Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models

Chun-Wei Chiang
chiang80@purdue.edu
Purdue University
West Lafayette, Indiana, USA

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

ABSTRACT

Today, machine learning (ML) technologies have penetrated almost every aspect of people's lives, yet public understandings of these technologies are often limited. This highlights the urgent need of designing effective methods to increase people's machine learning literacy, as the lack of relevant knowledge may result in people's inappropriate usage of machine learning technologies. In this paper, we focus on an ML-assisted decision-making setting and conduct a human-subject randomized experiment to explore how providing different types of user tutorials as the machine learning literacy interventions can influence laypeople's reliance on ML models, on both in-distribution and out-of-distribution examples. We vary the existence, interactivity and scope of the user tutorial across different treatments in our experiment. Our results show that user tutorials, when presented in appropriate forms, can help some people rely on ML models more appropriately. For example, for those individuals who have relatively high ability in solving the decision-making task themselves, receiving a user tutorial that is interactive and addresses the specific ML model to be used allows them to reduce their over-reliance on the ML model when they could outperform the model. In contrast, low-performing individuals' reliance on the ML model is not affected by the presence or the type of user tutorial. Finally, we also find that people perceive the interactive tutorial to be more understandable and slightly more useful. We conclude by discussing the design implications of our study.

CCS CONCEPTS

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

KEYWORDS

Machine learning, human-AI interaction, appropriate reliance, user education, AI literacy

ACM Reference Format:
Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3490099.3511121

IUI '22, March 22–25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9144-3/22/03.
https://doi.org/10.1145/3490099.3511121

1 INTRODUCTION

In recent years, decision aids driven by machine learning (ML) models have become increasingly ubiquitous in diverse domains, from entertainment [8] to medical diagnosis [19]. As a result, today, a growing population is assisted by ML technologies in making better decisions, both at home and at work. On the other hand, despite their great success in uncovering hidden insights from massive data to enhance decision making, the current ML technologies have their own limitations, such as their tendency to reflect and reinforce existing biases and discrimination in the society [3, 11], their potential to pick up spurious correlations rather than causal relationships [12, 49], and their poor performance in generalizing to the external data [43, 69]. Unfortunately, those people who are assisted by ML-driven decision aids often lack sufficient understandings of the ML technologies and are not fully aware of the limitations of ML. Even worse, the lack of ML-related knowledge sometimes results in people's inappropriate usage of the ML technologies, such as interacting with the ML model's decision recommendations in a way that further increases the decision disparities among different demographic groups [26, 53], or overly relying on an ML model even when dataset shift occurs and the model's performance substantially degrades [13].

In light of this, researchers and practitioners alike have recently advocated for increasing people's literacy in machine learning and AI [41]. To this end, educators have proposed a diverse set of guidelines for effectively teaching machine learning and AI to children in K-12 classrooms [17, 32, 58, 72, 73], with the goal of increasing the AI literacy of the next generation in long term. However, for the large number of laypeople who have already utilized ML models—which are often commercial—in their decision making today, how can we improve their machine learning literacy? For example, are there any short-term ML literacy interventions that can be provided to them to raise their awareness of the limitations of ML and influence how they interact with ML models? And how do the designs of these interventions change their effectiveness in promoting appropriate usage of ML models and impact people's perceptions of them?

In this work, we provide some initial answers to these questions through an experimental study. We envision that the developers of the commercial ML models could provide users with a machine learning literacy intervention, in the form of a *user tutorial*, right before people's use of the ML model in decision making. The emphasis of this tutorial is to inform users of the development and

evaluation procedure of the ML model and warn them about the ML model's limitations. In this study, we focus on designing tutorials that communicate to users one particular limitation of ML, that is, the ML model's possible performance disparities on different data.

Specifically, we considered two design variables for the user tutorial. The first variable is the *interactivity* of the tutorial (i.e., static vs. interactive), with the interactive tutorial providing people with a sandbox environment in which they can construct customized testing datasets to evaluate the ML model's performance on different data. The second variable is the *scope* of the ML model discussed in the tutorial (i.e., general vs. specific)—while directly providing information in the tutorial about the specific commercial ML model that people will use in their decision making is the ideal, sometimes model developers are constrained in doing so due to various concerns (e.g., confidential/proprietary information). Thus, it's interesting to explore the alternative way of providing information about general ML models in the tutorial and see whether people can apply the learned knowledge about ML models in general to their specific use contexts. Corresponding to these two variables, we designed $2 \times 2 = 4$ versions of user tutorials. We recruited lay human subjects from Amazon Mechanical Turk to complete a set of house price prediction tasks with the help of an ML model, and we varied the presence and the type of user tutorial in different experimental treatments. We examined how the provisions and designs (i.e., tutorial scope and interactivity) of different types of user tutorials affect subjects' reliance on the ML model, both on tasks that come from the same distribution as the model's training data (i.e., "in-distribution examples") and on tasks that come from a different distribution than the model's training data (i.e., "out-of-distribution examples"). We further looked into whether the reliance changes brought up by the user tutorial are appropriate or not. Finally, we also explored how subjects' perceptions of the user tutorial vary with the designs of the tutorial.

Our results show that some ML literacy interventions that we designed in this study can indeed change subjects' reliance on the ML model. For example, overall, subjects who received the interactive user tutorial slightly decreased their reliance on the ML model on out-of-distribution examples compared to subjects who received the static version of the user tutorial, while there was an interaction effect between the scope and interactivity of the user tutorial in influencing subjects' reliance on the model on in-distribution examples. Interestingly, a closer look into the data suggests that the impacts of ML literacy interventions on people's reliance on the ML model are different for subjects with varying levels of decision-making performance themselves. For those high-performing subjects who are relatively capable in solving the decision-making tasks themselves, the provision of most types of user tutorials led to a reduction of reliance on the ML model on out-of-distribution examples.

In particular, when the tutorial addressed the specific ML model to be used and contained interactive components, high-performing subjects' reduction of reliance on the model implied a more appropriate level of reliance when they themselves outperformed the ML model. In contrast, whether and how user tutorials were provided to low-performing subjects did not affect how much they relied on the ML model on either in-distribution or out-of-distribution examples. Finally, we also found that subjects perceived interactive user tutorials to be significantly more understandable and marginally more useful, compared to the static ones.

Taken together, our findings bring insights into how to influence people's interactions with the ML models and help people rely on ML models more appropriately through designing short-term ML literacy interventions. We conclude by providing the design implications of our results, discussing the limitations of our study, and highlighting future opportunities in promoting appropriate usage of ML models through advancing people's ML literacy.

2 RELATED WORK

With the rapid development of ML-driven decision aids, a growing body of empirical studies have been conducted to understand what factors will affect people's reliance on ML models. For example, various factors related to the properties of the ML model, such as the ML model's accuracy [48, 66, 68], confidence [48, 70], and the level of agreement between the ML model and one's own judgment or reasoning [42, 71], are shown to affect how much a person will be willing to rely on an ML model. People's reliance on an ML model will also be heavily shaped by their interaction experience with the model, including how good their first impression of the model is [56], whether they have witnessed the model make mistakes [16], what the type of mistakes the model makes are [51], when the mistake happens [34], and whether the model's behavior aligns with their mental model of the model [6]. In addition, contextual factors such as the cultural background of a person [14] and the domain of the prediction tasks [29] can also affect people's trust in and reliance on the model.

Together with a deeper understanding of possible factors that will affect people's reliance on ML models, empirical studies also reveal more evidence suggesting that people often rely on ML models inappropriately. For example, a few studies have shown that people tend to reject recommendations provided by an ML model even when the model outperforms human predictions [16, 65], leading to the phenomenon of "algorithm aversion." Suresh et al. [55] showed that the presence of information about an ML system's training data, model architecture, and performance all results in people's increased blind trust in the system. More recently, it was found that people are more likely to delegate the decision-making right to an ML model when covariate shift occurs and the model operates poorly in a novel environment, compared to when the model operates in a static environment [13].

