Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making

Xinru Wang  
xinruw@purdue.edu  
Purdue University  
West Lafayette, Indiana, USA

Ming Yin  
mimgyin@purdue.edu  
Purdue University  
West Lafayette, Indiana, USA

ABSTRACT

This paper contributes to the growing literature in empirical evaluation of explainable AI (XAI) methods by presenting a comparison on the effects of a set of established XAI methods in AI-assisted decision making. Specifically, based on our review of previous literature, we highlight three desirable properties that ideal AI explanations should satisfy—improve people’s understanding of the AI model, help people recognize the model uncertainty, and support people’s calibrated trust in the model. Through randomized controlled experiments, we evaluate whether four types of common model-agnostic explainable AI methods satisfy these properties on two types of decision making tasks where people perceive themselves as having different levels of domain expertise in (i.e., recidivism prediction and forest cover prediction). Our results show that the effects of AI explanations are largely different on decision making tasks where people have varying levels of domain expertise in, and many AI explanations do not satisfy any of the desirable properties for tasks that people have little domain expertise in. Further, for decision making tasks that people are more knowledgeable, feature contribution explanation is shown to satisfy more desiderata of AI explanations, while the explanation that is considered to resemble effective AI explanation, concerning these desiderata on two different types of decision making tasks.

1 INTRODUCTION

In recent years, numerous AI-driven decision aids have been developed to assist people in making better decisions in diverse domains ranging from financial investment to criminal justice. To overcome the black-box nature of many complex AI models underlying the decision aids, various explainable AI (XAI) methods are designed to inform people of the reasoning processes underneath the algorithmic decisions. For example, sophisticated techniques such as LIME [60] and SHAP [48] have been used to illustrate how much each feature contributes to a model’s final prediction. Meanwhile, reviews of the social science literature reveal that humans tend to provide contrastive explanations when explaining their decisions to each other, suggesting that explaining an AI using counterfactual examples can be most understandable to humans as they share similar conceptual framework as human explanations [12, 54].

The rapid development of XAI methods raises a few key questions in rigorously evaluating and systematically comparing these methods: What properties characterize an effective AI explanation? Do the established explanation methods satisfy these properties? Is the extent to which an AI explanation satisfies these properties dependent on the property of the decision making task, such as how much the human decision makers feel they know about the task domain a priori? To answer these questions, researchers have been advocating for moving beyond defining what constitutes a “good” explanation using model designer’s intuition but actually examining how useful an explanation is with human users [24, 59]. In responding to this call, there is recently a growing line of literature on empirically evaluating the effectiveness of XAI methods (e.g., [14, 16, 41, 69, 71]). Yet, principles required for an explanation to be considered helpful in AI-assisted decision making, arguably, still remain to be articulated and comprehensively assessed.

In this paper, we contribute to the XAI research by positing three desirable properties that ideal AI explanations should satisfy in order to be considered helpful in AI-assisted decision making, and presenting a human-subject experiment which empirically compares to what extent established XAI explanation methods satisfy these desiderata on two different types of decision making tasks.

We start by reviewing the existing literature in empirical evaluation of XAI methods and summarizing from them three desirable properties that characterize an effective AI explanation, concerning how well the explanation can help people 1) understand the AI model, 2) recognize the uncertainty underlying an AI prediction, and 3) calibrate their trust in the model. These properties are stated mainly from the point of view of a human decision maker who is assisted by an AI-driven decision aid, and are certainly not comprehensive. However, they provide an initial set of concrete standards
upon which we can compare the strengths and weaknesses of different XAI methods.

We then conduct randomized human-subject experiments on Amazon Mechanical Turk to understand to what extent various types of established XAI methods (e.g., feature importance, feature contribution, nearest neighbors, counterfactual examples) satisfy these desirable properties, when they are used to assist human decision makers in two different decision making contexts (i.e., recidivism prediction and forest cover prediction) where people perceive themselves as having varying levels of domain expertise. Our results suggest that the effectiveness of different XAI methods largely depends on the properties of the decision making task. In particular, for tasks that people have limited domain expertise in (e.g., forest cover prediction), none of the three desiderata is reliably satisfied by any of the XAI methods that we have looked into. On the other hand, on decision making tasks that people feel that they have some domain expertise in (e.g., recidivism prediction), different XAI methods are shown to satisfy the three desiderata to different degree. For instance, showing each feature’s contribution to the model’s prediction in individual cases seems to have the potential to satisfy more of the desiderata. In contrast, the two example-based XAI methods we have examined, including providing counterfactual examples which is believed to resemble human explaining processes, seem to lack the ability to support trust calibration.

