Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons

XINRU WANG and MING YIN, Purdue University, USA

Recent years have witnessed the growing literature in empirical evaluation of **explainable AI (XAI)** methods. This study contributes to this ongoing conversation by presenting a comparison on the effects of a set of established XAI methods in AI-assisted decision making. 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 three randomized controlled experiments, we evaluate whether four types of common model-agnostic explainable AI methods satisfy these properties on two types of AI models of varying levels of complexity, and in two kinds of decision making contexts where people perceive themselves as having different levels of domain expertise. Our results demonstrate that many AI explanations do not satisfy any of the desirable properties when used on decision making tasks that people have little domain expertise in. On decision making tasks that people are more knowledgeable, the feature contribution explanation is shown to satisfy more desiderata of AI explanations, even when the AI model is inherently complex. We conclude by discussing the implications of our study for improving the design of XAI methods to better support human decision making, and for advancing more rigorous empirical evaluation of XAI methods.

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

Additional Key Words and Phrases: Interpretable machine learning, explainable AI, trust, trust calibration, human-subject experiments

ACM Reference format:
Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Trans. Interact. Intell. Syst.* 12, 4, Article 27 (November 2022), 36 pages.
https://doi.org/10.1145/3519266

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

The reviewing of this article was managed by special issue associate editors Tracy Hammond, Bart Knijnenburg, John O'Donovan, and Paul Taele.
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.
Authors' address: X. Wang and M. Yin, Purdue University, 305 N. University St., West Lafayette, Indiana, USA, 47907; emails: {xinruw, mingyin}@purdue.edu.

2160-6455/2022/11-ART27
https://doi.org/10.1145/3519266

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 [70] and SHAP [55] 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 [13, 61].

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*? And how will the effectiveness of the explanation methods vary with the property of the AI model (if they are applicable to any AI models), such as the inherent complexity of the model? 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 [27, 68]. In responding to this call, there is recently a growing line of literature on empirically evaluating the effectiveness of XAI methods (e.g., [16, 18, 48, 81, 83]). 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 first positing three desirable properties that ideal AI explanations should satisfy in order to be considered helpful in AI-assisted decision making, and then presenting three randomized human-subject experiments which empirically compare to what extent established AI explanation methods satisfy these desiderata on different types of decision making tasks and different types of AI models.

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. We consider two decision making contexts in our experimental studies where the human decision makers may perceive themselves as having varying levels of domain expertise (i.e., recidivism prediction and forest cover prediction). We also apply the XAI methods to two types of AI models of varying levels of complexity (i.e, logistic regression model and multi-layer neural network model).

Our experimental results suggest that the effectiveness of different XAI methods largely depends on the properties of both the decision making task and the AI model. In particular, when decision aids are developed based on models of low complexity (e.g., logistic regression) and used on decision making tasks that people feel that they have some domain knowledge in (e.g., recidivism prediction), each of the XAI methods that we have examined is able to satisfy some desiderata, though different 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. On the other hand, when XAI methods are used for decision making tasks that people have limited domain knowledge in (e.g., forest cover prediction tasks), or when XAI methods are applied to explain models of high complexity (e.g., neural network models), the effectiveness of various XAI methods decreases substantially — on the forest cover prediction tasks, none of the three desiderata is reliably satisfied by any of the XAI methods that we have looked into, while for complex AI models, the only XAI method that satisfies some desiderata is to show each feature's contribution to the model's prediction. Interestingly, we find that the specific interpretability methods that are used to compute the contribution values of different features may also influence how effective the feature contribution explanation is when the AI model is inherently complex.

We note that this work is an extended version of Wang and Yin [78]. Our contributions in this work include:

- Through reviewing the existing literature of XAI research, we posit three principles as essential to effective AI explanations in AI-assisted decision making, and we identify the gap in existing empirical evaluations of XAI methods in the context of these three principles (Section 2).
- Via three randomized human-subject experiments, as detailed in Sections 3–5, we conduct a comprehensive comparison of several established model-agnostic XAI methods to study whether they satisfy the three principles of effective AI explanations. The type of decision making tasks and the type of AI models are varied across the experiments, which allows us to obtain a nuanced understanding of how the effectiveness of XAI methods are moderated by these two factors. In particular, the third experiment in this article (Section 5), in which we evaluate the effectiveness of various XAI methods when they are used to explain an inherently complex AI model, is a new contribution of this article compared to [78].
- Based on the results of our experimental studies, we provide discussions on the design and selection of effective XAI methods that are most suitable for the use case, as well as fair and transparent reporting of empirical evaluation results of XAI methods (Section 6).

2 LITERATURE REVIEW

2.1 Overview of AI Explanation Methods

Earlier literature on AI explanations often concerns the communication of uncertainty in AI decisions [52, 53]. 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, 27, 30]. 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 [17, 39, 50, 77], as well as visualizing what deep neural network has learned in its intermediate layers and how its predictions are affected by different parts of the inputs (e.g., through saliency map) [43, 72, 80]. On the other hand, typical model-agnostic methods include providing information on global-level feature importance [32], computing feature contribution on individual predictions [55, 70], using examples in the training dataset or counterfactual examples to explain model predictions [42, 45, 76], and conducting model distillation [10, 36].

2.2 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 [27, 68], and the "people" here can be different parties in the AI ecosystems including model developers, regulators, and end-users [74, 75]. 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 context.[1] Based on tasks that satisfy these criteria, we summarized three desiderata of AI explanations as follows:

- ***Desideratum 1 (Understanding)*:** Explanations of an AI model should improve people's understanding of it.
- ***Desideratum 2 (Uncertainty awareness)*:** Explanations of an AI model 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 model 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 asking people to rank the input data features based on their influence to overall predictions [21, 37], to indicate the direction of change in the model's prediction when a feature's value is altered [18, 21, 33], to simulate the model's predictions [18, 27, 47, 54, 68], to answer "what-if" questions about the model behavior [7, 18, 27, 37, 61], and to detect mistakes of the model and debug the model [68, 70].

Desideratum 2 connects to the needs of communicating the uncertainty inherent in AI model predictions to people [83]. 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 [81, 83]. Note that when an explanation simply improves the human-AI joint decision making accuracy, it does not necessarily mean this desideratum is satisfied. This is because people could trust an AI model inappropriately yet still achieve a higher level of decision accuracy (e.g., blindly trust a model which has a higher accuracy than oneself).

[1] An example of a task that may not be applicable to some decision making context is to have the human decision maker detect fairness problems of the AI model or utilize the AI model in a fair way [26, 34].

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)

Publications	Tasks	AI models	XAI methods	Desideratum 1 (Understanding)	Desideratum 2 (Uncertainty awareness)	Desideratum 3 (Trust calibration)
Poursabzi-Sangdeh et al. [68]	house price prediction	linear regression	intrinsically interpretable model	mixed results	N/A	✗?
Alqaraawi et al. [3]	image classification	neural network (CNN)	saliency map	mixed results	N/A	N/A
Chu et al. [20]	age prediction	neural network (CNN)	saliency map	N/A	N/A	✗?
Cheng et al. [18]	student admission	linear regression	feature contribution	✓	N/A	N/A
Lai and Tan [49]	deception detection	linear SVM	feature contribution	N/A	N/A	✓?
Cai et al. [15]	drawing recognition	neural network (RNN)	example-based	mixed results	N/A	N/A
Zhang et al. [83]	income prediction	gradient-boosted trees	feature contribution	N/A	✗	✗?
Carton et al. [16]	toxicity content detection	neural network (LSTM)	feature contribution	N/A	N/A	✗?
Lai et al. [48]	deception detection	SVM, neural network (BERT)	feature contribution	N/A	N/A	✓?
Yang et al. [81]	leaf classification	linear SVM	example-based	N/A	N/A	✓
Bansal et al. [6]	sentiment analysis	neural network (BERT)	feature contribution	N/A	N/A	✗
Kenny et al. [41]	digit recognition	neural network (CNN)	example-based	N/A	N/A	✓?

Note: "N/A" means the study does not examine the desideratum. ✓ (or ✗) means the study finds (or does not find) evidence suggesting the explanation method it examines satisfies a desideratum. In the ✓? (or ✗?) 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.

2.3 Empirical Studies on the Effectiveness of AI Explanations

A small but growing number of empirical studies have been recently carried out to evaluate whether and how various AI explanations can provide necessary assistance to human decision makers in their decision making. Table 1 shows a brief summary of the results of these studies with respect to whether different desiderata of AI explanations have been satisfied.

On the one hand, we find that the current empirical evaluation results are incomplete. Few studies explicitly examine Desideratum 2, and most studies touching upon Desideratum 3 only report the human-AI joint decision making accuracy, which is not sufficient to fully understand people's ability of calibrating their trust in the AI model. On the other hand, the results, overall, are quite mixed, which may be caused by many reasons. For example, different types of AI explanations may naturally show distinctive impact on human decision makers, so a systematic comparison of how well various state-of-the-art explanation methods can satisfy the desiderata of AI explanations is needed. Moreover, we note that the effect of explanations in AI-assisted decision making may also be moderated by the properties of the decision making task. For example, people may have different levels of domain expertise in the decision making task, which could potentially change the difficulty for them to understand the model explanation, or utilize the model explanation to infer about model uncertainty and correctness, and eventually influence the effectiveness of AI

explanations. Another factor that may have contributed to the mixed results is the complexity of the AI model. As more complex models tend to have the capacity to capture highly sophisticated patterns in the data, explanations of those complex models might not be as straightforward for people to process as explanations of simple models.

In light of the gap and limitations that we have identified in the existing literature, we present in this paper a comparative evaluation on how different XAI methods satisfy the desiderata when people are assisted by AI in making decisions on different types of tasks, and when these XAI methods are used on different types of AI models.

3 EXPERIMENT 1: RECIDIVISM PREDICTION, LOGISTIC REGRESSION

We set out to conduct experimental studies to gain in-depth understanding 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?