One common approach used for promoting more appropriate reliance on ML models is to provide explanations of model predictions [2, 33, 47, 61, 63], though these attempts show mixed success. In particular, Liu et al. [39] explored the effects of providing interactive local explanations on people' reliance on ML models, and they found that these explanations do not help people to rely on the ML models more appropriately on either in-distribution or out-of-distribution examples. Other approaches have also been studied for reducing people's over-reliance on ML models, such as explicitly signaling to people that the model is not perfect [4], or designing cognitive forcing interventions to nudge people into engaging more thoughtfully with the ML model's recommendations and explanations [10]. More recently, realizing that one of the root causes for people's inappropriate reliance on ML models could be their

lack of understanding of basic ML-related knowledge, researchers and educators have advocated for the need of improving people's literacy in AI and machine learning [41, 72]. As such, a growing line of research has been carried out to explore how to effectively increase children's AI literacy through K-12 education [20, 36, 72]. For example, Touretzky et al. developed the guideline for teaching AI to K-12 students, with the goal of ensuring a more informed populace that understands the AI technologies and inspiring the next generation of AI researchers and developers [57]. As another example, researchers utilized gamified design to help students better understand the inner workings of ML models [31, 59].

Different from these earlier works that aim at comprehensively and systematically improving people's AI and machine learning knowledge in the long-term, in this study, we focus on exploring the possibility of providing short-term ML literacy interventions to those people who have the need to interact with some ML model (often commercial) in their decision making. These interventions can be provided by the developers of the ML model, and their ultimate goal is to help people better utilize the model (e.g., rely on the model more appropriately). In this sense, such ML literacy intervention share similarities with the ideas of "user guide" or "user manual" in other domains like automotive vehicles [22, 25, 44].

To help people make better use of the ML model, a critical task of the ML literacy intervention is to inform users of the ML model's potential limitations. In this study, we focus on designing ML literacy interventions that can convey to people one particular limitation of ML models, i.e., the ML model is learned from a set of training data and may not generalize to other distributions of data well. We borrowed existing frameworks like model cards [46] and methodologies like disaggregated evaluation of an ML model [7] during the designs of our ML literacy interventions. We also adopted the concept of "sandbox" for designing an interactive ML literacy intervention—sandbox is an almost realistic but isolated environment that people can explore the system without influencing the real world [5], and it has been widely used in education in different domains like chemistry [5], computer security [50], and entrepreneurial thinking [23]. In our context, we used sandbox as a way for allowing users to construct customized testing datasets to explore an ML model's performance disparity on different data by themselves.

3 STUDY DESIGN

To explore the impacts of different types of machine learning literacy interventions on laypeople's reliance on machine learning models, we conduct a human-subject randomized experiment on Amazon Mechanical Turk (MTurk).

3.1 Experimental Task

In this experiment, human subjects were recruited to evaluate the value of houses with the help of an ML model. In each task, subjects saw the profile of a house, which included information on eight features such as living area size, quality, and number of bedrooms, and they were asked to predict the final sale price of the house. The house presented in each task was selected from the Ames Housing Dataset, a public dataset containing information of properties sold in Ames, Iowa, between 2006 and 2018 [15].

We decided to focus on the domain of real estate valuation in our experiment for several reasons. First, real estate valuation is a realistic domain where commercial ML models have been developed to help humans make better decisions. Second, predicting house prices is a kind of task that laypeople may need to complete in their real life (e.g., when purchasing or selling a home). Thus, laypeople may find it possible to utilize their day-to-day knowledge to make the predictions and to gauge whether to rely on the ML models. Lastly, as we are interested in examining how ML literacy interventions affect laypeople's reliance on ML models on both *in-distribution* and *out-of-distribution* examples, the Ames Housing Dataset provided us with unique opportunities to simulate the differences between in-distribution and out-of-distribution data, as well as the ML model's performance gap on these two types of data distributions. Specifically, we used the K-means clustering algorithm to split all the houses in the Ames Housing Dataset into two clusters—Cluster 1 mainly consisted of low-quality houses with small living areas while Cluster 2 primarily consisted of bigger and higher-quality houses. Then, we randomly sampled 800 houses in Cluster 1 to be used as the training dataset. We trained a linear regression model, **M**, based on the training dataset, and the rest of the houses in Cluster 1 were used as our hold-out validation dataset. The predictions given by the model **M** were then provided to our human subjects on each task in the experiment. In other words, since the ML model used throughout our experiment was **M**, we can treat houses from Cluster 1 as the in-distribution examples, while houses from Cluster 2 are the out-of-distribution examples. The performance of **M** is much better on Cluster 1 as compared to that on Cluster 2 (e.g., the R^2 of **M** on Cluster 1 validation dataset and Cluster 2 are 0.47 and 0.17, respectively); this reflects that the performance of ML models may degrade when applied to a new distribution of data that is different from the model's training data.

3.2 Experimental Treatments

In our experiment, we operationalized the ML literacy interventions as the user tutorials of ML models that were presented to human subjects before they started to work on the house price prediction tasks. We intended to use this user tutorial to help our subjects improve three of the key competencies that they need in order to effectively interact with and evaluate ML technologies, which were previously identified in [41]—understand the steps involved in machine learning ("ML Steps"), recognize that machine learning algorithms learn from data ("Learning from Data"), and be aware of that the examples provided in the training dataset of a machine learning model can affect the results of the model ("Critically Interpret Data"). As a result, the user tutorial we designed always included the same five parts content-wise—Part 1 addressed what an ML model is; Part 2 described how to get an ML model from a training dataset; Part 3 discussed how to evaluate an ML model's performance using a testing dataset; Part 4 emphasized that an ML model can exhibit systematic performance disparities when evaluated on different data; and Part 5 explained that the ML model's systematic performance disparity on different data could be partly caused by the composition of the model's training dataset.

We varied the designs of the user tutorial along two dimensions—the *scope* of the ML model discussed in the tutorial, and whether

interactive components were included in the tutorial—and created 4 versions of tutorials in total (see the supplementary materials for the detailed content of these 4 tutorials). Specifically, we first created two *static* versions of the user tutorials without any interactive components. The key concepts of machine learning were communicated to subjects either as a *general* fact that applies to any ML models, or in the context of the *specific* house price prediction model that subjects would need to use in the tasks. These two versions of user tutorials were presented to subjects in the following two experimental treatments, respectively:

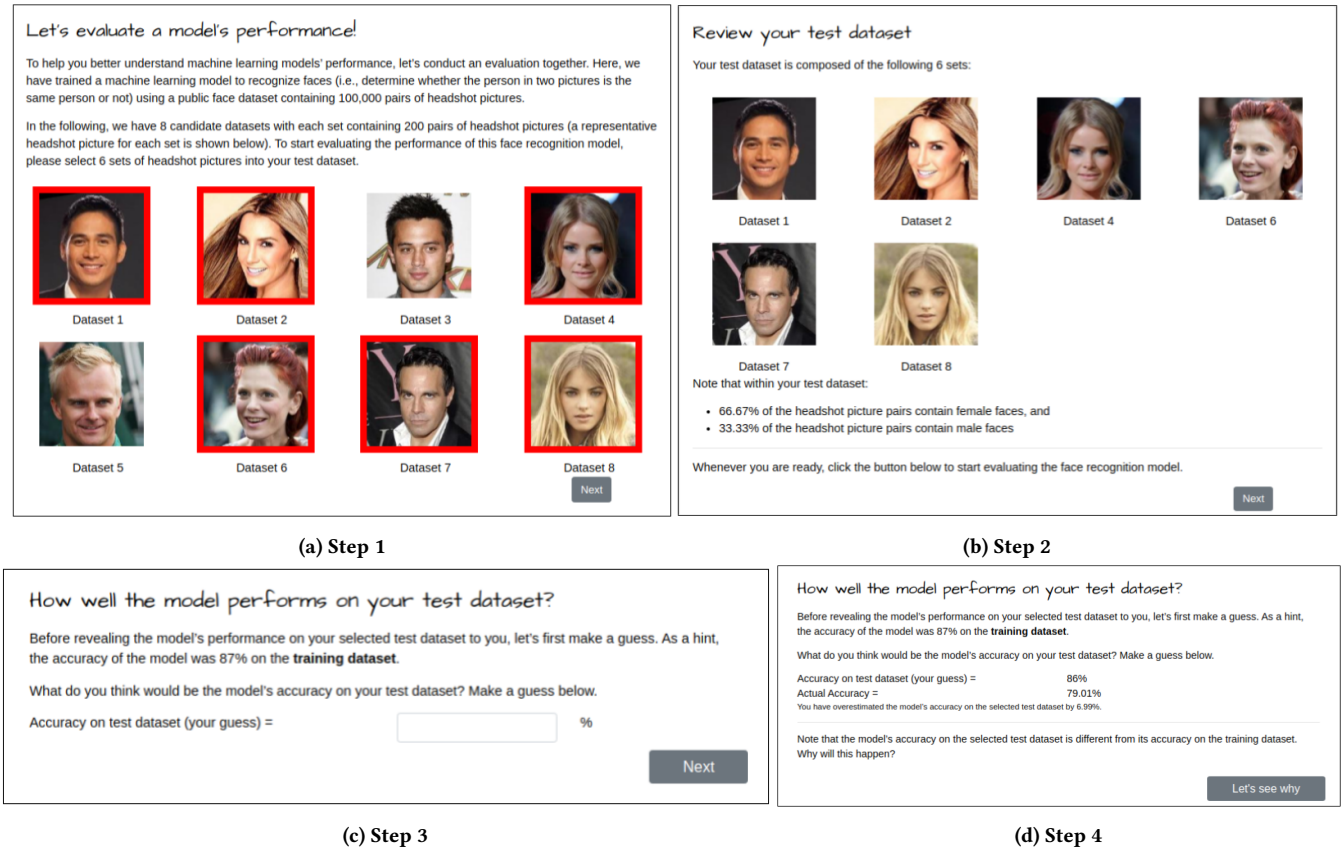
- **General Static**: In this treatment, the user tutorial was static, and it was presented as explaining the general properties of any supervised ML models. To help explain concepts (e.g., prediction, data, patterns, training dataset, testing dataset) in concrete terms, we used ML-powered face recognition models as a running example throughout the tutorial. In particular, in Part 4 of the tutorial, in order to illustrate that an ML model's performance can be different on different data, we revealed to subjects the findings of a recent academic paper which suggested that today's commercial face recognition systems have different levels of performance for people from different demographic groups [60]. Further, in Part 5 of the tutorial, we explained that such findings might be attributed to an unbalanced training dataset—when a face recognition model was trained on a dataset that is overwhelmingly composed of faces from a certain demographic group, even the model performed well on that group, its performance on faces from other demographic groups can be limited. We stressed that this phenomenon was not unique for face recognition models and reminded subjects to be mindful of an ML model's possible performance discrepancies on different data when utilizing it.
- **Specific Static**: In this treatment, the user tutorial was static, and it was presented as explaining the properties of the specific house price prediction model (i.e., **M**) that subjects would use in the task. We followed the "model card" template [46]—a transparent model reporting framework—to design this tutorial. Specifically, the user tutorial covered information on the type, developer, and intended use of the model in Part 1. Next, in Part 2, we explained what the training dataset of the model is, but did not directly describe the distributions of the training dataset to reflect commercial model developers' constraints on sharing training data details. In Part 3, we further described what the evaluation dataset of the model is (i.e., a sample of 200 houses from the hold-out validation dataset[1]), the metrics adopted to evaluate the model's performance (i.e., absolute percentage error, APE), and the relevant factors along which the model's performance may vary (i.e., living area size and house quality). Then, in Part 4, we performed a disaggregated evaluation of the model's performance on the evaluation dataset according to the two chosen factors[2], and visualized the average APE of the model's predictions as well as the 95% confidence interval for each subset of the evaluation data. As houses in the evaluation dataset were selected from Cluster 1, some of its subsets corresponding to the intersection of the chosen factors may contain very few or even no data points (e.g., houses with large size and high quality). Thus, for all subsets that contained fewer than 5 houses, we told subjects that we did not have sufficient data to conduct a reliable evaluation of the model's performance within those subsets, hence we indicated the model's performance as "N/A" in the visualization accordingly. Finally, in Part 5, we explained that a possible reason for the model's performance disparity on different subsets of evaluation data could be the composition of the training dataset, and asked subjects to be mindful of this when utilizing the house price prediction model in their tasks.

Beyond these two static versions of user tutorials, we made a further attempt to increase the interactivity of the tutorials to make them more engaging. In particular, in Part 4 of the tutorial, we provided subjects with a *sandbox* environment in which they can construct customized testing datasets to evaluate the ML model's performance on different data. We thus created two interactive versions of user tutorials which were presented to subjects in the following two experimental treatments:

- **General Sandbox**: The user tutorial used in this treatment was the same as that used in the General Static treatment, except for the designs in Part 4. Specifically, here in Part 4, we actually trained an imperfect face recognition model (with relatively high performance on male faces but low performance on female faces), and we invited subjects to evaluate the performance of this model by themselves (see Figure 1). We had prepared for our subjects 8 candidate datasets with each set containing 200 pairs of face images, while 4 of these sets contained only pairs of male faces and the other 4 sets contained only pairs of female faces. Subjects were then asked to select 6 sets out of the 8 candidate sets to compose their own testing dataset (Figure 1a). Once the subject finalized her selection, we provided her with a brief summary of the gender decomposition for the face image pairs in the selected testing dataset (Figure 1b). We then informed the subject of the model's accuracy on the training dataset, and prompted her to make a guess about the model's accuracy on the selected testing dataset in predicting whether each pair of face images belong to the same person (Figure 1c). After the subject made her guess, we revealed to her the correct answer and displayed the feedback on the errors in her estimate (e.g., "You have overestimated/underestimated the model's accuracy on the selected testing dataset by $x\%$", Figure 1d). Lastly, we nudged the subject to think about why the face recognition model's performance on the training and testing dataset can be quite different, before providing the explanation of training data composition in Part 5. The design of the interactive components in this tutorial was inspired by earlier research findings which showed that prompting people to reflect on their prior knowledge by making predictions and providing self-explanations could improve the recall and comprehension of the information [35].
- **Specific Sandbox**: The user tutorial used in this treatment was the same as that used in the Specific Static treatment except

[1] We used houses from the hold-out validation dataset as the evaluation dataset, because in reality model developers may not be able to get access to a large set of out-of-distribution examples when conducting model evaluations and preparing ML literacy interventions based on the evaluation results.

[2] Specifically, we divided the living area size of a house into three categories: small ($<$1200 square feet), medium (1200 $-$ 2000 square feet), and large ($>$2000 square feet). We also divided the quality of a house (which was a score between 1 and 10) into three categories: low ($<$5), medium (5 $-$ 7), and high ($>$7). We then split the evaluation dataset into subsets based on each factor individually or based on the combinations of the two factors, and we evaluated the model's performance within each subset.



(a) Step 1 **(b) Step 2**

(c) Step 3 **(d) Step 4**

Figure 1: The interface for Part 4 of the tutorial used in the General Sandbox treatment. Subjects were first prompted to constructed their own testing dataset (a). After the testing dataset was selected, subjects were given a summary of its gender decomposition (b). Subjects were then asked to guess the performance of the model on the selected testing dataset (c), before the model's actual performance on the selected testing dataset was revealed to them (d).

for the designs in Part 4. Further, the interactive designs of Part 4 of this tutorial was very similar to that of the tutorial in the General Sandbox treatment. The only differences were that in this tutorial, (1) subjects were invited to evaluate the performance of the house price prediction model **M** that they would use in the tasks by themselves; (2) subjects needed to select 5 houses from a set of 9 candidate houses (6 houses from Cluster 1 and 3 houses from Cluster 2) to compose their own testing dataset; and (3) the summary of the selected testing dataset reported information on the size distribution of the houses in the testing dataset.

Finally, we also included a *control* treatment in which subjects would *not* receive any user tutorial about ML models. Together with the previous four treatments, in total, we had 5 experimental treatments in this experiment.

3.3 Experimental Procedure

Figure 2 provides an overview of the procedure of our experiment. On the high-level, our experiment was designed to reflect the following real-life scenario: the developers of a commercial ML model want to help users utilize the model properly, but are constrained in the kind of information they can share (e.g., they can't show details of the proprietary training data). Thus, they first present the user with a tutorial as an ML literacy intervention, and offer the user with a chance to try out the ML model on some training data so that users can get a sense of how well the model performs. Afterward, the user starts to make decisions in the wild with the help of the model, and the decision-making cases the user encounters may come from the same distribution as the model's training data, but they may also come from a different distribution.

More specifically, our experiment was posted as a Human Intelligence Task (HIT) on Amazon Mechanical Turk (MTurk)[3]. Upon arriving at the HIT, each subject was randomly assigned to one of the five experimental treatments. We started by asking the subject to complete a brief survey reporting her expertise in real estate valuation and machine learning on a five-point Likert scale. Next, the subject was presented with the instructions of the house valuation task, informing them that their task in the HIT was to predict house sale prices with the help of an ML model. For those subjects in the four treatments with user tutorial, after they finished reading the task instructions, we told them that to help them better utilize the ML model in the house valuation task, we had prepared a brief

[3]Screenshots of the HIT interface can be found in the supplementary materials.