Our findings provide important implications on designing and selecting effective XAI methods that are most suitable for the type of decision making task and for the intended purposes, as well as fairly and transparently reporting empirical evaluation results of XAI methods. We conclude by discussing these implications.

2 LITERATURE REVIEW

Overview of AI explanation methods. Earlier literature on AI explanations often concerns the communication of uncertainty in AI decisions [45, 46]. More recently, the surge of interests in increasing the interpretability and transparency of AI has brought about the development of a variety of techniques for explaining the rationale of AI decisions, and different taxonomies of these techniques also emerge. For example, methods that aim at explaining the behavior of the entire AI model is categorized as global explanations, while methods that provide reasons for specific model predictions are categorized as local explanations [2, 24, 26]. In addition, depending on whether the explanation is designed for a particular type of model, explanations can also be divided into model-specific methods and model-agnostic methods. Model-specific methods often involve the development of intrinsically interpretable models such as generalized additive models and decision sets [15, 34, 43, 66], as well as visualizing what deep neural network has learned in its intermediate layers and how its predictions are affected by different part of the inputs (e.g., through saliency map) [37, 61, 68]. On the other hand, typical model-agnostic methods include providing information on global-level feature importance [28], computing feature contribution on individual predictions [48, 60], using examples in the training dataset or counterfactual examples to explain model predictions [36, 39, 65], and conducting model distillation [10, 32].

Desiderata of AI explanations. In contrast to the rapid development of explainable AI methods, systematic understandings of what is an effective AI explanation fall far behind. Most recently, researchers argued that the interpretability of an AI model should not be defined using the model designer’s intuition. Instead, it should be defined by user behavior, that is, whether model explanations can improve people’s abilities in completing various tasks [24, 59], and the “people” here can be different parties in the AI ecosystems including model developers, regulators, and end-users [63, 64]. Researchers have proposed many tasks that AI explanations should assist people in. We reviewed these tasks and used two criteria to narrow down the scope of the tasks from which we extracted the desiderata of AI explanations—first, we focused on those tasks related to the ability of human decision makers in making decisions when they are assisted by an AI model; second, we required the tasks to be easily applicable to any kind of decision making context1. Based on tasks that satisfy these criteria, we summarized three desiderata of AI explanations as follows:

- **Desideratum 1 (Understanding):** Explanations of an AI should improve people’s understanding of it.
- **Desideratum 2 (Uncertainty awareness):** Explanations of an AI should help people recognize the uncertainty underlying an AI prediction and nudge people to rely on the model more on high confidence predictions when the model’s confidence is calibrated.
- **Desideratum 3 (Trust calibration):** Explanations of an AI should empower people to trust the AI appropriately.

Desideratum 1 is the most straightforward one, and researchers have proposed various methods to assess people’s understanding of an AI model. Typical methods include ask people to rank the input data features based on their influence to overall predictions [18, 33], to indicate the direction of change in the model’s prediction when a feature’s value is altered [16, 18, 29], to simulate the model’s predictions [16, 24, 40, 47, 59], to answer “what-if” questions about the model behavior [7, 16, 24, 33, 54], and to detect mistakes of the model and debug the model [59, 60].

Desideratum 2 connects to the needs of communicating the uncertainty inherent in AI model predictions to people [71]. Ideal AI explanations inform people of when the model is confident in its predictions and provide insights into when it is uncertain; thus they allow people to act upon different predictions differently. In particular, when the AI model’s confidence is calibrated (i.e., the model’s confidence accurately reflects the model’s correctness likelihood), the explanations should provide useful cues for people to infer the model’s confidence on each case and adjust their reliance on the model’s predictions based on the inferred model confidence.

Finally, Desideratum 3 concerns the ultimate goal of AI-assisted decision-making, that is, to maximize the joint human-AI team performance [5, 6, 11]. An essential step towards this goal is to use explanations to guide people to trust an AI model when it is right and not to trust it when it is wrong. In other words, with the assistance of model explanations, people should have better capability of calibrating their trust in the model [69, 71]. Note that when an explanation simply improves the human-AI joint decision making accuracy, it does not necessarily mean this desideratum is satisfied.

---

1An example of task that may not be applicable to some decision making context is to examine human decision maker’s ability in detecting fairness problems of the AI model or utilizing the AI model more fairly, in the presence of explanations [23, 30].
We set out to conduct an experimental study to gain in-depth understandings of whether and to what extent various established AI explanation methods can bring about human’s desirable behavior in AI-assisted decision making. Corresponding to the desiderata listed in Section 2, we ask the following research questions:

- **RQ1**: How do different types of explanation impact people’s understandings of an AI model?
- **RQ2**: How do different types of explanation influence people’s capability of differentiating a model’s high confidence predictions from the low confidence ones?
- **RQ3**: How do different types of explanation change people’s ability of calibrating their trust in an AI model?