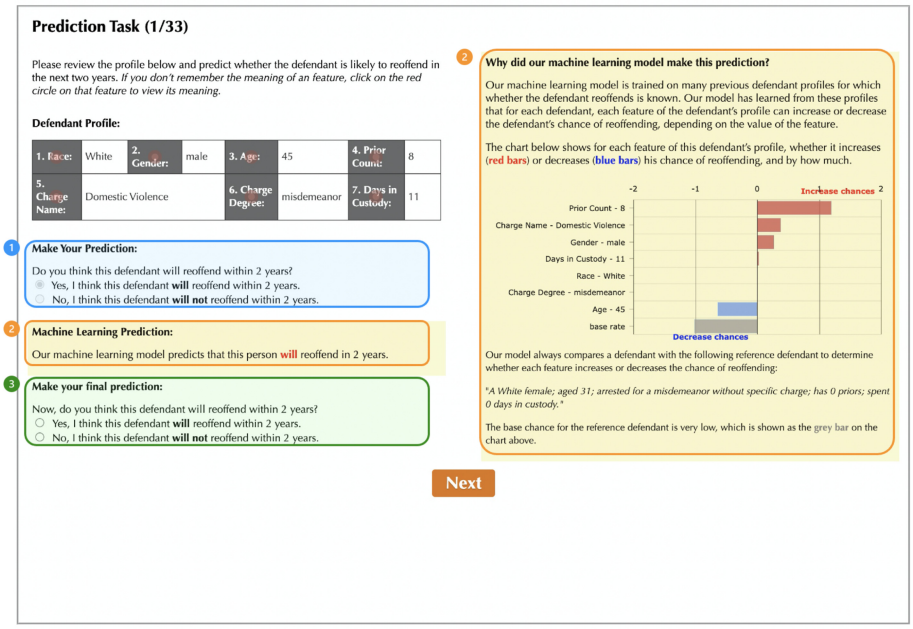
We focus on *model-agnostic* explanation methods in our studies, as these methods can be applied to any kind of AI models. To start, we conduct our first experiment on a decision making task that people may have some domain knowledge in (i.e., the recidivism prediction task), and we build a decision aid based on an AI model of low complexity (i.e., a logistic regression model) for this task. In our later experiments, we will vary the type of decision making task and the complexity of the AI model to explore the generalizability of our results, which we will detail in Sections 4 and 5.

3.1 Decision Making Task

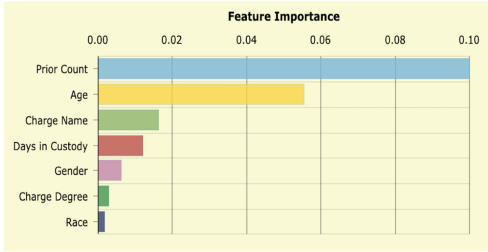
In this experiment, we asked participants to complete a sequence of recidivism prediction tasks with the help of a decision aid powered by a **machine learning (ML)** model. Specifically, in each task, participants were asked to review a profile of a criminal defendant consisting of seven features — 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 re-offend within two years. The defendant profiles were selected from the COMPAS dataset, which contained information of 7,214 criminal defendants in Broward County, Florida, USA, between 2013 and 2014 [4, 51]. This dataset was widely used by researchers to understand how people interact with machine assistance in their decision making [26, 35].

We built a decision aid based on a machine learning model to help people make these recidivism predictions. In particular, in each decision making task, the participant was asked to first review the profile of the defendant to make her own prediction. After the participant made this initial prediction, we would present to her the model's prediction, possibly together with some explanation on why the model made such prediction (see more detail in Section 3.2.2). Lastly, the participant needed to make a *final* prediction. Figure 1(a) shows an example of the task interface.

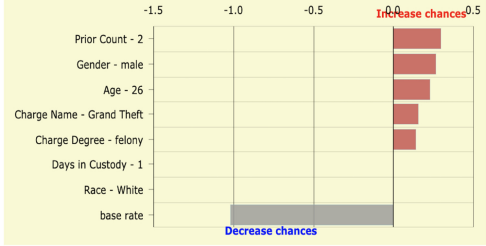
We chose the recidivism prediction task for our first experiment for two main reasons. First, recidivism prediction represents a realistic type of task that AI-based decision aids are developed to help both experts like judges and laypeople like jurors in decision making for social justice



(a) Task Interface for the treatment with the feature contribution explanation. Participants went through the 3-step procedure to make their recidivism predictions with the help of the ML model, and they were only presented with content in Steps 2 and 3 after making their initial predictions in Step 1.



(b) Explanation: Feature importance



(c) Explanation: Feature contribution

	Current defendant	Defendant A (same)	Defendant B (different)
Machine Learning Prediction	will reoffend	will reoffend	will not reoffend
1. Race:	White	Black	White
2. Gender:	male	male	male
3. Age:	26	26	26
4. Prior Count:	2	2	2
5. Charge Name:	Grand Theft	Grand Theft	arrest case no charge
6. Charge Degree:	felony	felony	felony
7. Days in Custody:	1	1	1

(d) Explanation: Nearest neighbors

For this defendant, our model would have made the opposite prediction (i.e., predict this defendant "will not reoffend") in the each of following cases:

- **Race**: If the defendant's Race had been **Hispanic** instead of White
- **Gender**: If the defendant's Gender had been **female** instead of male
- **Age**: If the defendant's Age had been **29** instead of 26
- **Prior Count**: If the defendant's Prior Count had been **1** instead of 2
- **Charge Name**: If the defendant's Charge Name had been **Driving with a Suspended License** instead of Grand Theft
- **Charge Degree**: If the defendant's Charge Degree had been **misdemeanor** instead of felony

In contrast, changing the value for each of the following features while keeping other features unchanged would not make our model predict differently:

- Days in Custody

(e) Explanation: Counterfactuals

Fig. 1. Examples of the recidivism prediction task interface and the four types of model explanations that we showed to participants in Experiment 1.

[9, 22, 44]. Thus, the need of communicating AI recommendations to decision makers properly to promote high-quality decision making is pressing for tasks like this. Second, we conjecture that people may perceive themselves as having a degree of domain expertise in solving the recidivism prediction task, because they can apply their day-to-day, common sense knowledge in their predictions. In fact, earlier research suggested that the accuracy of laypeople in predicting recidivism is similar to that of commercial risk assessment software like COMPAS [29]. In addition, we found that when being asked about how a defendant's features relate to that defendant's chance of re-offending in a pilot study, people exhibited highly consistent beliefs on the relationship between some features and the re-offending risks (e.g., 91.5% of participants in our pilot study believed defendants with a larger number of non-juvenile prior crimes are more likely to re-offend). This again supports our conjecture that people may consider themselves as having some knowledge about how to make recidivism predictions.

3.2 Experimental Design

3.2.1 Machine Learning Model. In this experiment, we trained a *logistic regression* model using a subset of the original COMPAS dataset, and this model was used as the underlying machine learning model of the decision aid that predicts the likelihood of a defendant re-offending. We chose the logistic regression model in this experiment because it has a relatively low level of complexity and is generally considered as intrinsically interpretable. On the hold-out test dataset consisting of 1,000 task instances with 47.7% of the instances associated with a positive label (i.e., the defendant in the task instance reoffended within two years), the accuracy of our logistic regression model is 69.1% and the AUC is 0.749. This level of model performance is comparable with previous predictive models trained on the COMPAS dataset [38, 51].

3.2.2 Experimental Treatments. We adopted a between-subject design in our experiment. We created five 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 [32] 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 the model we used in this experiment was a logistic regression model, 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 rate[2] (Figure 1(c)).
- **Nearest neighbors:** In this treatment, we explained the model's prediction to participants by showing to them the 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*

[2]The base rate is the log odds value for a hypothetical profile in which the value of each feature takes the reference level.

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, 26].[3] 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.

3.2.3 Selection of Task Instances. Participants in different treatments worked on the *same* set of 32 task instances. To better answer our research question **RQ1**–**RQ3**, we carefully selected these 32 task instances from the hold-out test dataset. In particular, we categorized the machine learning model's confidence on a task instance as high or low depending on whether the model's probability estimate of the *predicted label* (which is in the range of $[0.5, 1]$) is higher than 0.7, and we confirmed this probability aligned well with the model's correctness likelihood on each instance.[4] 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 (i.e., number of non-juvenile prior crimes and age), 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) [42, 62].

3.3 Experimental Procedure

We conducted our experiment 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 five 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.

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's prediction (and possibly 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.

[3]While the format of feature contribution explanation we used in this experiment was specific to logistic regression models, explaining model predictions by showing the contribution of each feature is applicable for other models [55, 70].

[4]We used 0.7 as the threshold because previous studies suggest that people tend to perceive probabilities higher than 70% as at least "likely" [65], and it leads to two similar-sized subsets of the hold-out test dataset (35.6% of the task instance in the hold-out test dataset is associated with a confidence that is higher than 0.7).

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. 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 $0.03 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 experiment was $2.26.

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 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 model is assessed in existing literature (see Section 2), we designed a set of nine 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 x_2 instead of x_1, 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 Appendix A. Moreover, we asked participants to report their own perceived understanding of the model by indicating their agreement on the following two statements (adapted from earlier literature [15, 18]) from 1 ("strongly disagree") to 7 ("strongly agree"): (1) I understand how the model works to predict whether a defendant will re-offend; and (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 model (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 [82, 83], 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.

Finally, for **RQ3**, we evaluated participants' capability of calibrating their trust in the ML model using three main dependent variables, including their *appropriate trust* [58–60, 63] (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 the participants' final decision accuracy), *overtrust* [25, 66] (i.e., the fraction of tasks where participants used a model's prediction when it was wrong) and *undertrust* [25, 66] (i.e., the fraction of tasks where participants did not use a model's prediction when it was correct). 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 and/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 [24, 28]. That is, we first 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 ($R = 5000$). Then, we interpreted our results based on the range of confidence intervals, and measured the effect sizes using Cohen's d [23].

To take the impact of covarietes (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.

3.5 Experimental Results

In total, 1,062 participants completed our experiment HIT. After filtering the data from 280 participants who did not pass our attention check, we were left with valid data from 782 participants (62.9% male, the average age is 38). We analyzed these data to answer our research questions.

3.5.1 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). 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).