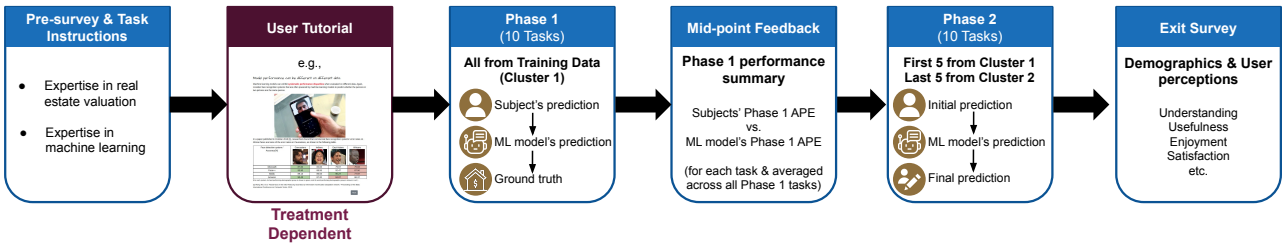


Figure 2: An overview of the procedure of our experiment.

machine learning tutorial for them. The user tutorial of the treatment that the subjects were assigned would then be shown to them. Subjects were required to go through the tutorial at least once, but they could also repeat it multiple times if wish.

Once the subjects completed the instructions and the tutorial, they started to work on the *same* set of 20 tasks to predict house sale prices, which were divided into two phases of 10 tasks each. Phase 1 was designed to reflect the "trial" of the model—subjects can use the model on a few training examples to better understand how well they can predict house prices compared to the ML model, and to get a sense of the kind of data distributions that the ML model was trained on. In particular, subjects were told that the 10 houses in Phase 1 were all selected from the training dataset of the ML model. This means that all Phase 1 houses were selected from Cluster 1 (i.e., small and low quality houses). In each task, the subject was asked to first review the house information and make her prediction on the sale price of the house. Then, the subject would be able to check the prediction given by the ML model **M** as well as the actual sale price of the house. The order of the 10 houses that the subject saw in Phase 1 was randomized. After completing all Phase 1 tasks, the subject was presented with a mid-point feedback page which showed a summary of her own prediction performance as well as the ML model **M**'s prediction performance in Phase 1, as measured by the absolute percentage error (APE) of the subject's or the model's prediction on each of the 10 tasks. By design, the model **M**'s average APE across the 10 tasks in Phase 1 was 15.4%.

After completing Phase 1, the subject moved on to Phase 2 to make price predictions for another 10 houses. Phase 2 was designed to reflect the decision making "in the wild." Thus, we explicitly told subjects that the 10 houses in Phase 2 do not belong to the ML model's training dataset, so it was also the first time for the model to make predictions on them. In each task, the subject was asked to first make her own initial prediction of the house price, P_i, after reviewing the house information. Then, the ML model **M**'s predicted sale price for the house, P_M, would be disclosed to the subject. Finally, the subject needed to submit her final price prediction, P_f. Unlike in Phase 1, subjects would not receive immediate feedback on the actual sale price of the house in each task after completing that task. Again, all subjects saw the exactly *same* set of 10 houses in Phase 2. In addition, the first 5 tasks in Phase 2 contained houses that were selected from the validation dataset of the model (i.e., selected from Cluster 1), while the last 5 tasks in Phase 2 contained houses that were selected from Cluster 2. In other words, while making predictions in Phase 2, subjects would experience distribution shift such that the houses shown in the tasks gradually changed from in-distribution examples to out-of-distribution examples for the ML model **M**. The model **M**'s average APE across the first 5 and last 5 tasks in Phase 2 were 12.3% and 50.4%, respectively, and it systematically underestimated the prices for the out-of-distribution examples (i.e., the last 5 tasks). Note that we did not explicitly tell subjects about the change of data distributions in Phase 2, as users seldom get such hints in the real-world ML-assisted decision making settings.

Finally, after the subject completed all tasks, she filled out a brief exit-survey. In particular, if the subject was presented with a user tutorial in the HIT, she was asked to rate on a 5-point Likert scale about how much she agreed with each of the following statements from 1 (Strongly disagree) to 5 (Strongly agree):

- **(Understanding)** The tutorial about machine learning models that I have received at the beginning of this HIT helps me better understand the house price prediction model that I use in this HIT.
- **(Usefulness)** The tutorial about machine learning models that I have received at the beginning of this HIT helps me properly make use of the house price prediction model in this HIT.
- **(Enjoyment)** I enjoy the tutorial about machine learning models that I have received at the beginning of this HIT.
- **(Satisfaction)** I'm satisfied with the tutorial about machine learning models that I have received at the beginning of this HIT.

Our HIT was open only to U.S. workers on MTurk, and each worker was allowed to take the HIT at most once. The base payment of our HIT was $1.0. To motivate subjects to make accurate predictions and carefully consider how much to rely on the ML model in their predictions, we further provided performance-based bonus opportunities to subjects—In Phase 2, if the average percentage error (APE) of a subject's *initial* prediction was less than 30%, she could earn extra bonuses (APE < 10%: $0.15, 10% ≤ APE < 20%: $0.10, 20% ≤ APE < 30%, $0.05). The same bonus rule also applied to the subject's *final* prediction in each Phase 2 task. Thus, the max amount of bonuses a subject could earn in our HIT was $3.0.

4 DATA

In total, 498 unique subjects participated in our experiment[4]. For each subject, we recorded her responses to the survey questions both at the beginning and at the end of the HIT. For each task in Phase 1, we recorded her house price prediction, while for each task in Phase 2, we recorded her initial prediction P_i as well as her final prediction P_f.

Since subjects only made "ML-assisted decisions" (i.e., they could update their predictions after seeing the ML model's predictions)

[4]On average, a subject spent 10 minutes on our HIT and was compensated $1.8, leading to an effective hourly wage of $10.7.

in Phase 2, our analyses on people's reliance on the ML model were conducted only on Phase 2. We adopted the *weight of advice (WOA)*—a measure initially introduced in the literature of advice-taking [24, 27, 64] and more recently used in several studies of human trust in algorithms [21, 29, 40, 47]—to quantify the extent to which a subject relied on the ML model in each Phase 2 task. Specifically, WOA is defined as $\frac{P_f - P_i}{P_M - P_i}$, which reflects how much the subject weighed the advice P_M that she received from the ML model in a task. As the WOA is highly sensitive to outliers, following the convention established in the literature [24, 30, 52, 54], we truncated the value of WOA by setting values smaller than 0 to 0 and values greater than 1 to 1. So, a WOA of 0 suggests that the subject did not move her prediction towards the ML model's prediction at all after receiving the model's prediction, while a WOA of 1 suggests that the subject adjusted her final prediction to fully match or even move beyond that of the ML model's after seeing the model prediction.

While the WOA values help us quantify subjects' reliance on the ML model, they do not directly provide implications on whether such reliance is *appropriate* or not. This is because the "ideal" WOA value depends on the accuracy comparison between the subject's initial prediction P_i and the model's prediction P_M. In light of this, we adopted another metric—the subject's *prediction performance gain* in each task after seeing the ML model's prediction—to quantify the degree to which the subject's reliance on the ML model is appropriate. Formally, a subject's prediction performance gain (i.e., "prediction error reduction") in a task is defined as the absolute percentage error (APE) difference between the subject's initial prediction and final prediction on the task, i.e., $|\frac{P_i - P_A}{P_A}| - |\frac{P_f - P_A}{P_A}|$ (P_A is the actual sale price for the house in the task). Intuitively, the larger the prediction performance gain, the more appropriate the subject's reliance on the model was, and this is true regardless of how accurately P_i was compared to P_M.

Finally, we conducted data cleaning to filter out potential spammers. Previous literature suggests that spammers on MTurk tend to quickly go through a HIT and enter short random answers to maximize their earnings [38]. Thus, following a similar approach adopted in previous studies [13], we considered keep predicting extremely low prices for houses as a spamming behavior and deemed subjects whose house price predictions were lower than $10,000 on over half of the tasks in Phase 1 as not paying attention[5].