We focus on *model-agnostic* explanation methods in this study, as these methods can be applied to any kind of AI models. Further, we conduct our study on different decision making tasks to see whether and how the answers to these questions may change for tasks with different properties, such as tasks that people perceive themselves as having varying levels of domain expertise in.

### 3.1 Decision Making Tasks

In our study, we asked participants to complete a set of decision making tasks with the help from decision aids that were powered by machine learning models. Specifically, we considered two types of decision making tasks:

- **Recidivism prediction**: In this task, participants were asked to review a profile of a criminal defendant consisted of 7 features—the defendant’s race, sex, age, the number of non-juvenile prior crimes, name of the currently charged crime, degree of the current charge, and the number of days the defendant spent in custody for the current charge. After reviewing the profile, participants were asked to make a prediction on whether this defendant would reoffend within two years. The defendant profiles were selected from the COM-PAS dataset, which contained information of 7,214 criminal defendants in Broward County, Florida, USA, between 2013 and 2014 [4, 44]. This dataset was widely used by researchers to understand how people interact with machine assistance in their decision making [23, 31].
- **Forest cover prediction**: In this task, participants were asked to review a profile of a (30m × 30m) cell containing 8 features—the area’s elevation, aspect, slope, hillshade index, the horizontal/vertical distance to nearest surface water, the horizontal distance to nearest roadway, and the horizontal distance to nearest wildfire ignition points. After reviewing the profile, participants were asked to make a prediction on whether this area is primarily covered by the spruce-fir forest. These geological profiles

<table>
<thead>
<tr>
<th>Publications</th>
<th>Decision making tasks</th>
<th>AI Explanation methods</th>
<th>Desideratum 1 (Understanding)</th>
<th>Desideratum 2 (Uncertainty awareness)</th>
<th>Desideratum 3 (Trust calibration)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foursahib-Sangheeth et al. [9]</td>
<td>house price prediction</td>
<td>interpretable model</td>
<td>mixed results</td>
<td>N/A</td>
<td>$\nabla$</td>
</tr>
<tr>
<td>Alqaraawi et al. [3]</td>
<td>image classification</td>
<td>feature contribution</td>
<td>mixed results</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Chu et al. [17]</td>
<td>age prediction</td>
<td>saliency map</td>
<td>N/A</td>
<td>N/A</td>
<td>$\nabla$</td>
</tr>
<tr>
<td>Cheng et al. [16]</td>
<td>student admission</td>
<td>feature contribution</td>
<td>$\nabla$</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Zhang et al. [71]</td>
<td>income prediction</td>
<td>feature contribution</td>
<td>$\nabla$</td>
<td>N/A</td>
<td>$\nabla$</td>
</tr>
<tr>
<td>Bansal et al. [6]</td>
<td>sentiment analysis</td>
<td>feature contribution</td>
<td>N/A</td>
<td>N/A</td>
<td>$\nabla$</td>
</tr>
<tr>
<td>Carton et al. [14]</td>
<td>toxicity content</td>
<td>feature contribution</td>
<td>N/A</td>
<td>N/A</td>
<td>$\nabla$</td>
</tr>
<tr>
<td>Lai and Tan [42]</td>
<td>detection</td>
<td>feature contribution</td>
<td>N/A</td>
<td>N/A</td>
<td>$\nabla$</td>
</tr>
<tr>
<td>Lai et al. [41]</td>
<td>deception detection</td>
<td>feature contribution</td>
<td>N/A</td>
<td>N/A</td>
<td>$\nabla$</td>
</tr>
<tr>
<td>Cai et al. [13]</td>
<td>drawing recognition</td>
<td>example-based</td>
<td>mixed results</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Yang et al. [69]</td>
<td>leaf classification</td>
<td>example-based</td>
<td>N/A</td>
<td>N/A</td>
<td>$\nabla$</td>
</tr>
</tbody>
</table>

Table 1: Summary of recent empirical studies examining the effects of explanations in AI-assisted decision making (top panel: studies using model-specific explanations; bottom panel: studies using model-agnostic explanations).