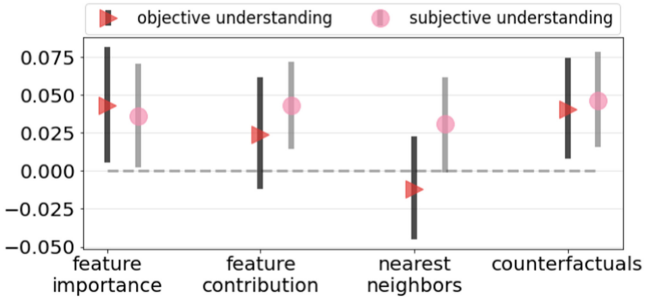


Fig. 2. Comparing how different types of explanations change participants' objective and subjective understanding of the model compared to when no model explanation is provided in Experiment 1. Error bars represent 95% bootstrap confidence intervals.

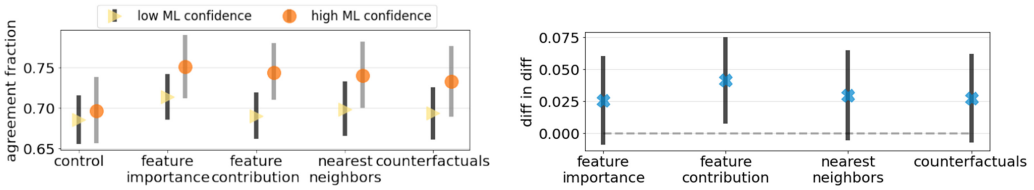
Further, we constructed mixed effect regression models to predict the correctness of a participant's answer on each multiple-choice question or a participant's rating on each subjective understanding survey question. We treated the type of explanation a participant received as the fixed effect, the participant as the random effect, and controlled for the participant's age, gender, and education as covariates.[5] Our regression results were consistent with what we have observed in Figure 2. For example, we found that 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: $\beta = 0.04[0.02, 0.06]$, feature contribution: $\beta = 0.04[0.03, 0.06]$, nearest neighbors: $\beta = 0.03[0.01, 0.05]$, counterfactuals: $\beta = 0.05[0.03, 0.07]$). We also found that 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.

Put together, for participants in our Experiment 1 who make recidivism predictions with the help of a simple, logistic regression model, it is shown that all four types of model explanations that we have examined can improve their understanding of the model to some extent. Notably, the feature importance explanation and the counterfactual explanation not only increase participants' subjective perceptions of understanding, but also result in increases in their objective understanding of the model behavior.

3.5.2 RQ2: Effects on Recognizing Model Uncertainty. We now move on to **RQ2** to examine how the presence of different model explanations affects people's ability to tell apart high confidence model predictions from low confidence model predictions. Figure 3(a) compares participants' reliance on the model (as measured by the agreement fraction) on tasks where the model has high confidence and tasks where the model has low confidence. Visually, it appears that in the control treatment, when participants were assisted by a logistic regression model to make recidivism predictions without seeing the model explanations, they did not rely on the model's high confidence predictions and low confidence predictions much differently. The provision of different types of model explanations, however, nudged participants into relying on the model's high confidence predictions more than the model's low confidence predictions.

Such observation becomes more clear when we look into the *difference in difference* — the difference in participants' reliance on high vs. low confidence model predictions in a treatment with

[5]We 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.



(a) Reliance (b) Reliance (difference in difference)

Fig. 3. Comparing how different types of explanations change participants' capability of recognizing model confidence in Experiment 1. (a): Reliance on high/low confidence model predictions, for participants in different treatments. (b): The difference of participants' reliance on high vs. low confidence model predictions in each treatment with a model explanation, compared against such difference in the control treatment. Error bars represent 95% bootstrap confidence intervals.

model explanation, minus the difference in participants' reliance on high vs. low confidence model predictions in the control treatment. We plot our estimated difference in difference values as well as the 95% bootstrap confidence intervals in Figure 3(b). We found that all four types of model explanations — especially the feature contribution explanation — seem to enable participants to rely on the model's high vs. low confidence predictions to a much more different extent than those participants who did not receive any model explanations (e.g., Cohen's $d = 0.20$, 95% CI $[-0.02, 0.42]$ when aggregating all explanation types).

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 concluded that 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: $\beta = 0.17[0.02, 0.31]$). These results, thus, confirm that in this experiment, participants might have utilized model explanations to infer model uncertainty, which allows them to rely on high confidence model predictions more, and this is true regardless of the type of explanation methods used.

3.5.3 RQ3: Effects on Trust Calibration. Finally, we look into how different explanations influence people's capability of calibrating their trust in the AI model. We measured participants' appropriate trust, overtrust, and undertrust in the model for each treatment. The difference in the mean values of these measures between a treatment with model explanation and the control treatment, as well as their 95% bootstrap confidence intervals, are shown in Figure 4.

Overall, our results suggest that both the feature importance and feature contribution explanation appear to help participants slightly increase their appropriate trust (Cohen's $d = 0.19[-0.05, 0.41]$ for feature importance, and $d = 0.19[-0.03, 0.40]$ for feature contribution) and decrease their undertrust (feature importance: $d = -0.21[-0.44, 0.02]$, feature contribution: $d = -0.15[-0.37, 0.06]$) in the model, although for participants receiving the feature importance explanation, this seems to be achieved at the price of a slight increase of overtrust in the model (Cohen's $d = 0.15[-0.08, 0.36]$). On the other hand, the effects of nearest neighbors and counterfactual explanations in influencing participants' trust calibration were inconclusive. Taking a closer look at the data by examining how model explanations affect trust calibration on tasks

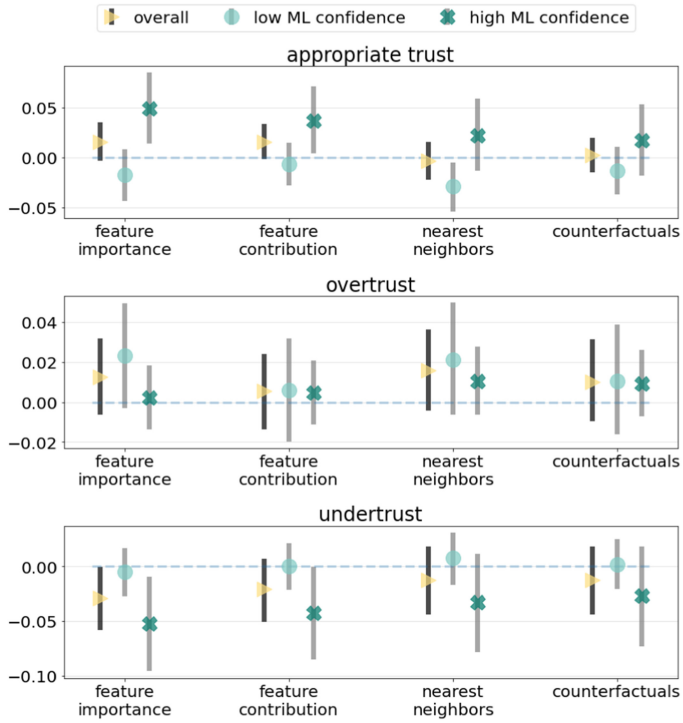


Fig. 4. Comparing how different types of explanation support participants' trust calibration in the AI in Experiment 1. For appropriate trust, the larger the value the better. For overtrust and undertrust, the smaller the value the better.

where the model has high or low confidence separately, we found that both the feature importance and feature contribution explanations support participants' trust calibration on high confidence model predictions (e.g., for appropriate trust, feature importance: $d = 0.30[0.07, 0.53]$, feature contribution: $d = 0.23[0.02, 0.45]$), but the feature importance explanation also results in a slight increase of participants' overtrust on the model's low confidence predictions ($d = 0.19[-0.03, 0.42]$).

Similar as before, we again built mixed effect models to predict whether a participant could trust the model appropriately on each task, whether she would over-trust the model on tasks that the model was wrong, and whether she would under-trust the model on tasks that the model was correct. The type of explanation the participant received was included as the fixed effect, and the participant was the random effect. Again, we found that only the feature contribution explanation increases participants' appropriate trust *without* incurring a higher level of overtrust or undertrust (estimated coefficients β for feature contribution — appropriate trust: $0.01[-0.003, 0.03]$, undertrust: $-0.03[-0.05, -0.01]$, and overtrust: not reliably different from 0). Interestingly, participants who reported to have a higher level of education consistently showed a lower level of appropriate trust, but a higher level of undertrust in the model.

To briefly summarize, when participants make recidivism predictions with the help of a logistic regression model, the only type of model explanation that could help participants increase their appropriate trust in the model without resulting in an increase of overtrust or undertrust in the model is the feature contribution explanation. Moreover, the effect of the feature contribution explanation in promoting trust calibration is particularly salient on those tasks where the model makes high confidence predictions.

4 EXPERIMENT 2: FOREST COVER PREDICTION, LOGISTIC REGRESSION

In our second experiment, we are interested in exploring that in AI-assisted decision making, whether and how the effectiveness of various AI explanation methods varies with the properties of decision making tasks involved, such as the level of domain expertise people usually have in the tasks. Therefore, in this experiment, we replicate Experiment 1 on a different type of decision making task to examine whether the effects of various AI explanation methods change with the task type.

4.1 Decision Making Task

The decision making task we used in our second experiment was the forest cover prediction task. Specifically, in this task, participants were shown a geological profile of a wilderness area (in a $30m \times 30m$ cell) containing eight 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 were selected from the UCI cover type dataset [8, 31], 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 six 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. Again, participants first needed to review the geological profile and make their own prediction. Then, they would be presented with the machine learning model's prediction before they could submit a final prediction in the task. Figure 5 shows the interface of this task.

We chose the forest cover prediction tasks in our second experiment for two main reasons. First, such task reflects realistic use cases of AI-driven decision aids, as machine learning models have been developed to assist people in making better decisions in forest management [56, 57]. Second, we speculate that compared to the recidivism prediction task, most people may perceive themselves as having a lower level of domain expertise in the forest cover prediction task. To confirm this intuition, we conducted a pilot study, in which we introduced both the recidivism prediction task and the forest cover prediction task to participants that we recruited from 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 make more accurate predictions, and they would be more confident about their predictions. Among 98 MTurk workers who participated in this pilot study, 82.6% of them reported themselves to be more knowledgeable on the recidivism prediction tasks, 63.3% (or 71.4%) of them believed they (or a normal person) can make more accurate predictions for the recidivism prediction tasks, and 71.4% of them felt they would be more confident in making recidivism predictions. In other words, consistent with our conjecture, most laypeople perceived themselves as lacking domain expertise in the forest cover prediction task, compared to the recidivism prediction task.