After filtering out the data from inattentive subjects, we were left with the valid data from 458 subjects. Our analyses were conducted on these valid data[6]. Table 1 shows the average level of self-reported expertise in real estate valuation and machine learning for subjects across different treatments (1 is the lowest level and 5 is the highest level). We did not find significant differences in subjects' domain expertise in real estate valuation and machine learning across treatments. We also looked into subjects' prediction performance in Phase 1 to understand their own ability in making house price predictions. Again, we found no significant differences in subjects' prediction performance in Phase 1 across treatments. The median value for subjects' average APE in Phase 1 was 35%, which was close to the bonus threshold we set in the experiment (30%). Thus, after completing the tasks in Phase 1, a significant portion of subjects might perceive themselves as having some degree of capability to make accurate house sale price predictions and earn bonus payments in the HIT even without utilizing the ML model.

[5] $10,000 was chosen as the threshold to ensure subjects' predictions are on the right scale in Phase 1. The actual sale prices for houses in Phase 1 were always 5 or 6 digit numbers that are above $80,000. Since the actual prices were revealed to subjects as feedback in Phase 1, if subjects were paying attention in the HIT, they should be able to learn the magnitude of the house prices while completing the tasks.

[6] We also explored other data cleaning methods (e.g., conduct additional data cleaning steps to remove subjects who kept making high or the same predictions), and our analysis results are qualitatively similar.

5 RESULTS

As discussed earlier, our experiment had a $2 \times 2 + 1$ (control) design. Following the recommendations on how to analyze the experimental data when the control treatment does not fit into the factorial design [28], we first made the comparison between the control treatment and each treatment with a user tutorial to understand whether each type of ML literacy intervention changes laypeople's reliance on ML models, as compared to when the intervention is absent. Then, we focused on analyzing the four treatments resulting from the factorial design to understand how the two factors—the scope and interactivity of the user tutorial—influence laypeople's reliance on ML models. We next looked into whether the changes in laypeople's reliance on the ML models brought up by the ML literacy interventions lead to more appropriate reliance on the ML model or not. Finally, we explored the individual differences in the reliance on ML models across people with different levels of decision-making performance, as well as people's self-reported perceptions of different ML literacy interventions.

5.1 Do ML literacy interventions change people's reliance on ML models?

We start by looking into that for our *entire population of subjects*, whether providing ML literacy interventions to them prior to they start making ML-assisted decisions had any impact on their reliance on the ML model, both on the out-of-distribution and in-distribution examples. Since the dependent variable that we used to capture the subject's reliance on the ML model—the weight of advice (WOA) in Phase 2—did not follow normal distributions, we visualized the median values of the WOA as well as the 95% bootstrap confidence intervals across different treatments in Figure 3.

Formally, we used Mann-Whitney U tests with Bonferroni corrections to test whether the difference in the median WOA values between each treatment with a user tutorial and the control treatment was significant or not. The results showed that on the out-of-distribution examples (Figure 3a), subjects in the Specific Sandbox treatment significantly decreased their reliance on the ML model compared to those in the control treatment ($p = 0.025$). Further, when focusing on the in-distribution examples (Figure 3b), we found that subjects in the General Static and Specific Sandbox treatments decreased their reliance on the ML model on in-distribution examples as compared to subjects in the control treatment (General Static vs. control: $p = 0.031$, Specific Sandbox vs. control: $p = 0.001$). These observations suggest that the Specific Sandbox version of user tutorial may be especially effective in conveying that ML models could suffer from performance drop when distribution shift occurs, so it leads to subjects' decreased reliance on the model on those out-of-distribution examples. However, when

	Baseline	General Static	Specific Static	General Sandbox	Specific Sandbox
N	100	89	91	88	90
Expertise in house price prediction (mean)	3.3	3.3	3.3	3.5	3.6
Expertise in machine learning (mean)	3.5	3.7	3.4	3.5	3.7
Phase 1 average APE (median)	0.36	0.33	0.30	0.36	0.39

Table 1: Comparing subjects' expertise and Phase 1 prediction performance across treatments. Subject's prediction on the first task in Phase 1 was excluded when we computed the average Phase 1 APE for each subject, since subjects may need to use the feedback of the house's actual price they received from the first task to calibrate their predictions.

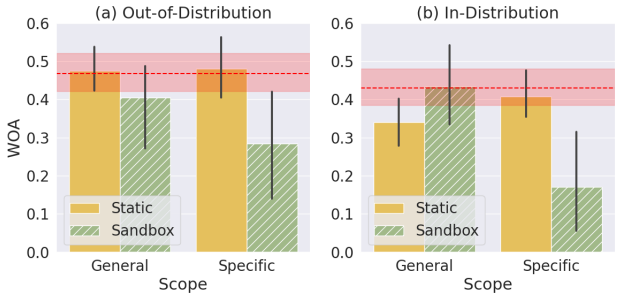


Figure 3: Comparing the median values of the weight of advice (WOA) on (a) out-of-distribution examples and (b) in-distribution examples in Phase 2 across different treatments, for the entire population of subjects. Error bars represent 95% bootstrap confidence intervals. For the control treatment, the dashed horizontal lines represent the median values, and the 95% bootstrap confidence intervals of the median are shown by the red shaded areas.

subjects received the Specific Sandbox version of user tutorial, they might not be engaged in a comprehensive evaluation of the model's performance (e.g., by constructing many different variations of the testing dataset and evaluating the model's performance on each of them) to thoroughly understand the model's strength and weakness, thus they also decreased their reliance on the ML model on in-distribution examples. On the other hand, we speculate that when subjects received the General Static version of user tutorial, they might form a strong negative impression of the ML model's *overall* capability (e.g., perceive all ML models as unable to make consistently accurate predictions), which may explain their decrease of reliance on the model on in-distribution examples.

5.2 The effects of the scope and interactivity of the ML literacy interventions

We next focused on the 4 experimental treatments with a user tutorial, and examined how the scope and interactivity of the ML literacy interventions changed laypeople's reliance on the ML model. For example, as shown in Figure 3a, subjects who received a user tutorial with the interactive components seemed to have a lower reliance on the ML model on the out-of-distribution examples compared to subjects who received a static version of the user tutorial, while the scope of the tutorial did not seem to have an obvious effect here. Since the WOA measures were not normally distributed, we used the aligned rank transformation ANOVA (ART ANOVA) [37, 62], a non-parametric version of the factorial ANOVA, to analyze how the scope and interactivity of the user tutorial influence subjects' reliance on the ML model on out-of-distribution examples. A marginally significant main effect of the interactivity of the tutorial was found in influencing subjects' reliance on the ML model on out-of-distribution examples ($F(1, 1778) = 3.028, p = 0.082$). A post-hoc pairwise comparison with Tukey adjustment [18] suggests that such influence is particularly salient when the ML literacy intervention concerns the specific ML model that people will use, as the WOA values for subjects in the Specific Sandbox treatment on the out-of-distribution examples were significantly lower than those subjects in the Specific Static treatment ($p = 0.046$). Moreover, we did not find any significant main effect of the scope of the ML literacy intervention ($F(1, 1778) = 0.015, p = 0.900$), or any significant interaction between the scope and interactivity of the ML literacy intervention in influencing people's reliance on the model on out-of-distribution examples ($F(1, 1778) = 0.381, p = 0.537$).

Interestingly, when examining how the scope and interactivity of the ML literacy intervention affect subjects' reliance on the ML model on in-distribution examples (Figure 3b), the ART ANOVA test detected a significant interaction effect between these two factors, $F(1, 1779) = 4.734, p = 0.030$. That is, when the user tutorial described properties of ML models in general, the addition of the sandbox in the tutorial did not seem to affect people's reliance on the ML model on in-distribution examples much. However, when the user tutorial discussed properties of the specific ML model that people would use in their tasks, the addition of the sandbox in the tutorial substantially decreased people's reliance on the ML model on in-distribution examples ($p = 0.041$). Again, we speculate that this is because when the user tutorial was interactive, subjects did not engage in a comprehensive evaluation of the model's strength and weakness in the interactive sandbox environment. Yet, subjects in the General Sandbox treatment did not attempt to generalize the limited performance of the face recognition model that they saw in the sandbox to their current use context (i.e., predict house prices), while subjects in the Specific Sandbox treatment might have done so.

5.3 Do ML literacy interventions lead to more appropriate reliance?

To see whether changes in subjects' reliance on the model brought up by the ML literacy interventions result in more appropriate reliance, we compared the subject's prediction performance gain across different treatments. Specifically, we used the Mann-Whitney U tests with Bonferroni corrections to test if the subject's median prediction performance gain was significantly different between the control treatment and each treatment with a user tutorial.