Note: ‘N/A’ means the study does not examine the desideratum; ‘$\nabla$’ means the study finds (or does not find) evidence suggesting the explanation method it examined satisfies a desideratum. In the $\nabla$ (or $\nabla$) cases, the study only reports human’s decision making accuracy is increased (or not changed) after receiving model explanation, which is not sufficient for us to draw conclusions on trust calibration.
were selected from the UCI cover type dataset [8, 27], which recorded the geological information collected from 581,012 observation areas located in the Roosevelt National Forest of northern Colorado, USA. In the original dataset, the primary forest cover for each area is one of the 6 types of tree species, including spruce/fir. To simplify the task, we only asked participants to make a binary prediction on whether the primary tree species in an area is spruce/fir or not.

For both types of decision making tasks, we trained a logistic regression model to help people make predictions. In particular, in each decision making task, the participant was asked to first review the profile (of a defendant or of a wilderness area) to make her own prediction. Then, we would present to the participant the model’s prediction, possibly together with some explanation on why the model made such prediction (see more detail in Section 3.2). Lastly, the participant needed to make a final prediction. Figure 1(a) shows an example of the task interface.

We chose the recidivism prediction and forest cover prediction tasks in our study for two main reasons. First, both tasks reflect realistic use cases of AI-driven decision aids, as machine learning models have been developed to assist people in making better decisions in social justice [9, 19, 38] as well as forest management [49, 50]. Second, people’s perceptions of the amount of domain expertise they have in these two types of decision making tasks can be quite different. In particular, we speculate that most people may perceive themselves as having a higher degree of domain expertise in the recidivism prediction tasks, because they can easily apply their day-to-day, common sense knowledge in their predictions. In contrast, people may find the forest cover prediction task to require more domain expertise that they are lack of. To confirm this intuition, we conducted a pilot study, in which we introduced both types of decision making tasks to participants that we recruited from Amazon Mechanical Turk (MTurk), and we asked them to decide on which of these two tasks, they felt themselves to be more knowledgeable. We also asked participants to indicate among these two tasks, on which task they feel they (or a normal person) can be more accurate predictions, and they would be more confident in making recidivism predictions. In other words, consistent with our conjecture, most laypeople perceived themselves to have a higher level of domain expertise in the recidivism prediction tasks, compared to the forest cover prediction tasks.

### 3.2 Experimental Design

#### 3.2.1 Experimental Treatments.

We adopted a between-subject design in our experiment. For each type of decision making task, we created 5 treatments by varying whether and how the model’s predictions were explained:

- **No explanation (Control):** Participants would not receive any explanation on the model’s prediction on each task.
- **Feature importance:** In this treatment, we explained the model’s prediction to participants by showing to them the overall “importance” of different features in influencing the model’s predictions. Specifically, we adopted the permutation feature importance method [28] to compute the importance of each feature as the increase of the model’s prediction error after permuting the values on that feature, and we visualized different feature’s importance using a bar chart (Figure 1(b)).
- **Feature contribution:** In this treatment, we explained the model’s prediction to participants by showing to them the contribution of each feature to the prediction. Since we used logistic regression models in this study, we computed a feature’s contribution to a prediction as the log-odds influence of that feature. We then provided a bar chart in each task to visualize the contributions of all features in that task as well as the base rate1 (Figure 1(c)).
- **Nearest neighbors:** In this treatment, we explained the model’s prediction to participants by showing to them the nearest neighbors of a hypothetical specimen presented to the participant. We visualized the nearest neighbors using a bar chart (Figure 1(d)).
- **Counterfactuals:** In this treatment, we explained the model’s prediction to participants by showing to them the

---

1The base rate is the log-odds value for a hypothetical profile in which the value of each feature takes the reference level.
model’s predictions on other similar data points (i.e., profiles) in the training dataset. For each task, we looked into all profiles in the training dataset on which the model’s predictions were correct, and we selected two of them—the one most similar to the current profile on which the model made the same prediction as that in the current task, and the one most similar to the current profile on which the model made a different prediction than that in the current task. We then presented these two training profiles in a table, side by side with the profile of the current task (Figure 1(d)).

- **Counterfactuals**: In this treatment, we explained the model’s prediction on each task by exploring what changes in feature values result in an opposite model prediction. For each feature, we either displayed the smallest change that is needed on that feature to flip the model’s prediction (when other feature values are unchanged), or we told participants that changing that feature’s value would not affect the model’s prediction (Figure 1(e)).

Together, these treatments covered a diverse set of classical model-agnostic explanations that are commonly used for explaining AI models [7, 23]. For example, the feature importance explanation is a global explanation while the other three are local explanations. Further, the feature importance and feature contribution explanations aim to explain the model by summarizing feature-based statistics, while the other two explanations are example-based.