4.2 Experimental Design and Procedure

We followed the same experimental design as that in Experiment 1 (see Section 3.2). In particular, we again trained a *logistic regression* model based on a subset of the UCI cover type dataset to help people make forest cover predictions. On the hold-out test dataset consisting of 1,000 task instances with 43.0% of the instances associated with a positive label (i.e., the geographical area in

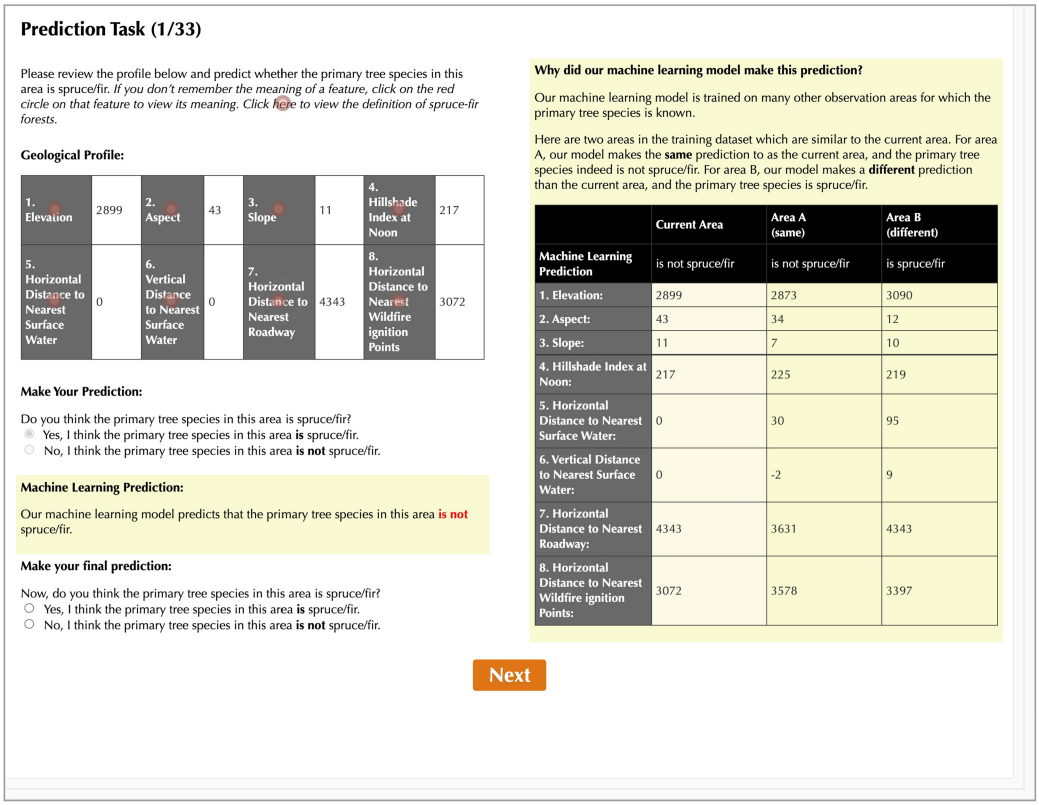


Fig. 5. An example of the forest cover prediction task interface (with the nearest neighbors explanation) that we showed to participants in Experiment 2. Participants needed to first make their initial prediction before they were presented with the model prediction (and possibly explanation), and were asked to make the final prediction.

the task instance is primarily covered by spruce-fir forests), the accuracy of our logistic regression model is 69.5% and the AUC is 0.686; this level of model performance is reasonable and is similar to the performance of the logistic regression model that we used in Experiment 1 for the recidivism prediction task.

We again conducted our experiment by recruiting participants from MTurk, and participants who had previously participated in Experiment 1 were not allowed to participate in this experiment. Regarding the experimental procedure, in addition to following the same procedure of Experiment 1, as that explained in Section 3.3, we further added a training component in the tutorial of the experiment to help participants get familiar with the forest cover prediction task — 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 the machine learning model, and they learned about the correct answer after each task. Such training component was adapted from those used in previous studies when asking participants to work on tasks that they may have limited expertise in [81, 83]. In addition, we increased the base payment of the experiment to $2.00 to account for the longer periods of time that participants needed to spend on this experiment due to the addition of the training tasks.

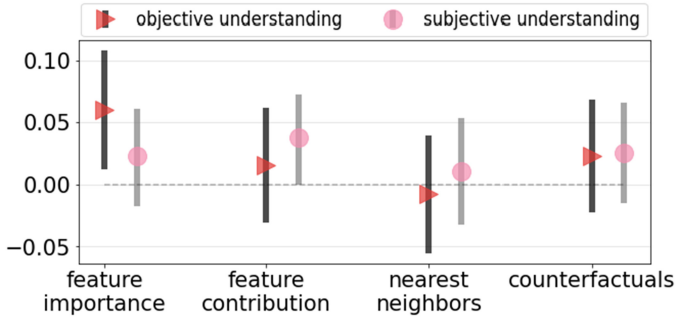


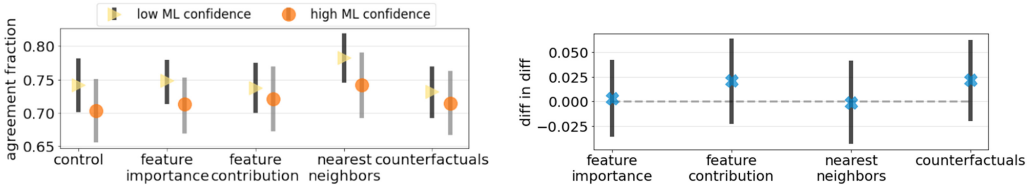
Fig. 6. Comparing how different types of explanations change participants' objective and subjective understanding of the model compared to when no model explanation is provided in Experiment 2. Error bars represent 95% bootstrap confidence intervals.

4.3 Experimental Results

After filtering data from 147 participants who did not answer the attention check questions correctly, we obtained valid data from 561 participants (64.3% male, the average age is 39). We analyzed these valid data, following the methods that we have described in Section 3.4.

4.3.1 RQ1: Effects on Understanding AI Models. Figure 6 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). It appears that when participants made forest cover predictions with the help of a logistic regression model, the improvement in participants' understanding of the model brought up by the provision of model explanations was more limited compared to when participants made recidivism predictions. Indeed, 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]). The results of our mixed effect regression models were consistent with what we have observed in Figure 6. More specifically, we found that in Experiment 2, 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. Similar to that in Experiment 1, we again found that females 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.

4.3.2 RQ2: Effects on Recognizing Model Uncertainty. Moving on to examine how the presence of different model explanations affects people's ability to recognize the uncertainty of a model's predictions on the forest cover prediction task, Figure 7(a) shows participants' reliance on the model on tasks where the model has high confidence and tasks where the model has low confidence. In addition, Figure 7(b) shows the difference in difference in reliance when using participants' reliance difference on high vs. low confidence model predictions in the control treatment as the reference. Here, we found that participants working on the forest cover prediction task did not seem to be affected by the model explanations in adjusting how much they would rely on the model differently based on the model confidence. Analyzing the data again using the mixed effect regression models to account for various covariates, we still found that the coefficients for all of



(a) Reliance (b) Reliance (difference in difference)

Fig. 7. Comparing how different types of explanations change participants' capability of recognizing model confidence in Experiment 2. (a): Reliance on high/low confidence model predictions, for participants in different treatments. (b): The difference of participants' reliance on high vs. low confidence model predictions in each treatment with a model explanation, compared against such difference in the control treatment. Error bars represent 95% bootstrap confidence intervals.

the four interaction terms between model explanation type and model confidence scores were not reliably different from zero. This means that participants making predictions on forest cover did *not* seem to act upon model predictions with varying levels of confidence differently in the presence of model explanations. In other words, various types of model explanations fail to enable participants to recognize the model's uncertainty on different tasks in Experiment 2.

4.3.3 RQ3: Effects on Trust Calibration. Finally, we look into how different explanations affect people's trust calibration. The mean value differences of participants' appropriate trust, overtrust, and undertrust in the model between a treatment with model explanation and the control treatment, as well as their 95% bootstrap confidence intervals, are shown in Figure 8. Inspecting Figure 8, we concluded that none of the model explanations helps improve participants' trust calibration in the AI model for the forest cover prediction task, regardless of the model's confidence in its predictions. We again built mixed effect models to predict whether a participant could trust the model appropriately on each task, and whether she would over-trust (under-trust) the model on tasks that the model was wrong (correct), and the results confirmed that none of the explanations supports participants to calibrate their trust in the model. We also noticed that participants who reported to have a higher level of education consistently showed a lower level of appropriate trust and a higher level of undertrust in the model, similar to what we've observed in Experiment 1.

5 EXPERIMENT 3: RECIDIVISM PREDICTION, DEEP NEURAL NETWORK

In our last experiment, we aim to explore whether and how the effectiveness of various AI explanation methods varies with the properties of the AI model used, such as the model's inherent complexity. Therefore, in this experiment, we replicate Experiment 1 and again study the effects of AI explanation methods on the recidivism prediction task, but we replace the machine learning model used in the experiment to a different type.

