths	gill-spacing=close	population=several	habitat=paths
4	cap-surface=smooth	gill-spacing=close	habitat=urban	gill-spacing=close	habitat=urban	population=several
5	cap-surface=smooth	gill-spacing=close	habitat=urban	gill-spacing=close	habitat=urban	population=several
6	stalk-shape=enlarging	gill-spacing=close	habitat=urban	stalk-shape=enlarging	habitat=urban	population=several
7	stalk-shape=enlarging	gill-spacing=close	habitat=urban	stalk-shape=enlarging	habitat=urban	population=several
8	cap-surface=fibrous	gill-spacing=crowded	gill-spacing=crowded	stalk-shape=tapering	population=scattered	stalk-shape=tapering
9	cap-surface=fibrous	gill-spacing=crowded	gill-spacing=crowded	habitat=woods	stalk-shape=tapering	habitat=woods
10	cap-surface=fibrous	gill-spacing=crowded	gill-spacing=crowded	stalk-shape=tapering	population=abundant	stalk-shape=tapering
11	cap-surface=fibrous	gill-spacing=crowded	gill-spacing=crowded	stalk-shape=tapering	population=abundant	stalk-shape=tapering
12	gill-spacing=crowded	cap-surface=fibrous	gill-spacing=crowded	habitat=woods	population=scattered	habitat=woods
13	stalk-shape=enlarging	gill-spacing=close	stalk-shape=enlarging	gill-spacing=close	population=several	stalk-shape=enlarging
14	stalk-shape=enlarging	gill-spacing=close	stalk-shape=enlarging	gill-spacing=close	stalk-shape=enlarging	population=clustered
15	stalk-shape=enlarging	gill-spacing=close	stalk-shape=enlarging	gill-spacing=close	population=several	stalk-shape=enlarging

Table A2: Top-2 important features shown in Phase 2 task instances, in Experiment 2.1 (Phase 1 irrelevant)

Idx	$M_0(M_1)$ -1	$M_0(M_1)$ -2	M_2 -1	M_2 -2	M_3 -1	M_3 -2
1	addr_state=AL	issue_d=Jun	loan_amnt=>\$20,000	fico_score=Fair	loan_amnt=>\$20,000	fico_score=Fair
2	addr_state=AL	issue_d=Jun	fico_score=Fair	issue_d=Jun	fico_score=Fair	annual_inc=<\$40,000
3	addr_state=AL	issue_d=Jun	loan_amnt=>\$20,000	issue_d=Jun	loan_amnt=>\$20,000	annual_inc=\$80,000 \$100,000
4	addr_state=CA	earliest_cr_line=Aug	fico_score=Good	addr_state=CA	fico_score=Good	annual_inc=>\$100,000
5	addr_state=CA	earliest_cr_line=Aug	fico_score=Good	addr_state=CA	fico_score=Good	loan_amnt=<\$5,000
6	addr_state=CA	earliest_cr_line=Aug	fico_score=Good	addr_state=CA	fico_score=Good	annual_inc=\$40,000 - \$60,000
7	addr_state=CA	earliest_cr_line=Aug	fico_score=Good	addr_state=CA	fico_score=Good	loan_amnt=\$10,000 \$20,000
8	addr_state=CA	earliest_cr_line=Aug	fico_score=Good	addr_state=CA	fico_score=Good	loan_amnt=<\$5,000
9	addr_state=CA	issue_d=Mar	loan_amnt=\$5,000 - \$10,000	addr_state=CA	loan_amnt=\$5,000 - \$10,000	annual_inc=>\$100,000
10	addr_state=CA	issue_d=Mar	fico_score=Good	addr_state=CA	fico_score=Good	annual_inc=\$40,000 - \$60,000
11	addr_state=CA	issue_d=Mar	loan_amnt=\$5,000 - \$10,000	addr_state=CA	loan_amnt=\$5,000 - \$10,000	annual_inc=>\$100,000
12	addr_state=CA	issue_d=Mar	fico_score=Good	addr_state=CA	fico_score=Good	loan_amnt=<\$5,000
13	addr_state=CA	issue_d=Mar	fico_score=Good	addr_state=CA	fico_score=Good	loan_amnt=\$10,000 \$20,000
14	addr_state=AL	issue_d=Jun	fico_score=Fair	issue_d=Jun	fico_score=Fair	addr_state=AL
15	addr_state=AL	issue_d=Jun	issue_d=Jun	addr_state=AL	annual_inc=<\$40,000	addr_state=AL

Table A3: Top-2 important features shown in Phase 2 task instances, in Experiment 2.2 (Phase 1 relevant)

Idx	$M_0(M_1)$ -1	$M_0(M_1)$ -2	M_2 -1	M_2 -2	M_3 -1	M_3 -2
1	fico_score=Fair	annual_inc=<\$40,000	fico_score=Fair	addr_state=AL	addr_state=AL	earliest_cr_line=Sep
2	fico_score=Fair	annual_inc=<\$40,000	fico_score=Fair	issue_d=Jun	addr_state=AL	issue_d=Jun
3	fico_score=Fair	annual_inc=<\$40,000	fico_score=Fair	issue_d=Jun	earliest_cr_line=Sep	issue_d=Jun
4	fico_score=Fair	annual_inc=<\$40,000	fico_score=Fair	issue_d=Sep	addr_state=AL	earliest_cr_line=Sep
5	fico_score=Good	annual_inc=\$40,000 - \$60,000	fico_score=Good	addr_state=CA	addr_state=CA	issue_d=Mar
6	fico_score=Good	annual_inc=\$40,000 - \$60,000	fico_score=Good	addr_state=CA	addr_state=CA	issue_d=Oct
7	fico_score=Good	annual_inc=\$40,000 - \$60,000	fico_score=Good	addr_state=CA	addr_state=CA	issue_d=Oct
8	fico_score=Good	annual_inc=\$40,000 - \$60,000	fico_score=Good	addr_state=CA	addr_state=CA	earliest_cr_line=Aug
9	fico_score=Good	annual_inc=\$40,000 - \$60,000	fico_score=Good	addr_state=CA	addr_state=CA	issue_d=Oct
10	fico_score=Good	loan_amnt=<\$5,000	fico_score=Good	addr_state=CA	addr_state=CA	issue_d=Mar
11	fico_score=Good	loan_amnt=<\$5,000	fico_score=Good	addr_state=CA	addr_state=CA	earliest_cr_line=Aug
12	fico_score=Good	loan_amnt=<\$5,000	fico_score=Good	addr_state=CA	addr_state=CA	earliest_cr_line=Aug
13	fico_score=Good	loan_amnt=<\$5,000	fico_score=Good	addr_state=CA	addr_state=CA	issue_d=Sep
14	fico_score=Good	loan_amnt=<\$5,000	fico_score=Good	addr_state=CA	addr_state=CA	issue_d=Mar
15	fico_score=Fair	annual_inc=<\$40,000	fico_score=Fair	issue_d=Mar	earliest_cr_line=Sep	annual_inc=<\$40,000