3.2.2 **Selection of task instances.** Within one type of decision making task, participants in different treatments worked on the same set of 32 task instances, and the model predictions they saw on each task were produced by the same logistic regression model that we trained using a subset of the original COMPAS or UCI cover type dataset. On the hold-out test datasets consisting of 1,000 task instances, the accuracy of our logistic regression models is 69.1% for the recidivism prediction task and 69.5% for the forest cover prediction task, which suggests a reasonable predictive validity.

To better answer our research question RQ1–RQ3, we carefully selected from the test datasets the 32 task instances that we presented to our participants. In particular, we categorized the logistic regression model’s confidence on a task instance as high or low depending on whether the model’s probability estimate of the predicted label is higher than 0.7, and we confirmed this probability aligned well with the model’s correctness likelihood on each instance. We included in our task set 16 instances that the model’s confidence is low and 16 instances that the model’s confidence is high. To ensure the representativeness of the selected instances, we projected all task instances in the test dataset onto the two features with the largest predictive power, and the 16 low (or high) confidence task instances we included in our final task set were the “prototypes” that can cover the centers of the data distributions for all data instances in the test set where the model’s confidence is low (or high) [36, 55].

3.3 **Experimental Procedure**

We conducted our study on both types of decision making tasks by posting human intelligence tasks (HITs) on Amazon Mechanical Turk (MTurk) and recruiting MTurk workers as our participants. Upon arrival, participants were randomly assigned to one of the 5 experimental treatments. They first completed a survey on their background, including their demographics, technical literacy, and expertise in machine learning. Then, we presented participants with an interactive tutorial to walk them through the task interface. If a participant was assigned to a treatment with model explanation, we also included examples of the model explanation in the tutorial and provided instructions to help the participant understand the explanations. We included a few qualification questions in the tutorial to make sure that participants correctly understand all the information. For those participants working on predicting forest cover, we further helped participants get familiar with the task, following a similar procedure as those used in previous literature [69, 71]—we provided participants with a brief introduction about the characteristics of spruce-fir forests as well as a set of 10 training tasks, in which participants needed to make predictions on the forest cover type without the assistance from a model, and they learned about the correct answer after each task.

After completing the tutorial, the participant then moved on to work on the set of 32 decision making tasks with the assistance from the machine learning model, and the order of the tasks was randomized across participants. In each task, the participant followed the three-step procedure as we have described in Section 3.1—make an initial independent prediction, review the model prediction (and explanation), and make a final prediction. The participant was not given any accuracy feedback on either her prediction or the model’s prediction on any of these tasks.

Finally, before submitting the HIT, the participant needed to complete an exit survey, which included a set of multiple-choice questions testing her objective understanding of the model behavior. In addition, the participant was also asked to report her perceived understanding of the model by answering a few survey questions (see Section 3.4 for more details). We included two attention check questions in the HIT in which the participant was instructed to select a pre-specified option, which 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 for the recidivism prediction tasks and $2.00 for the forest cover prediction tasks. To incentivize participants to carefully read about the model’s explanation in each task and adjust their behavior accordingly, we further provided them with additional performance-contingent bonuses—if the overall accuracy of the participant’s final predictions on the 32 tasks was at least 60%, she can earn a bonus of $0.05 for each of her correct final predictions; and for each correct answer the participant submitted to a multiple-choice question about the model behavior in the exit survey, she could also earn a $0.10 bonus. The maximum amount of bonuses a participant could earn in this study was $2.26.

---

3While the format of feature contribution explanation we used in our study was specific to logistic regression models, explaining model predictions by showing the contribution of each feature is applicable for other models [46, 60].

4The base payment for the forest cover prediction tasks was higher because participants spent more time on them due to the addition of training tasks.
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 existence and type of model explanations that the participant received.

3.4.2 Dependent variables. For RQ1, we used two main dependent variables: (1) a participant’s objective understanding of the ML model, as measured by the number of multiple-choice questions in the exit survey that she answered correctly, and (2) the participant’s subjective understanding of the ML model, as measured by her self-report in the exit survey. Specifically, based on our literature review on how people’s understanding of an AI is assessed in existing literature (see Section 2), we designed a set of 9 multiple-choice questions that aim at evaluating participants’ knowledge of the model behavior from various aspects, including:

- **Compare feature importance**: participants were asked to select among a list of features which one was most/least influential on the model’s overall predictions (2 questions)
- **Specify a feature’s marginal effect on predictions**: participants were asked questions like “if the value of feature X of this profile is x2 instead of x1, would it increase or decrease the chance for the model to predict Y?” (1 question)
- **Counterfactual thinking**: participants were given a reference profile, and they were asked to select from a list of changes in feature values the ones that they believed would result in an opposite model prediction (2 questions)
- **Simulate model behavior**: participants were given a profile, and they were asked to predict what the model would predict on this profile (2 questions)
- **Error detection**: participants were given a profile, the model’s prediction on the profile, and the model’s explanation (if applicable), and they were asked to determine whether the model’s prediction was correct (2 questions)