Overall, when considering subjects' prediction performance gain on all Phase 2 tasks, we found that subjects in the Specific Sandbox treatment obtained a significantly larger prediction performance

gain than subjects in the control treatment ($p = 0.007$), indicating that people's reliance on the ML model become more appropriate when the user tutorial addresses the specific ML model to be used and contains interactive components. A closer look into the data suggests that such improvement in appropriate reliance was mainly observed on the out-of-distribution examples—Indeed, on the out-of-distribution examples, compared to that in the control treatment, subjects' decreased reliance on the ML model in the Specific Sandbox treatment leads to a larger prediction performance gain ($p = 0.001$) when the subject's initial prediction was more accurate than the model's prediction, and it does not significantly affect the prediction performance gain when the subject's initial prediction was less accurate than the model's prediction. This means that the Specific Sandbox version of the user tutorial helped subjects reduce their *over-reliance* on the ML model on out-of-distribution examples when subjects could outperform the model. In contrast, on the in-distribution examples, we did *not* find subjects' decreased reliance on the ML model in the Specific Sandbox or General Static treatment was associated with any significant change in their prediction performance gain, regardless of how accurately the subject's initial prediction was compared to the model.

5.4 Individual differences in the effects of ML literacy interventions

Whether and how much people are willing to rely on an ML model may be highly dependent on their perception of their own decision-making performance. In this subsection, we are interested in exploring that for subjects with different levels of performance in making house price predictions themselves, how their reliance on an ML model would be changed by different ML literacy interventions. We thus split all subjects in our experiment into two groups based on their own prediction performance in Phase 1: If the mean APE of a subject's own predictions across the 10 tasks in Phase 1 was lower than 30% (which was the bonus threshold we chose in the experiment), the subject might perceive herself as being able to earn bonus payments even without the assistance of the ML model after seeing the mid-point performance feedback page; we thus considered such subject as a "high-performing" subject. In contrast, if the mean APE of a subject's own predictions across the 10 tasks in Phase 1 was higher than 30%, we considered such subject as a "low-performing" subject. Following this split, 40.4% of subjects in our experiment were high-performing subjects, while the rest 59.6% of the subjects were low-performing subjects.

5.4.1 The effects of ML literacy interventions on high-performing subjects. We first analyzed the data from the high-performing subjects. Figure 4a and Figure 4b showed high-performing subjects' reliance on the ML model on out-of-distribution examples and in-distribution examples, respectively.

Specifically, on the out-of-distribution examples, the Mann-Whitney U tests with Bonferroni correction suggested that high-performing subjects who received any type of user tutorial, except for the Specific Static one, significantly decreased their reliance on the ML model compared to those high-performing subjects who did not receive any user tutorial (i.e., General Static vs. control: $p = 0.030$, General Sandbox vs. control: $p = 0.002$, Specific Sandbox vs. control: $p = 0.039$). An ART ANOVA test further

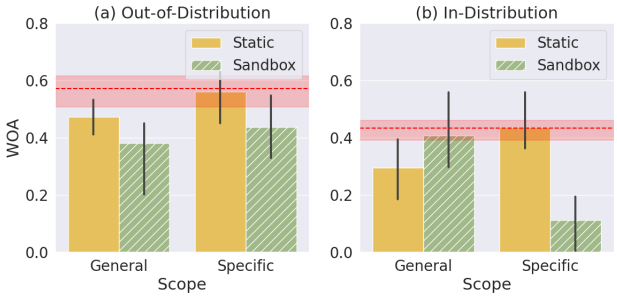


Figure 4: Comparing high-performing subjects' median WOA on (a) out-of-distribution examples and (b) in-distribution examples in Phase 2 across different treatments. Error bars represent 95% bootstrap confidence intervals. For the control treatment, the dashed horizontal lines represent the median values, and the 95% bootstrap confidence intervals of the median are shown by the red shaded areas.

showed that providing a user tutorial that describes the properties of ML models in general leads to a significant decrease in high-performing subjects' reliance on the ML model on out-of-distribution examples, as compared to providing a user tutorial that discusses the specific house price prediction model to be used ($F(1, 716) = 7.319, p = 0.007$). Moreover, increasing the interactivity of the user tutorial also results in a significant decrease in how much high-performing subjects rely on the model on out-of-distribution examples ($F(1, 716) = 4.799, p = 0.028$).

When focusing on high-performing subjects' reliance on the ML model on in-distribution examples, we found that compared to those high-performing subjects who did not receive any ML literacy interventions, the ones who received the General Static ($p = 0.008$) or Specific Sandbox ($p < 0.001$) tutorial tended to significantly decrease their reliance on the model on in-distribution examples. Moreover, a significant interaction between the scope and the interactivity of the ML literacy intervention was detected in influencing high-performing subjects' reliance on the ML model on in-distribution examples ($F(1, 715) = 20.078, p < 0.001$).

Finally, we examined whether these changes in subjects' reliance on the ML model lead to more appropriate reliance. Overall, we found that high-performing subjects in the Specific Sandbox treatment achieved a significantly higher level of prediction performance gain across all prediction tasks in Phase 2 compared to high-performing subjects who did not receive any user tutorial ($p = 0.001$). A closer examination of the data suggests that subjects in the Specific Sandbox treatment mainly relied on the ML model more appropriately by reducing their over-reliance on the ML model when they could outperform the model—their decreased reliance on the model results in a significantly larger prediction performance gain on out-of-distribution examples ($p = 0.034$), and a marginally larger prediction performance gain on in-distribution examples ($p = 0.081$), when their initial prediction was more accurate than the model. However, when a high-performing subject's initial prediction was less accurate than the model, the subject's decrease of reliance on the ML model, as a result of the presence of the Specific Sandbox version of user tutorial, actually leads

to a smaller prediction performance gain on in-distribution examples ($p = 0.020$). Similarly, we also found that when they underperformed the model in their initial predictions, high-performing subjects in the General Sandbox treatment had a significantly smaller prediction performance gain on out-of-distribution examples ($p = 0.006$) compared to those in the control treatment, due to their decreased reliance on the model.

5.4.2 The effects of ML literacy interventions on low-performing subjects. When we conducted similar analyses on the data obtained from low-performing subjects, we had substantially different findings. As shown in Figures 5a and 5b, providing the ML literacy interventions to low-performing subjects did not have significant impacts on their reliance on the model on either out-of-distribution or in-distribution examples, as compared to those low-performing subjects who did not receive any ML literacy interventions. In addition, neither the scope of the ML literacy intervention nor the interactivity had any impact on how much low-performing subjects would rely on the ML model for both out-of-distribution and in-distribution examples. As a result, the provision of ML literacy interventions did not significantly change low-performing subjects' prediction performance gain in the tasks, either.

5.4.3 Summary. Put together, these results suggested that people's own capability in making accurate predictions largely moderates how the ML literacy interventions impact their reliance on the ML model. When people perceive themselves as having low prediction performance themselves, we saw minimal evidence suggesting that whether and how ML literacy interventions were provided would change their reliance on the ML model, possibly because they felt a strong "need" for the assistance from the ML model (hence mostly ignore the limitations of the model). However, for people who perceive themselves as having some capability in making accurate predictions even without the help of the ML model, receiving ML literacy interventions did change their reliance on the ML models. When the ML literacy interventions are presented in proper formats (e.g., the Specific Sandbox tutorial), these changes may lead to reductions in high-performing subjects' over-reliance on the ML model (e.g., when they can outperform the ML model on a task) and thus result in more appropriate reliance.