5.1 Experimental Design and Procedure

5.1.1 Machine Learning Model. In this experiment, we trained a *multi-layer neural network* model based on the COMPAS dataset to help participants make recidivism predictions. Specifically, the model was trained using the Python scikit-learn package [12, 67], with the solver for weight optimization set to be "lbfgs" and the L2 penalty set to be 0.00001. We conducted a 5-fold cross-validation on the training dataset to identify the best hyper-parameter values for the neural network (e.g., the number of hidden layers and the number of neurons per hidden layer) using grid search. Eventually, the optimal architecture we ended up with contained 10 hidden layers with 50 neurons in each layer. On the hold-out test dataset consisting of 1,000 task instances (the same

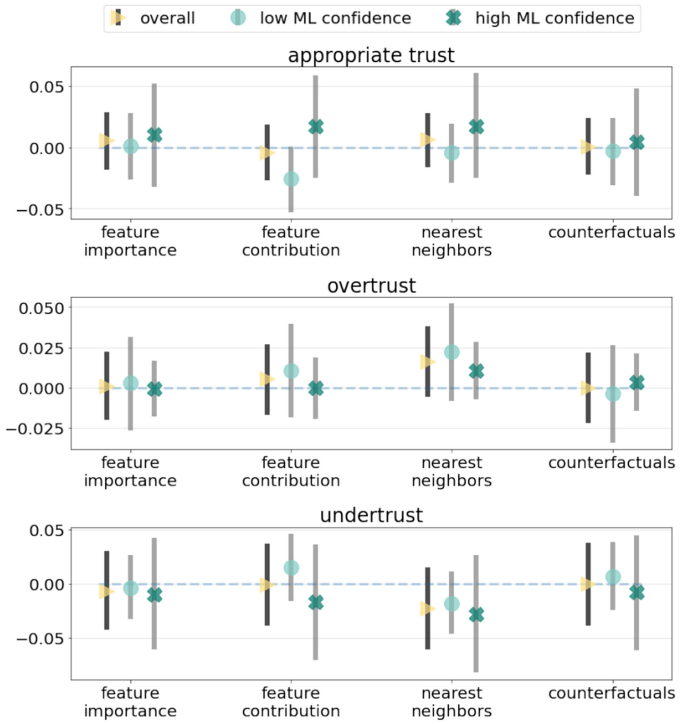
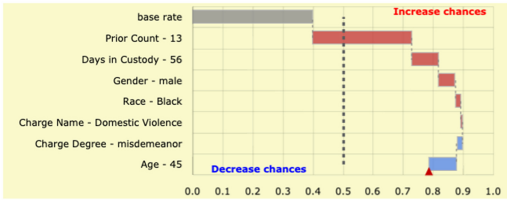


Fig. 8. Comparing how different types of explanation support participants' trust calibration in the AI in Experiment 2. For appropriate trust, the larger the value the better. For overtrust and undertrust, the smaller the value the better.

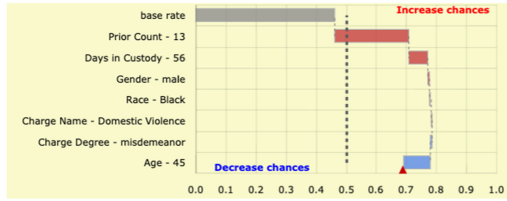
hold-out dataset as that described in Section 3.2.1), the accuracy of this neural network model is 68.3% and the AUC is 0.761. In contrast to the logistic regression model that we used in Experiment 1, this neural network model represents a class of black-box models with high level of complexity that is not intrinsically interpretable. Indeed, the large number of hidden nodes and layers used in this neural network implies a much more complex relationship between the input features and output predictions, compared to the linear relationship expressed by the logistic regression model between the input features and log-odds of the binary predictions.

5.1.2 Experimental Treatments. We adopted almost exactly the same experimental treatments as those used in Experiments 1 and 2 (see Section 3.2.2), except for that we included two treatments with the feature contribution explanation instead of one. In these two treatments, we adopted two state-of-the-art model-agnostic interpretability methods to compute the contribution that each feature makes to the neural network model's prediction on a task instance:

- **Feature contribution – LIME:** In this treatment, we explained the model's prediction to participants by showing to them the contribution of each feature to the prediction, and these contribution scores were computed based on the LIME algorithm [70]. Specifically, LIME trains an interpretable model such as a Lasso regression model at the neighborhood around the data instance of interest, in order to locally approximate the decision boundary of the black-box model, and then explains each feature's contribution to the model's prediction based on this local surrogate model. For each task, we first followed the LIME algorithm to compute each feature's contribution to the neural network model's prediction on the task,



(a) Feature contribution – LIME



(b) Feature contribution – SHAP

Fig. 9. Examples of the two types of feature contribution explanations that we showed to participants for the same task instance in Experiment 3.

and then provided a waterfall chart to visualize the contributions of all features as well as the base rate[6] (Figure 9(a)).

- **Feature contribution – SHAP:** In this treatment, we explained the model's prediction to participants by showing to them the contribution of each feature to the prediction, and these contribution scores were computed based on the SHAP algorithm [55]. SHAP computes the contribution of each feature to a model's prediction based on the Shapley values, a concept from coalitional game theory that enables a "fair" distribution of payoffs among players. For each task, we first followed the SHAP algorithm to compute each feature's contribution to the neural network model's prediction on the task, and then provided a waterfall chart to visualize the contributions of all features and the base rate[7] (Figure 9(b)).

Note that the feature contribution values generated by LIME and SHAP algorithms should be interpreted in the probability space, instead of in the log-odds space as that for the logistic regression model. We thus chose to use waterfall charts to visualize the feature contribution explanations in this experiment, and we informed participants that the length of the bar associated with a feature reflects how much that feature increases or decreases the model's estimated probability in the defendant re-offending. The LIME and SHAP explanations for a task can be different, though. In particular, summing up feature contribution scores in a SHAP explanation leads to the exact predicted probability of the defendant re-offending as that given by the neural network model, whereas summing up feature contribution scores in a LIME explanation only arrives at a approximation of the predicted probability.

5.1.3 Experimental Procedure. Other than that the underlying machine learning model for the decision aid is switched from a logistic regression model to a multi-layer neural network model, and participants who previously participated in Experiment 1 or 2 were excluded from participating in this experiment, the procedure of this experiment is exactly the same as that of Experiment 1 (see Section 3.3).

5.2 Experimental Results

After filtering data from 115 participants who did not answer the attention check questions correctly, we obtained valid data from 665 participants (65.1% male, the average age is 38), and we again followed the methods as described in Section 3.4 to analyze these data.

5.2.1 RQ1: Effects on Understanding AI Models. Figure 10 shows the impact of different types of explanations on participants' understanding of the machine learning model. Even though we

[6]The base rate in LIME is the re-offending chance that the machine learning model predicts for a defendant when it does not know any feature information about the defendant.
[7]The base rate in SHAP is the average re-offending chance that the machine learning model predicts for all instances in the training dataset.

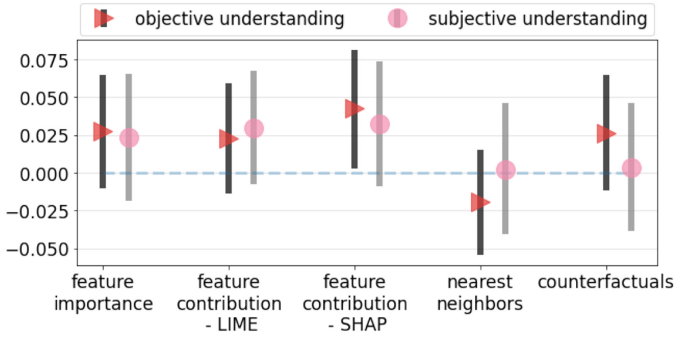


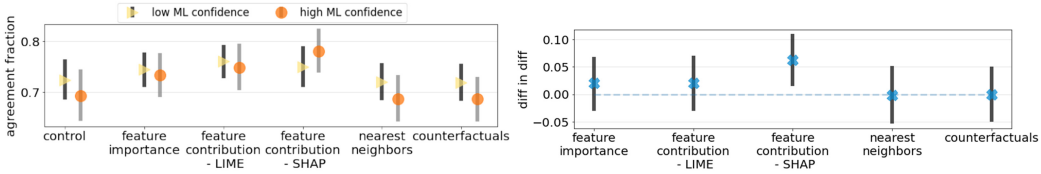
Fig. 10. Comparing how different types of explanations change participants' objective and subjective understanding of the model compared to when no model explanation is provided in Experiment 3. Error bars represent 95% bootstrap confidence intervals.

still asked our participants to make recidivism predictions in this experiment, we observed that increasing the complexity of the underlying machine learning model of the decision aid seems to affect the effectiveness of different explanation methods in influencing people's understanding of the model.

In terms of objective understanding, we found only the SHAP feature contribution explanation brought about an overall positive effect (Cohen's d = 0.30, 95% CI [0.02, 0.57])[8]. For subjective understanding, it appears that only the two feature contribution explanations could slightly help increase participants' perceptions on how much they understand the model (Cohen's d = 0.22, 95% CI [−0.05, 0.50] for LIME, and 0.22 [−0.06, 0.50] for SHAP). When taking the impact of participants' demographics into consideration, the results of the mixed effect regression models again suggested the same story — Among all explanation methods, only the SHAP feature contribution explanation slightly improves participants' objective understanding of the model ($\beta = 0.03[-0.004, 0.07]$), while the LIME and SHAP feature contribution explanation could both help increase participants' subjective understanding of the model (LIME: $\beta = 0.03[0.01, 0.05]$, SHAP: $\beta = 0.03[0.01, 0.06]$) Aligning with our previous findings, We again found that females 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.

5.2.2 RQ2: Effects on Recognizing Model Uncertainty. Figure 11(a) compares participants' reliance on the neural network model on the model's high confidence predictions and low confidence predictions, and Figure 11(b) displays the estimated difference in difference of reliance using participants' behavior in the control treatment as the reference. Visually, it is clear that only when the SHAP feature contribution explanation was presented, participants could effectively differentiate the neural network model's high confidence predictions from its low confidence predictions and therefore relied on the high confidence predictions more (Cohen's d = 0.37, 95% CI [0.08, 0.64]). This is confirmed by our further analysis using mixed effect regression models, which showed that the only type of model explanation that had a reliably positive interaction with model confidence in influencing participants' willingness to rely on the model was the SHAP feature contribution explanation ($\beta = 0.36[0.18, 0.53]$).

[8]A closer look at the data suggests that the SHAP feature contribution explanation mainly helps increase participants' ability to simulate model behavior. More details on how different types of explanations change each aspect of participants' objective understanding of the neural network model (e.g., their ability to simulate model behavior, ability to detect model errors) can be found in Appendix B.



(a) Reliance (b) Reliance (difference in difference)

Fig. 11. Comparing how different types of explanations change participants' capability of recognizing model confidence in Experiment 3. (a): Reliance on high/low confidence model predictions, for participants in different treatments. (b): The difference of participants' reliance on high vs. low confidence model predictions in each treatment with a model explanation, compared against such difference in the control treatment. Error bars represent 95% bootstrap confidence intervals.