The full list of multiple-choice questions is included in the supplementary material. Moreover, we asked participants to report their own perceived understandings of the model by indicating their agreement on the following two statements (adapted from earlier literature [13, 16]) from 1 (“strongly disagree”) to 7 (“strongly agree”): (1) I understand how the model works to predict whether a defendant will reoffend [whether the primary tree species in an area is spruce/fir]; (2) I can predict how the model will behave. The participant’s subjective understanding of the model is then computed as her average ratings on these two statements. We expect that if an AI explanation improves people’s understanding of the AI (i.e., satisfy Desideratum 1), the participant’s objective and subjective understanding scores would both increase.

For RQ2, we looked into participants’ capability in differentiating the model’s high confidence predictions from its low confidence predictions by examining how people’s reliance on the model changes with the model’s confidence. Following earlier literature [70, 71], we quantified people’s reliance on the model using the fraction of tasks in which the participant’s final prediction was the same as the model’s prediction (i.e., agreement fraction). If an AI explanation can expose the uncertainty of AI predictions to people (i.e., satisfy Desideratum 2), given that the confidence of our model is calibrated, we expect participants’ reliance on the model to be higher on high confidence predictions.

For RQ3, we evaluated participants’ capability of calibrating their trust in the ML model using three main independent variables, including their appropriate trust [51–53, 56] (i.e., the fraction of tasks where participants used the model’s prediction when the model was correct and did not use the model’s prediction when the model was wrong; this is effectively participants’ final decision accuracy), overtrust [22, 58] (i.e., the fraction of tasks where participants used a wrong model prediction) and undertrust [22, 58] (i.e., the fraction of tasks where participants did not use a correct model prediction). If an AI explanation supports trust calibration (i.e., satisfy Desideratum 3), we expect that participants’ appropriate trust in the model would increase, while their overtrust or undertrust in the model would decrease.

3.4.3 Statistical methods. To avoid multiple comparison problems and control false discovery, we conducted our analyses using the interval estimate method [21, 25]. We visualized our data by plotting the mean values of the dependent variable of interest for each treatment (or the difference in the mean value of a dependent variable between a treatment with model explanation and the control treatment) along with the 95% bootstrap confidence intervals ($R = 5000$). We interpreted our results based on the range of confidence intervals, and measured the effect sizes using Cohen’s $d$ [20]. To take the impact of covariates (e.g., participants’ demographics) into account, we then constructed mixed-effect regression models which treated each participant as a random effect and controlled for covariates, while fixed effects were decided by the variables of interests in each research question. Results of these models are interpreted via the estimated coefficient values for the fixed effect variables as well as their 95% bootstrap confidence intervals.

4 RESULTS

After filtering data from inattentive participants, we obtained valid data from 782 participants on the recidivism prediction tasks (62.9% male, the average age is 38), and 561 participants on the forest cover prediction tasks (64.3% male, the average age is 39). We analyzed these data to answer our research questions.

4.1 RQ1: Effects on understanding AI models

We start with examining the impact of different types of explanation on people’s understanding of the machine learning model (RQ1). For each participant, we first normalized her objective and subjective understanding scores by dividing them by the maximum possible values. Figure 2 then shows the average changes of a participant’s normalized objective and subjective understanding scores between each treatment with a specific type of model explanation and the treatment without model explanation (i.e., the control treatment). Interestingly, on the task that people perceived themselves as having more domain expertise in (i.e., the recidivism prediction task, Figure 2(a)), we found that both the feature importance and counterfactual explanations increase participants’ objective understanding of the model (Cohen’s $d$=0.26, 95% CI [0.03, 0.48] for feature importance, and 0.27 [0.04, 0.48] for counterfactuals), and all four types of explanations increase participants’ subjective understanding of the model (Cohen’s $d$=0.28, 95% CI [0.05, 0.49])
when aggregating all explanation types). However, on the task that people have limited domain expertise in (i.e., the forest cover prediction task, Figure 2(b)), we were only able to conclude that the feature importance explanation increases participants’ objective understanding of the model (Cohen’s $d=0.33$, 95% CI [0.06, 0.59]), while the feature contribution explanation increases participants’ subjective understanding (Cohen’s $d=0.28$, 95% CI [0.01, 0.55]).