5.5 Perceptions of the ML literacy interventions

Lastly, we looked into subjects' responses in the exit-survey to explore whether the designs of the user tutorial had any impact on subjects' perceptions of the tutorial, including the perceived understandability and usefulness of the tutorial, as well as the extent to which subjects enjoyed the tutorial and be satisfied with the tutorial. As subjects' responses to these survey questions followed normal distributions, we used two-way ANOVAs to examine whether and how the scope and interactivity influenced subjects' perceptions of the ML literacy interventions. Our results showed that increasing the interactivity of the user tutorial leads to a significant increase in subjects' perceived understandability of the tutorial ($\Delta M = 0.18$, $F(1, 320) = 5.086, p = 0.024$), as well as a marginal increase in subject's perceived usefulness of the tutorial ($\Delta M = 0.16$,

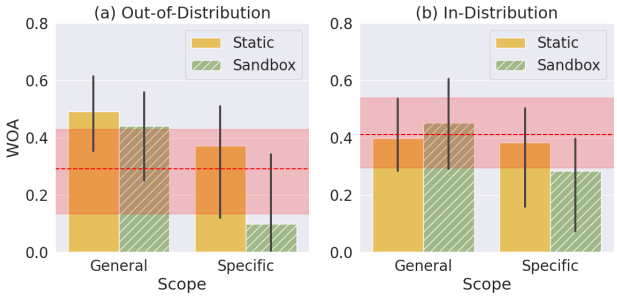


Figure 5: Comparing low-performing subjects' median WOA on (a) out-of-distribution examples and (b) in-distribution examples in Phase 2 across different treatments. Error bars represent 95% bootstrap confidence intervals. For the control treatment, the dashed horizontal lines represent the median values, and the 95% bootstrap confidence intervals of the median are shown by the red shaded areas.

$F(1, 320) = 3.432, p = 0.065$). In addition, although we find no impact of the interactivity of tutorial on how much subjects reported to enjoy the tutorial or found the tutorial satisfying, our data suggested that subjects receiving the interactive tutorial spent significantly more time on the tutorial ($F(1, 362) = 14.411, p < 0.001$) and marginally increased the number of times that they replayed the tutorial ($F(1, 362) = 2.816, p = 0.094$). On the other hand, the scope of the user tutorial was not shown to have any significant impacts on subjects' perceptions of the tutorial.

6 DISCUSSION

In this paper, we present an experimental study to understand whether and how various designs of machine learning literacy interventions will influence laypeople's willingness to rely on an ML model on the data that comes from the same distribution as the model's training data (i.e., in-distribution data) and on the data that comes from a different distribution than the model's training data (i.e., out-of-distribution data). We briefly summarize our experimental results in Table 2. In this section, we reflect on the findings of our study and provide design implications of our results.

Towards interactive ML literacy interventions. A key finding of our experiment is that increasing the interactivity of an ML literacy intervention by, for example, providing people with a sandbox environment to evaluate the ML model's performance on different customized testing datasets, can change people's reliance on the ML model, especially for those people who have a relatively high level of performance on the prediction tasks themselves. In particular, the access to the Specific Sandbox tutorials actually nudged the high-performing subjects in our experiment into relying on the ML model more *appropriately* overall, especially on those cases where they could outperform the model. Moreover, our analyses on subjects' survey responses also indicated that increasing the interactivity of ML literacy interventions may improve people's subjective perceptions of them—in general, people perceive interactive ML literacy interventions to be more understandable and slightly more useful. Together, these results highlight the importance of providing interactive components in the ML literacy interventions to increase the effectiveness of these interventions in influencing

	Out-of-distribution examples			In-distribution examples		
	All	*High-performing*	*Low-performing*	*All*	*High-performing*	*Low-performing*
Do tutorials change reliance?	Specific Sandbox ↓	General Static ↓ General Sandbox ↓ Specific Sandbox ↓	No	General Static ↓ Specific Sandbox ↓	General Static ↓ Specific Sandbox ↓	No
Does tutorial scope affect reliance?	No	General ↓	No	Interaction with tutorial interactivity	Interaction with tutorial interactivity	No
Does tutorial interactivity affect reliance?	Sandbox ↓ (marginal)	Sandbox ↓	No	Interaction with tutorial scope	Interaction with tutorial scope	No
Do tutorials lead to more appropriate reliance?	Specific Sandbox ↑ (human > ML)	Specific Sandbox ↑ (human > ML); General Sandbox ↓ (human < ML)	No	No	Specific Sandbox ↑ (marginal, human > ML); Specific Sandbox ↓ (human < ML)	No

Table 2: A summary of our main experimental results. In the first three rows, ↓ means subjects' reliance decreased in the specified treatment. In the last row, ↑ (↓) means subjects' reliance on the ML model became more (less) appropriate in the specified treatment under the specified scenario, which is shown in parentheses as either the subject's initial prediction was more accurate than the model's prediction ("human > ML") or the subject's initial prediction was less accurate than the model's prediction ("human < ML").

both user understanding and user behavior, which is consistent with findings on how to design effective user education in other contexts like automated driving systems [22].

In our study, the interactive component was designed to help people understand the ML model's possible performance disparities on different data, with the hope that these interactions could allow people to utilize the model more appropriately on different distributions of data. Our experimental results show promises to this end, but we also note that our current designs of the interactive tutorials are still far from the ideal. For example, we found a few undesirable scenarios during our experiment that the inclusion of interactive user tutorials actually results in a less appropriate reliance on the ML model, when people's own prediction is less accurate than the model, especially on in-distribution examples. This may be caused by subjects' insufficient interactions with the user tutorial, so that they did not get a full picture of the model's strength and weakness. It's also possible that subjects had limited knowledge on how their own prediction performance compared with the ML model on different decision making tasks, making it difficult for them to determine how to rely on the model appropriately even after fully understanding the model's strength and weakness. Thus, future research should be conducted to explore how to provide guidance to people during their evaluations of the model in interactive tutorials, so that they can obtain a more comprehensive understanding of both the model's and their own performance on different distributions of data. Finally, how to design suitable interactive components in ML literacy interventions that can serve other purposes (e.g., help people understand how the choice of the objective function for the ML model affects its performance and fairness properties [67]) is another interesting future direction.

On the scope of ML literacy interventions. We explored whether the scope of ML models addressed by the ML literacy interventions changes people's reliance on the ML models, because we suspected that in practice, it is not always possible for developers of commercial models to directly provide information about their model in an ML literacy intervention due to privacy or intellectual property concerns. In this case, it would be ideal if the ML literacy interventions could be provided in the context of other ML models, while people could still generalize the knowledge that they learn from other ML models to the specific model that they would use. In our study, some subjects expressed the difficulty to do so in the exit survey. For example, when asked about what feedback they have for the design of the user tutorial, one subject in the General Static treatment said: "I think that it was not focused enough on examples relative to the HIT to really make it worthwhile. Comparing the ML results of faces is not the same as housing prices to me."

Surprisingly, when examining whether the scope of the user tutorial affects people's reliance on the ML model, we did not detect any main effect of it for the overall subject population. Moreover, for the high-performing subjects, we even found that those who received a user tutorial discussing properties of ML models in general reduced their reliance on the model on out-of-distribution examples to a larger extent than those who received a user tutorial specifically discussing the house price prediction model. However, such larger reduction did not directly result in more appropriate reliance for subjects who received the general tutorials, possibly due to their limited ability to differentiate when they outperform/under-perform the model (e.g., high-performing subjects in the General Sandbox treatment mainly decreased their reliance on the model when their own predictions were less accurate than the model). Nevertheless, our results still suggest that people seem to have some ability to generalize their learned knowledge about ML models from one context to another, showing the promise of designing effective ML literacy interventions even if the information about the specific ML model cannot be fully revealed.

On improving the designs of model cards. We followed the "model card" template [46] to design the user tutorial in the Specific Static treatment of our experiment. Through the exit-survey, some subjects in this treatment expressed how they benefited from the user tutorial to gauge on how much they should rely on the ML model (e.g., "I was glad to know that it rated certain houses with more accuracy than others so I didn't rely on it for those other houses."), and some other subjects also seemed to realize the house price prediction model was trained on an unbalanced dataset (e.g., "I also don't think the tutorial prepared me for the very large houses with very high quality ratings."). On the other hand, we found this static version of the user tutorial on the house price prediction model had very limited impacts on most subjects' reliance on the ML model for both in-distribution and out-of-distribution data, and this is true even for the high-performing subjects.

We had a few conjectures on why the model card seemed to lack a degree of effectiveness in influencing laypeople's reliance on the ML models. First, the target user populations of model cards are mostly people with domain expertise in machine learning, including ML and AI practitioners, model developers, and software developers. As such, in model cards, much of the information related to the performance of ML models was communicated through sophisticated visualizations, such as bar charts with error bars representing confidence intervals. It is thus unclear that for laypeople who may lack sufficient background in STEM, whether they can correctly interpret the information contained in these visualizations. In fact, previous research has found that most non-expert users cannot fully understand statistical information contained in visualizations [45]. Thus, to improve the effectiveness of model cards as an ML literacy intervention in influencing laypeople's reliance on ML models, the designers of the model cards may need to either communicate the model's performance information on different data to people using more accessible language or diagrams, and/or help them increase their literacy in data visualization [9].