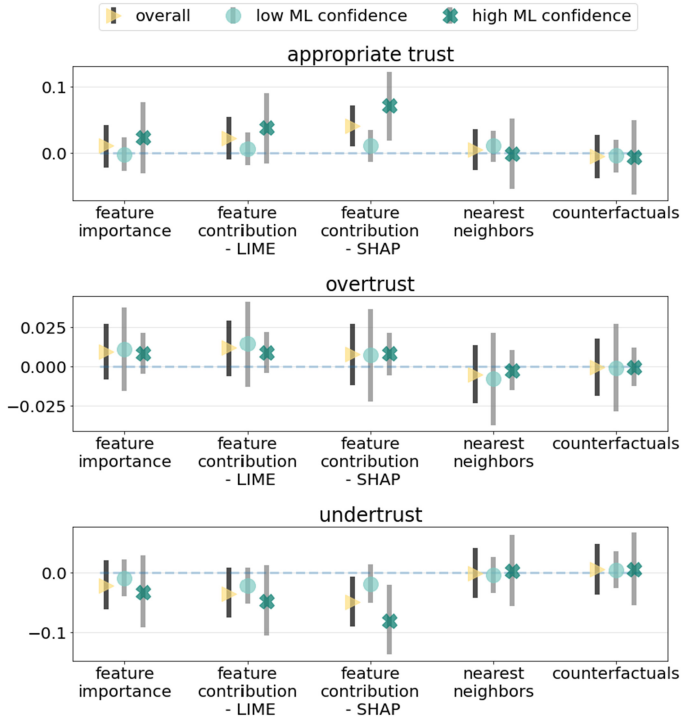


Fig. 12. Comparing how different types of explanation support participants' trust calibration in the AI in Experiment 3. For appropriate trust, the larger the value the better. For overtrust and undertrust, the smaller the value the better.

5.2.3 RQ3: Effects on Trust Calibration. Figure 12 compares participants' appropriate trust, overtrust, and undertrust in the model between a treatment with model explanation and the control treatment.

Overall, when assisted by the neural network model, participants in the SHAP feature contribution explanation treatment exhibited more calibrated trust in the AI model, as their appropriate trust increased (Cohen's $d = 0.38[0.07, 0.65]$) and undertrust decreased (Cohen's $d = -0.33[-0.61, -0.01]$). The LIME feature contribution explanation had a similar, yet less conclusive effect. In addition, we found that the impact of the SHAP explanation on supporting participants'

Table 2. Summary of the Effects of AI Explanations on Different Types of Decision Making Tasks, when People are Assisted by a Decision aid that is based on a Logistic Regression Model

	Recidivism prediction (Experiment 1)			Forest cover prediction (Experiment 2)		
Explanation type	**Understanding**	**Uncertainty Awareness**	**Trust Calibration**	**Understanding**	**Uncertainty Awareness**	**Trust Calibration**
feature importance	✓	✓	✗	✓?	✗	✗
feature contribution	✓?	✓	✓	✓?	✗	✗
nearest neighbor	✓?	✓	✗	✗	✗	✗
counterfactuals	✓	✓	✗	✗	✗	✗

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).

Table 3. Summary of the Effects of AI Explanations for Different Types of AI Models, when People Make Recidivism Predictions with the Assistance from AI

	Logistic regression (Experiment 1)			Multi-layer neural network (Experiment 3)		
Explanation type	**Understanding**	**Uncertainty Awareness**	**Trust Calibration**	**Understanding**	**Uncertainty Awareness**	**Trust Calibration**
feature importance	✓	✓	✗	✗	✗	✗
feature contribution	✓?	✓	✓			
- LIME				✓?	✗	✗
- SHAP				✓	✓	✓
nearest neighbor	✓?	✓	✗	✗	✗	✗
counterfactuals	✓	✓	✗	✗	✗	✗

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).

trust calibration was mostly shown on those tasks where the model's confidence in its prediction was high (appropriate trust: $d = 0.39[0.09, 0.67]$, undertrust: $d = -0.39[-0.67, -0.07]$).

Similar as before, we again built mixed effect models to predict whether a participant could trust the model appropriately on each task, and whether she would over-trust (under-trust) the model on tasks that the model was wrong (correct). The regression results suggested that the SHAP feature contribution explanation increased participants' appropriate trust *without* incurring a higher level of overtrust or undertrust (estimated coefficients β for SHAP — appropriate trust: $0.03[0.01, 0.06]$, undertrust: $-0.07[-0.09, -0.05]$, and overtrust: not reliably different from 0), while LIME explanation helped decrease participants' undertrust but at the price of inducing a higher level of overtrust (undertrust: $\beta = -0.05[-0.07, -0.02]$, overtrust: $0.04[0.004, 0.08]$). In addition, providing feature importance explanation also appeared to help participants decrease their undertrust ($\beta = -0.03[-0.05, -0.005]$), but its effects on influencing participants' appropriate trust and overtrust in the model were inconclusive. Still, participants who reported to have a higher level of education consistently showed a lower level of appropriate trust, but a higher level of undertrust in the model.

6 DISCUSSION

We summarize our experimental results in Tables 2 and 3, and we highlight a few key findings:

- The effects of AI explanations are dramatically different on tasks where people have varying levels of domain expertise in, and when applied to AI models of differing levels of complexity.

In particular, when established XAI methods are used on decision making tasks that people are not knowledgeable about or on AI models of high complexity, most of them do not reliably satisfy any of the three desiderata.
- The only positive effect of model explanation that we have consistently observed across all three experiments is that feature contribution explanations increase people's subjective understanding of an AI model.
- Among the four 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, regardless of the complexity of the AI model. However, when the AI model is inherently complex, the specific algorithms used to compute the contribution values of different features may also influence the effectiveness of the feature contribution explanation.
- The two example-based explanations in our study seem to satisfy least desiderata of AI explanations. Notably, for the counterfactual explanation, which is considered to closely resemble how human explain decisions, it is shown that its impact on improving people's understanding of the AI model and increasing people's awareness of the model uncertainty is significantly weakened as the decision making task requires more domain expertise or the AI model becomes more complex, and its influence on promoting calibrated trust in AI is minimal.

In the following, we provide possible explanations of our results, and discuss implications and limitations of our study.

6.1 The Role of Domain Expertise in Moderating the Effects of AI Explanations

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 could be because without the domain knowledge that is learned from their day-to-day working and social experience and may have become part of the subconscious mind [40, 79], people have to process all the new information (i.e., the AI explanations) in their working memory, which takes up more cognitive capacity [64]. 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 [83]. Without these domain expertise, people may find themselves clueless to extract meaningful insights from the explanations.

6.2 The Role of AI Model Complexity in Moderating the Effects of AI Explanations

While model-agnostic explainable AI methods can be applied to any type of AI models, our study demonstrates that the effectiveness of these methods can be highly dependent on the properties of the AI models such as the model's inherent complexity — when the AI model has high complexity, most AI explanations become less effective as the number of desiderata that they can satisfy

decreases. This finding is consistent with observations that have been made in previous literature [48].

We suspect that one reason to explain why we get this finding is that complex models tend to generate nonlinear and highly complex decision boundaries, which makes it more challenging for people to make sense of the explanations of these complex models. People may find the explanation of a complex AI model on some task instance contradicts with their own intuition on how features of a task "should" relate to the prediction on that task. For example, our pilot study suggests that the majority of people believe defendants' risks of re-offending increase with the number of days the defendants have spent in custody. However, in our Experiment 3, there exists one defendant whose *Days in Custody* is 82 and the neural network model predicts he will re-offend, but the counterfactual explanation suggests that the model would have predicted that he will not re-offend if his value on the feature *Days in Custody* was 124. Upon observing such explanation that is contradictory to their own intuition, people may feel confused and find themselves unable to understand why the model captures this counterintuitive relationship, and they may further doubt the correctness of the model, despite the model's recommendation may not necessarily be wrong.

The highly complex decision boundary of sophisticated AI models also implies that the explanations of the model may not be consistent across different task instances. For example, in our Experiment 3, when the LIME algorithm is used to compute each feature's contribution to the neural network model's prediction on a defendant, within the set of Hispanic defendants we show to our participants, the contribution of the defendant's race (i.e., "Hispanic") ranges from decreasing the defendant's estimated chance of re-offending by 3.2% to increasing the estimated chance of re-offending by 1.1%. This inconsistency of the explanations may add an additional layer of difficulty for people to see how and why the model makes predictions based on different rationales in different regions of the feature space. On the other hand, the feature importance explanation, which is a consistent explanation by design (i.e., it is a global explanation and will not vary across task instances), is also not very effective when used for explaining complex models. This could potentially be caused by the fact that the global-level feature importance does not always align with the model's sophisticated behavior on individual cases.

6.3 Understanding the Difference between the Effectiveness of LIME and SHAP

LIME and SHAP are both established techniques that explain the prediction of an AI model by computing the contribution of each feature to the prediction, yet results of our Experiment 3 indicate that under our experimental settings, the feature contribution explanation based on SHAP can satisfy more of the desiderata compared to the feature contribution explanation based on LIME. Here, we provide two possible accounts for this finding.

First, given a black-box model, LIME computes feature contribution values for a data instance by generating perturbed training data within the local neighborhood of this instance and learning an interpretable model on the weighted version of these perturbed data. This interpretable model, thus, is only an approximation of the original black-box model at the local neighborhood, and does not necessarily produce the same prediction as the black-box model. In fact, researchers have found that the random perturbation process employed by LIME for generating local training data may result in considerable data and label shift, which could eventually decrease the fidelity of the local interpretable model for explaining the black-box model [69]. Indeed, for the neural network model that we have used in Experiment 3, we find that the mean absolute difference between the neural network model's predicted chance of a defendant re-offending and the chance estimated by the local interpretable model of LIME on the same defendant (which is the sum of all feature's

contributions and the base rate) is 4.8% across all 32 task instances that we show to participants, with the largest difference being 13.9%.[9] Moreover, when using a threshold of 0.7 to differentiate high confidence model predictions from the low confidence ones (as we have described in Section 3.2.3), we find that on three of the 32 task instances, the local model of LIME makes high (low) confidence predictions while the original neural network model makes low (high) confidence predictions.[10] We conjecture that the misalignment in the predicted probability between the actual model and the local interpretable model of LIME may have hampered people's capability in accurately recognizing the model's uncertainty and adjusting their trust in the model appropriately. Note, this is less of an issue for the feature contribution explanations produced by the SHAP algorithm, because by design, SHAP is guaranteed to produce the same probability estimate as the model to be explained.