Further, we constructed mixed effect regression models to predict the correctness of a participant’s answer on each multiple-choice question or a participant’s rating on each subjective understanding survey question. We treated the type of explanation a participant received as the fixed effect, the participant as the random effect, and controlled for the participant’s age, gender, and education as covariates5. Our regression results were consistent with what we have observed in Figure 2. For example, we found that on the recidivism prediction task, feature importance and counterfactual explanations lead to higher levels of objective understanding (estimated coefficient $\beta = 0.04 [0.007, 0.07]$ for feature importance, and $\beta = 0.04 [0.01, 0.07]$ for counterfactuals), and all 4 types of explanations result in higher levels of subjective understanding (feature importance: $\beta = 0.04 [0.03, 0.06]$, feature contribution: $\beta = 0.04 [0.03, 0.06]$, nearest neighbors: $\beta = 0.03 [0.01, 0.05]$, counterfactuals: $\beta = 0.05 [0.03, 0.07]$). On the forest cover prediction task, other than the positive coefficient associated with the feature importance explanation on influencing objective understanding ($\beta = 0.05 [0.01, 0.09]$) and the positive coefficient associated with the feature contribution explanation on influencing subjective understanding ($\beta = 0.04 [0.02, 0.06]$), the effects of other explanations are inconclusive. We also found that on both tasks, female had higher levels of objective understanding of the model compared to male participants, while participants who self-reported to have a higher level of education had lower objective understanding scores.

4.2 RQ2: Effects on recognizing model uncertainty

We next constructed mixed effect regression models to understand participants’ capability in recognizing model uncertainty in different treatments when accounting for various covariates. More specifically, regression models were built for estimating whether a participant would use the model’s prediction as her final prediction in a task, and we included the type of model explanation the participant received, the raw value of model confidence on the task, as well as the interaction between explanation type and model confidence as the fixed effects. We further treated each participant as the random effect and controlled for the participant’s demographic information. Doing so, we again detected that for recidivism prediction, the coefficients for the interaction terms between model confidence and each type of model explanation are reliably estimated to be positive (feature importance: $\beta = 0.20 [0.06, 0.35]$, feature contribution: $\beta = 0.27 [0.12, 0.42]$, nearest neighbor: $\beta = 0.17 [0.01, 0.33]$, counterfactual: $\beta = 0.17 [0.02, 0.31]$), indicating that participants might have utilized model explanations to infer model uncertainty and rely on high confidence model predictions more when they had some domain expertise in the task. For the forest cover prediction task, however, we did not get conclusive evidence suggesting that

---

5We also constructed mixed effect regression models when controlling for the participant’s technical literacy and expertise in machine learning as covariates in addition to the demographic background, and the results are qualitatively similar.
the coefficients for any of the four interaction terms were different from zero. That is, participants making prediction on forest cover did not seem to act upon model predictions with varying levels of confidence differently in the presence of model explanations.

4.3 RQ3: Effects on trust calibration

Finally, we look into how different explanations influence people’s capability of recognizing model confidence. We summarized our experimental results in Table 2, and we highlight a few key findings:

- The effects of model explanations are dramatically different on tasks where people have varying levels of domain expertise in. Notably, for decision making tasks that people are not knowledgeable about, most established AI explanations did not satisfy any of the three desiderata.
- The only positive effect of model explanation that we have consistently observed across different decision making tasks...
Figure 4: Comparing how different types of explanation support participants’ trust calibration in the AI. For appropriate trust, the larger the value the better. For overtrust and undertrust, the smaller the value the better.

Table 2: Summary of our experimental results.

<table>
<thead>
<tr>
<th>Explanation type</th>
<th>Recidivism prediction</th>
<th>Forest cover prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Understanding</td>
<td>Uncertainty Awareness</td>
</tr>
<tr>
<td>feature importance</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>feature contribution</td>
<td>✓?</td>
<td>✓</td>
</tr>
<tr>
<td>nearest neighbor counterfactuals</td>
<td>✓?</td>
<td>✓</td>
</tr>
</tbody>
</table>

Note: ✓ (or ✗) means our study finds (or does not find) supportive evidence suggesting the explanation method satisfies a desideratum. In the ✓? cases, we only find partial evidence supporting the explanation increases people’s understanding of the model (either measured by objective understanding or subjective understanding, but not both).

is that feature importance explanations increase people’s objective understanding of an AI model, while feature contribution explanations increase people’s subjective understanding of an AI model.

- Among the 4 types of explanations that we have examined, the feature contribution explanation seems to be able to satisfy more desiderata of AI explanations when people have some domain expertise in the decision making task.
- For the two example-based explanations in our study, we found minimal evidence on their capability to support trust calibration. In particular, for the counterfactual explanation, which is considered to closely resemble how human explain decisions, it indeed helps increase people’s understanding of the AI model, but only on tasks that people have some domain expertise in. The improved understanding of the model brought up by the counterfactual explanation, however, fail to help people calibrate their trust in the model.