Second, compared to the feature contribution explanations produced by the LIME algorithm, we find the explanations produced by the SHAP algorithm seem to have a higher level of consistency across different task instances, at least on the set of 32 task instances that we show to our participants in Experiment 3. For example, when using SHAP to compute feature contributions, for all categorical features (e.g., race, sex, name of the currently charged crime), the contribution values associated with a specific feature value (e.g., the value of "Hispanic" for the feature "race") always take the same sign, suggesting that setting the categorical feature to a particular value tend to always change the model's prediction to the same direction. In addition, the base rate of re-offending given by the SHAP algorithm is always the same for any task instance, which is the average re-offending chance for all task instances in the model's training dataset. In contrast, the LIME algorithm has different base rate prediction for different task instances, because LIME fits separate local models for each individual instance. It is thus possible that people find the more consistent explanation to be easier to understand and utilize, leading to a higher level of effectiveness of the SHAP explanation compared to the LIME explanation.

Finally, in Section 5.2.1, we found that both the LIME and SHAP explanation lead to higher levels of subjective understanding of the model among participants. However, for those participants who received the LIME explanation, their objective understanding of the model was not reliably improved in general. We conjecture that this could be caused by the fact that the LIME and SHAP explanation share similar formats (i.e., they both explain the model's prediction by specifying how much each feature contributes to the prediction), as people may generally find such format of explanation to be informative and accessible. This means that those receiving the LIME explanation may think they understand the model more than they actually do, or in other words, suffer from the *illusion of explanatory depth* [71], perhaps partly because the fidelity and internal consistency of the LIME explanation is imperfect. This could be a potentially harmful scenario in reality, which highlights the importance of conducting more future studies to explore when and why an illusion of explanatory depth occurs [19], and how to carefully design and present XAI methods like LIME to avoid such a scenario.

[9]On the 1,000 task instance in the hold-out test dataset, the mean absolute difference between the neural network model's predicted chance of a defendant re-offending and the chance estimated by the local interpretable model of LIME on the same defendant is 5.5%, with the largest difference being 37.7%.

[10]When looking into the 1,000 task instances in the hold-out test dataset, we find for 75 instances the local model of LIME makes high (low) confidence predictions when the original neural network model makes low (high) confidence predictions. We even find that on 72 task instances, the predicted probability of re-offending given by the local model of LIME would result in an opposite binary prediction compared to the predicted probability of the neural network model. On each of the 32 task instances that we show to participants in Experiment 3, the binary prediction of the neural network model and the local model of LIME are always the same, though.

6.4 Implications for Designing and Selecting XAI Methods

In light of the ineffectiveness of existing XAI methods, better explanations should be designed for those decision making contexts when people have limited knowledge in the task (e.g., recommending portfolios to beginning investors), and for those cases when the XAI methods are applied to inherently complex models (e.g., gradient boosting machines, deep neural networks).

Regarding designing more effective XAI methods for decision making tasks that require a high level of domain expertise, 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 [61, 73, 81], 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 without relying on their domain knowledge. 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 simply sum up the contribution of all features and the base rate in a case and then compare the sum to some threshold (i.e., if the sum indicates the probability of a binary prediction, use 0.5 as the threshold; the closer the sum is to the threshold, 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).

With respect to increasing the effectiveness of XAI methods when explaining complex AI models, we speculate that an important step to take is to go beyond just explaining what causes the model's prediction on each local data point. Instead, more *context* and *justifications* should be provided along with these explanations. For example, when the model captures a counterintuitive relationship between features and predictions, in addition to illustrating such a relationship to people, further information can be supplied to help people obtain deeper understandings on when and why such counterintuitive relationship exists (e.g., when fixing the value of feature B, does the counterintuitive relationship between feature A and the prediction still exist?).

Moreover, if a complex model's explanations show a degree of inconsistency across different task instances, it could be useful to give people a big picture of the range of applicability of different explanations (e.g., when does the value of Hispanic on the race feature result in increases in predicted re-offending risk and when does it result in decreases, and why?). To this end, combining global and local explanations to explain complex AI models may turn out to be a more effective approach. Similar recommendations have been previously made by other researchers, resulting in many innovative model exploration interfaces that incorporate both model-level (global) and instance-level (local) explanations to flexibly support the user's various needs in understanding the model behavior [14, 37, 46, 70]. Compared to providing only global or local explanations, combining both types of explanations may enable users to not only probe a high-level overview of the model but also drill down into detailed investigation on specific instances. Moreover, it may also help users get a sense of the partition of the feature space on which the model operates based on different rationales, so that they can avoid over-generalize their understandings of the model from one instance to another. In the future, more empirical studies need to be carried out to examine the effectiveness of those interfaces which combine global and local explanations of AI models, and to identify the best practices in combining these two types of explanations.

Finally, 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 use cases (e.g., whether the goal is to increase users' comprehension of the model or enhance user's decision making, what type of decision making task is involved, and what kind of AI models are used).

6.5 Limitations

Our study is limited by the particular formats of explanations we have adopted (e.g., visual designs of feature contribution, the way we select nearest neighbors), as well as the particular types of decision making tasks and AI models that we have considered. We caution the readers to not overgeneralize our results to other settings. The desiderata we have proposed in this study are not comprehensive, and future studies could be conducted to explore other aspects of the effects of AI explanations (e.g., influence user's satisfaction). In addition, the effects of AI explanations may also be moderated by other factors such as the accuracy of the AI model. Thus, investigating how these effects change with additional moderating factors is another important direction to explore.

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 on AI models with different properties. Importantly, results of any empirical evaluation of XAI methods should also be communicated along with sufficient contextual information on the properties of the decision making task, as well as the properties of the AI model.

7 CONCLUSION

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 model, recognize the uncertainty underlying the AI model, and calibrate their trust in the AI model in AI-assisted decision making. We further conduct three 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, and on two types of AI models which have various inherent complexity. We find that overall, the effectiveness of most XAI methods decreases when used on decision making tasks that people lack domain expertise in or on complex AI models. In particular, on tasks that people have little domain expertise in, none of the four AI explanations we have examined reliably satisfy any of the three desiderata. On tasks that people perceive themselves as more knowledgeable, our results provide evidence supporting that the feature contribution explanation has the potential to satisfy most of the desiderata, even when the AI model is inherently complex.

APPENDICES

A OBJECTIVE UNDERSTANDING QUESTIONS (EXPERIMENT 1)

The full list of multi-choice questions we used for evaluating participants' objective understanding of the model behavior in Experiment 1 is included in this appendix. In each question, we highlight the correct answer in bold text.

A.1 Feature Importance

Question 1: Among the following features, which one is the *most* important in influencing our machine learning model's prediction (that is, variations in the value of that feature will *most likely* change the model's prediction)?

A. **Age**
B. Charge Name
C. Charge Degree

Question 2: Among the following features, which one is the *least* important in influencing our machine learning model's prediction (that is, variations in the value of that feature will *most unlikely* change the model's prediction)?

A. Days in Custody
B. Sex
C. **Charge Degree**

A.2 Marginal Effect of Features on Predictions

Question 3: Consider a defendant with the following profile:

1. Race	White	2. Sex	female	3. Age	31	4. Prior Count	0
5. Charge Name	Arrest case no charge			6. Detailed Charge Degree	misde-meanor	7. Days in custody	0

When all other features are kept the same,

(1) *If the defendant's age was 41 instead of 31*, how would the machine learning model's prediction on the defendant's likelihood of reoffending change?
   A. The model would predict the 41-year-old defendant to be *more* likely to reoffend.
   B. **The model would predict the 41-year-old defendant to be *less* likely to reoffend.**
(2) *If the number of days the defendant spent in custody was 100 instead of 0*, how would the machine learning model's prediction on the defendant's likelihood of reoffending change?
   A. **The model would predict the defendant spending 100 days in the custody to be *more* likely to reoffend.**
   B. The model would predict the defendant spending 100 days in the custody to be *less* likely to reoffend.
(3) *If the defendant's charge name is "Driving under the Influence" instead of "Arrest with no case"*, how would the machine learning model's prediction on the defendant's likelihood of reoffending change?
   A. The model would predict the defendant who was charged with "Driving under the Influence" to be *more* likely to reoffend
   B. **The model would predict the defendant who was charged with "Driving under the influence" to be *less* likely to reoffend**
(4) *If the defendant's race is Hispanic instead of white*, how would the machine learning model's prediction on the defendant's likelihood of reoffending change?
   A. The model would predict the defendant whose race is Hispanic to be *more* likely to reoffend
   B. **The model would predict the defendant whose race is Hispanic to be *less* likely to reoffend**

(5) *If the defendant is a male instead of a female*, how would the machine learning model's prediction on the defendant's likelihood of reoffending change?
   A. **The model would predict the male defendant to be *more* likely to reoffend.**
   B. The model would predict the male defendant to be *less* likely to reoffend.

A.3 Counterfactual Thinking

Question 4: Consider a defendant with the following profile:

1. Race	White	2. Sex	male	3. Age	22	4. Prior Count	3
5. Charge Name	Possession of Cocaine			6. Detailed Charge Degree	felony	7. Days in custody	10

Our machine learning model currently predicts this defendant *will* reoffend. When all other features are kept the same, which of the following changes on the crime charge is **most likely** to change our model's prediction (i.e., make the model predict the defendant *will not* reoffend)?

A. **Change the charge name to "Driving Under the Influence"**
B. Change the charge name to "Driving with a Suspended License"
C. Change the charge name to "Battery"

Question 5: Consider a defendant with the following profile:

1. Race	Black	2. Sex	female	3. Age	24	4. Prior Count	2
5. Charge Name	Grand Theft			6. Detailed Charge Degree	felony	7. Days in custody	117

Our machine learning model currently predicts this defendant *will* reoffend. If we change only one feature of this profile but leave all other features unchanged, which of the following changes is **going** to change our model's prediction (i.e., make the model predict the defendant *will not* reoffend)? Please check all that apply.

A. **Change Race from Black to Hispanic**
B. Change Sex from female to male
C. **Change Age from 24 to 29**
D. **Change Priors Count from 2 to 0**
E. **Change Charge Name from "Grand Theft" to "Driving Under the Influence"**
F. **Change Days in Custody from 117 to 9**

A.4 Simulate Model Behavior

Question 6: Consider a defendant with the following profile:

1. Race	White	2. Sex	male	3. Age	26	4. Prior Count	2
5. Charge Name	Driving Under the Influence			6. Detailed Charge Degree	felony	7. Days in custody	1

What do you think **our machine learning model will predict** for this defendant?

A. The model will predict this defendant will reoffend within two years
B. **The model will predict this defendant will not reoffend within two years**

Question 7: Consider three defendants with the following profiles:
Defendant 1:

1. Race	Hispanic	2. Sex	male	3. Age	41	4. Prior Count	1
5. Charge Name	Driving Under the Influence			6. Detailed Charge Degree	misde-meanor	7. Days in custody	1

Defendant 2:

1. Race	White	2. Sex	male	3. Age	24	4. Prior Count	3
5. Charge Name	Driving Under the Influence			6. Detailed Charge Degree	misde-meanor	7. Days in custody	1

Defendant 3:

1. Race	White	2. Sex	male	3. Age	25	4. Prior Count	3
5. Charge Name	Driving with a Suspended License			6. Detailed Charge Degree	misde-meanor	7. Days in custody	172

For one of these three defendants, **our machine learning model predicts** that the defendant will reoffend. Which one do you think is this defendant?

A. Defendant 1
B. Defendant 2
C. **Defendant 3**

A.5 Error Detection

Question 8: Consider a defendant with the following profile:

1. Race	White	2. Sex	male	3. Age	22	4. Prior Count	0
5. Charge Name	Possession of Cocaine			6. Detailed Charge Degree	felony	7. Days in custody	1

Our machine learning model predicts that this defendant will reoffend [and also gives its explanation on the right side]. Do you believe this prediction is correct? (w/ or w/o ML explanation)

A. Yes, I think this prediction is correct
B. **No, I think this prediction is wrong**

Question 9: Consider a defendant with the following profile:

1. Race	Black	2. Sex	male	3. Age	52	4. Prior Count	7
5. Charge Name	Grand Theft			6. Detailed Charge Degree	misde-meanor	7. Days in custody	1

Our machine learning model predicts that this defendant will not reoffend [and also gives its explanation on the right side]. Do you believe this prediction is correct? (w/ or w/o ML explanation)

A. Yes, I think this prediction is correct
B. **No, I think this prediction is wrong**

B EFFECTS OF MODEL EXPLANATIONS ON INFLUENCING EACH ASPECT OF OBJECTIVE UNDERSTANDING OF THE MODEL (EXPERIMENT 3)

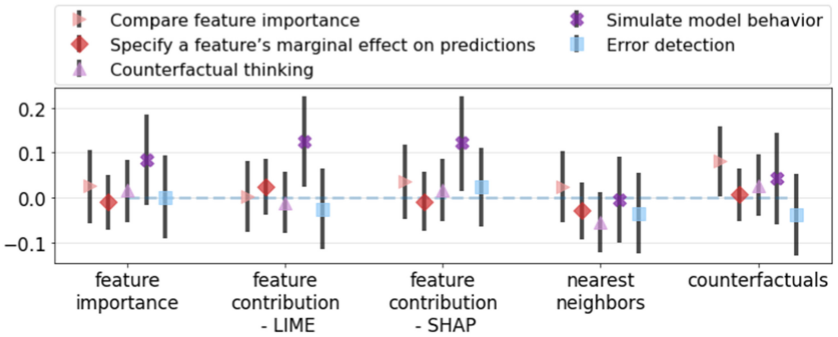


Fig. B.1. Comparing how different types of explanations change the five aspects of participants' objective understanding of the neural network model compared to when no model explanation is provided in Experiment 3. Error bars represent 95% bootstrap confidence intervals.

Since the multi-layer neural network model we used in Experiment 3 was inherently more difficult to understand, it is interesting to explore whether and how the provision of different model explanations changes participants' objective understanding of the model from different perspectives. To do so, we computed a participant's normalized score in each objective understanding survey component (i.e., compare feature importance, specify a feature's marginal effect on predictions, etc.) for each of the six experimental treatments in Experiment 3. In Figure B.1, for each component of objective understanding, we show the average changes of the normalized score between each treatment with a specific type of model explanation and the control treatment, along with the 95% bootstrap confidence intervals.

According to the figure, we find that when compared with the control treatment, the provision of the SHAP feature contribution explanations mainly helps increase participants' objective understanding of the model by increasing their capability in simulating the model behavior. In addition, the LIME feature contribution explanation can also increase participants' ability to simulate the model behavior, and the counterfactual explanation is shown to increase participants' ability in better understanding feature importance for the model, despite they both could not improve participants' overall objective understanding of the model when considering all components together. In contrast, we also notice that none of the explanation seems to affect participants' capability

in specifying a feature's marginal effect on the model's predictions, conducting counterfactual thinking, or detecting model errors.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers who provided many helpful comments.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: A user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. ProPublica. See https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[6] 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.

[7] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[8] Jock A. Blackard and Denis J. Dean. 1999. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture* 24, 3 (1999), 131–151.

[9] Marcus T. Boccaccini, Darrel B. Turner, Daniel C. Murrie, Craig E. Henderson, and Caroline Chevalier. 2013. Do scores from risk measures matter to jurors? *Psychology, Public Policy, and Law* 19, 2 (2013), 259.

[10] Cristian BuciluǍ, 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.

[11] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[12] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: Experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

[13] Ruth M. J. Byrne. 2019. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *IJCAI*. 6276–6282.

[14] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 46–56.

[15] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 258–262.

[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] 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.

[19] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think I get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.
[20] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction. arXiv:2007.12248 [cs.LG].
[21] DAC Collaris, Leo M. Vink, and J. J. van Wijk. 2018. Instance-level explanations for fraud detection: A case study. (2018).
[22] Cindy C. Cottle, Ria J. Lee, and Kirk Heilbrun. 2001. The prediction of criminal recidivism in juveniles: A meta-analysis. *Criminal Justice and Behavior* 28, 3 (2001), 367–394.
[23] Geoff Cumming. 2013. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Routledge.
[24] Geoff Cumming. 2014. The new statistics: Why and how. *Psychological Science* 25, 1 (2014), 7–29.
[25] Ewart J. de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. 2014. A design methodology for trust cue calibration in cognitive agents. In *International Conference on Virtual, Augmented and Mixed Reality*. Springer, 251–262.
[26] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285.
[27] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *Stat* 1050 (2017), 2.
[28] Pierre Dragicevic. 2016. Fair statistical communication in HCI. In *Modern Statistical Methods for HCI*. Springer, 291–330.
[29] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580.
[30] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.
[31] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.
[32] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* 20, 177 (2019), 1–81.
[33] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics* 24, 1 (2015), 44–65.
[34] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[35] 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.
[36] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Stat* 1050 (2015), 9.
[37] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[38] Northpointe Inc. 2012. COMPAS Risk & Need Assessment System. http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf.
[39] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2017. Simple rules for complex decisions. *Available at SSRN 2919024* (2017).
[40] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Macmillan.
[41] Eoin M. Kenny, Courtney Ford, Molly Quinn, and Mark T. Keane. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence* 294 (2021), 103459.
[42] 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.
[43] 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.
[44] John Logan Koepke and David G. Robinson. 2018. Danger ahead: Risk assessment and the future of bail reform. *Wash. L. Rev.* 93 (2018), 1725.
[45] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. (2017), 1885–1894.
[46] Josua Krause, Adam Perer, and Enrico Bertini. 2018. A user study on the effect of aggregating explanations for interpreting machine learning models. In *ACM KDD Workshop on Interactive Data Exploration and Analytics*.

[47] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*.
[48] 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.
[49] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.
[50] 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.
[51] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9, 1 (2016).
[52] Brian Y. Lim and Anind K. Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th International Conference on Ubiquitous Computing*. 415–424.
[53] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2119–2128.
[54] Zachary C. Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.
[55] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.
[56] Duncan Macmichael and Dong Si. 2017. Addressing forest management challenges by refining tree cover type classification with machine learning models. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 177–183.
[57] Duncan MacMichael and Dong Si. 2018. Machine learning classification of tree cover type and application to forest management. *International Journal of Multimedia Data Engineering and Management (IJMDEM)* 9, 1 (2018), 1–21.
[58] Maranda McBride and Shona Morgan. 2010. Trust calibration for automated decision aids. *Institute for Homeland Security Solutions*. [Online]. Available: https://www.ihssnc.org/portals/0/Documents/VIMSDocuments/McBride_Research_Brief.pdf.
[59] John M. McGuirl and Nadine B. Sarter. 2006. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors* 48, 4 (2006), 656–665.
[60] Stephanie M. Merritt, Deborah Lee, Jennifer L. Unnerstall, and Kelli Huber. 2015. Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors* 57, 1 (2015), 34–47.
[61] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.
[62] Christoph Molnar. 2019. *Interpretable Machine Learning*.
[63] Bonnie M. Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-machine Studies* 27, 5–6 (1987), 527–539.
[64] Barbara A. Oakley. 2014. *A Mind for Numbers: How to Excel at Math and Science (Even if you Flunked Algebra)*. Tarcher-Perigee.
[65] Bernie J. O'Brien. 1989. Words or numbers? The evaluation of probability expressions in general practice. *The Journal of the Royal College of General Practitioners* 39, 320 (1989), 98–100.
[66] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39, 2 (1997), 230–253.
[67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[68] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. (2021), 1–52.
[69] Amir Hossein Akhavan Rahnama and Henrik Boström. 2019. A study of data and label shift in the LIME framework. *arXiv preprint arXiv:1910.14421*.
[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] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* 26, 5 (2002), 521–562.

[72] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
[73] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: Empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 107–120.
[74] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*.
[75] Jennifer Wortman Vaughan and Hanna Wallach. 2020. A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence* (2020).
[76] 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.
[77] Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In *Artificial Intelligence and Statistics*. 1013–1022.
[78] 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.
[79] Sheri Lynn Warren. 2016. Make it stick: The science of successful learning. *Education Review* 23 (2016).
[80] 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.
[81] 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.
[82] 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.
[83] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. 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.

Received 29 July 2021; revised 16 December 2021; accepted 15 February 2022