5 CONCLUSION AND DISCUSSION

In this paper, we present a comparative study to understand the effectiveness of four types of XAI methods in supporting people to make better decisions. We first identify three desiderata of AI explanations as critical for people to understand the AI, recognize the uncertainty underlying the AI, and calibrate their trust in the AI in AI-assisted decision making. We further conduct randomized experiments to evaluate whether commonly-used model-agnostic XAI methods satisfy these desiderata on two types of decision making tasks where people have varying levels of domain expertise in. We found that on tasks that people have little domain expertise in, none of the four AI explanations we examined reliably satisfy any of the three desiderata. On tasks that people perceive themselves as more knowledgeable, our results provided evidence supporting that the feature contribution explanation has the potential to satisfy more desiderata. In the following, we provide possible explanations of our results, and discuss implications and limitations of our study.

The role of domain expertise. The ineffectiveness of various XAI methods in supporting human decision makers on tasks that they have limited domain expertise in raises an important question of understanding why. We conjecture that this may be due to a number of reasons. First, without the domain expertise, people may find the explanations to be rather foreign and mentally taxing to consume, thus their ability to absorb the information carried in the explanations decreases. This is particularly true in our study, as participants in the forest cover prediction task may not only have limited knowledge of how different features relate to the output, but they may even need to learn the meanings of some features. In addition, people’s domain expertise may play an important role in facilitating people’s inference of the uncertainty and correctness of an AI prediction. For example, when receiving a feature contribution explanation, people may attempt to gauge the uncertainty of a model prediction by examining whether a few features that they believe as predictive contribute to the model’s prediction in the same direction or not, and they may also compare the direction of
each feature’s contribution with their own rationale to evaluate the correctness of the prediction [21]. Without these domain expertise, people may find themselves clueless to extract meaningful insights from the explanations.

**Implications for designing and selecting XAI methods.** In light of the ineffectiveness of existing XAI methods, better designs should be developed for those decision making contexts when people have limited knowledge in the task (e.g., recommend portfolios to beginning investors). A key challenge is how to construct and communicate the explanation in a manner that places reasonable cognitive load on the explanation consumers. To this end, techniques for presenting explanations visually, selectively, and progressively [54, 62, 69], and methods for incorporating the consideration of cognitive load into the explanation generation process [1] should be explored. Moreover, new approaches can be developed to increase people’s ability in making full use of the information carried in AI explanations. For example, for explanations like feature contribution and counterfactual examples, people could have been able to infer the model uncertainty even without any knowledge about the domain—they can sum up the contribution of all features and the base rate in a case (i.e., the closer it is to zero, the more uncertain the model), or they can count the number of counterfactual examples and compute the magnitude of difference between each counterfactual and the original data (i.e., the larger number of counterfactual examples and the smaller the difference, the more uncertain the model).

Our study also indicates that the three desiderata we have posited for AI explanations may each capture distinct aspects of people’s usage of AI explanations—satisfying one desideratum is not always sufficient for satisfying the other desiderata, and one explanation can score high on some desideratum but not the others. This is in line with previous findings that XAI methods that help people simulate an AI may not necessarily increase people’s decision accuracy [11]. Further studies are needed to systematically understand the relationships between these desiderata. Explanation providers should also carefully select the type of explanations to present to users based on the specific needs (e.g., whether to increase users’ comprehension of the model or enhance user’s decision making).

**Limitations.** Our study is limited by the particular formats of explanations we adopted (e.g., visual designs of feature contribution, the way we selected nearest neighbors), and the choice of the logistic regression model which is inherently simple. We caution the readers to not over-generalize our results to other settings. The desiderata we have proposed in this study are not comprehensive, and the effects of AI explanations may also be moderated by other factors such as the accuracy of the AI model. Future studies should be conducted to explore other aspects of the effects of AI explanations (e.g., influence user satisfaction), as well as carefully examine how these effects change with the moderating factors. Nevertheless, we hope our study provides a starting point for comparing the effectiveness of various XAI methods in AI-assisted decision making along concrete standards, and inspires more empirical studies to advance our knowledge of the strengths and weaknesses of different explanations. Towards obtaining a rigorous and comprehensive understanding of the effectiveness of various XAI methods, we recommend future researchers to evaluate XAI methods across decision making tasks with different characteristics, and to communicate results of any empirical evaluation of XAI methods along with sufficient contextual information on the properties of the decision making task.

**ACKNOWLEDGMENTS**

We are grateful to the 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**