We suspected another reason may have also contributed to the ineffectiveness of the model cards: The evaluation dataset we used in the model card was a subset of the hold-out validation dataset of the ML model; as a result, we indicated the model's performance as "N/A" for those subsets of the evaluation dataset containing very few or even no data points. Our intention here was to signal to subjects that houses in those "N/A" subsets could potentially be out-of-distribution examples (we explicitly told subjects that the training dataset and the evaluation dataset of the house price prediction model were sampled from the same dataset). However, subjects in our experiment may interpret "N/A" differently than what we would expect—for example, subjects may simply consider "N/A" as suggesting the model's performance on those subsets was unknown, without realizing that the model's performance could be poor on those subsets as they represent the out-of-distribution data for the model. This highlights the critical challenge of how to transparently communicate and set the right expectation about an ML model's possible performance on out-of-distribution data without sufficient out-of-distribution examples in the model's evaluation dataset.

Finally, given that in our experiment, the user tutorial used in the Specific Sandbox treatment had a much salient impact on laypeople's reliance on the ML models compared to the model cards, especially for empowering subjects to rely on the model more appropriately, we recommend incorporating interactive components with model cards to improve its effectiveness as an ML literacy intervention. We note that the example model cards Google deployed online [1] did add a degree of interactivity into the original model card template—given an ML model, users can choose an evaluation dataset from a pre-defined list or upload their own testing dataset, select relevant factors for analysis, and view the model's performance on the selected evaluation dataset. Compared to Google's interactive model card, our Specific Sandbox tutorial allowed subjects to customize the composition of the testing dataset along pre-defined factors (e.g., gender, house size) and included more elements that aimed to increase subjects' recall and comprehension of information (e.g., guess the model's performance on the testing dataset). Further studies are thus needed to advance our understandings of whether and how these variations result in differences in laypeople's reliance on ML models.

Limitations and other future work. Our study was conducted with laypeople (i.e., subjects recruited from Amazon Mechanical Turk) on one specific type of prediction task. Therefore, cautions should be used when generalizing results in this work to different settings, such as how real-estate experts who lack ML-related knowledge would rely on the ML model after receiving the ML literacy interventions, or how laypeople will rely on ML models upon receiving the ML literacy interventions when working on prediction tasks that are significantly easier or harder. In addition, for the two versions of specific user tutorials used in our experiment, we assumed at least some key relevant variables for defining out-of-distribution examples (i.e., house sizes) for the ML model are known, but such information may not exist in reality. Therefore, an interesting future work is to design an effective ML literacy intervention for a specific ML model to help people rely on this model appropriately on its out-of-distribution examples, even when how to characterize its out-of-distribution examples is unclear. We also note that none of our ML literacy interventions could impact the low-performing subjects' reliance on the ML model. Future work should investigate more into how to help individuals with low decision-making performance themselves to utilize the ML models more appropriately. Finally, our study was motivated by the scenario that developers of *commercial* models design user tutorials to help users utilize the ML model more appropriately, thus the contents of these tutorials are often subject to constraints (e.g., training data details can't be shared). It's interesting in the future to explore that in a non-commercial setting where these constraints no longer exist, how to design effective ML literacy interventions to promote appropriate reliance on ML models.

7 CONCLUSION

In this paper, we present an experimental study to explore whether and how a short-term machine learning literacy intervention can be designed to help laypeople become aware of ML model's possible limitations in generalizing to new data distributions, and thus influence the ways that they interact with the model. We find that a brief tutorial about machine learning that highlights ML model's possible performance disparities on different data has the potential to help those people who have a relatively high level of ability in the decision-making tasks themselves to rely on the model more appropriately, when the tutorial addresses the specific model people will use in their decision making and contains interactive components. On the other hand, for people whose own decision-making performance is relatively low, how much they are willing to rely on an ML model is not significantly affected by whether they receive any ML literacy interventions and what types of interventions they receive. Finally, ML literacy interventions are perceived to be more understandable and slightly more useful when they are interactive. These results provide important implications for promoting appropriate use of ML models through enhancing people's machine learning and AI literacy, and we hope the findings we report in this paper can inspire more discussions in this line.

ACKNOWLEDGMENTS

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

REFERENCES

[1] [n. d.]. https://modelcards.withgoogle.com/face-detection
[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] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. ProPublica (2016). *URL: https://www. propublica. org/article/machine-bias-risk-asses sments-in-criminal-sentencing* (2016).
[4] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. 2021. AI-Assisted Human Labeling: Batching for Efficiency without Overreliance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.
[5] George F Atkinson. 1985. Professional education in the sandbox: Hazard, perceived risk, acceptable risk. *Journal of Chemical Education* 62, 12 (1985), 1070.
[6] 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.
[7] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. *arXiv preprint arXiv:2103.06076* (2021).
[8] James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. New York, 35.
[9] Katy Börner, Andreas Bueckle, and Michael Ginda. 2019. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1857–1864.
[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.
[11] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
[12] 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.
[13] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. 120–129.
[14] Shih-Yi Chien, Michael Lewis, Katia Sycara, Jyi-Shane Liu, and Asiye Kumru. 2018. The effect of culture on trust in automation: reliability and workload. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 4 (2018), 1–31.
[15] Dean De Cock. 2011. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education* 19, 3 (2011).
[16] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.
[17] Stefania Druga, Sarah T Vu, Eesh Likhith, and Tammy Qiu. 2019. Inclusive AI literacy for kids around the world. In *Proceedings of FabLearn 2019*. 104–111.
[18] Lisa A Elkin, Matthew Kay, James J Higgins, and Jacob O Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. *arXiv preprint arXiv:2102.11824* (2021).
[19] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.
[20] Alex Fang. 2019. Chinese colleges to offer AI major in challenge to US. *Nikkei Asian Review* (2019).
[21] Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. 2021. The Impact of Algorithmic Risk Assessments on Human Predictions and its Analysis via Crowdsourcing Studies. *arXiv preprint arXiv:2109.01443* (2021).
[22] Yannick Forster, Sebastian Hergeth, Frederik Naujoks, Josef Krems, and Andreas Keinath. 2019. User Education in Automated Driving: Owner's Manual and Interactive Tutorial Support Mental Model Formation and Human-Automation Interaction. *Information* 10, 4 (2019), 143.
[23] Xiang Fu, Simona Doboli, and John Impagliazzo. 2010. Work in progress—a sandbox model for teaching entrepreneurship. In *2010 IEEE Frontiers in Education Conference (FIE)*. IEEE, F2C–1.
[24] Francesca Gino. 2008. Do we listen to advice just because we paid for it? The impact of advice cost on its use. *Organizational behavior and human decision processes* 107, 2 (2008), 234–245.
[25] Osman Gök, Pervin Ersoy, and Gülmüş Börühan. 2019. The effect of user manual quality on customer satisfaction: the mediating effect of perceived product quality. *Journal of Product & Brand Management* (2019).
[26] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 90–99.
[27] Nigel Harvey and Ilan Fischer. 1997. Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational Behavior and Human Decision Processes* 70, 2 (1997), 117–133. https://doi.org/10.1006/obhd.1997.2697
[28] Samuel Himmelfarb. 1975. What do you do when the control group doesn't fit into the factorial design? *Psychological Bulletin* 82, 3 (1975), 363.
[29] Yoyo Tsung-Yu Hou and Malte F Jung. 2021. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.
[30] Mandy Hütter and Fabian Ache. 2016. Seeking advice: A sampling approach to advice taking. *Judgment & Decision Making* 11, 4 (2016).
[31] Sven Jatzlau, Tilman Michaeli, Stefan Seegerer, and Ralf Romeike. 2019. It's not Magic After All – Machine Learning in Snap! using Reinforcement Learning. In *2019 IEEE Blocks and Beyond Workshop (B B)*. 37–41. https://doi.org/10.1109/BB48857.2019.8941208
[32] Ken Kahn and Niall Winters. 2017. Child-friendly programming interfaces to AI cloud services. In *European Conference on Technology Enhanced Learning*. Springer, 566–570.
[33] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[34] Antino Kim, Mochen Yang, and Jingjng Zhang. 2020. When Algorithms Err: Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms. *Late Errors on Users' Reliance on Algorithms (July 2020)* (2020).
[35] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1375–1386.
[36] Irene Lee, Safinah Ali, Helen Zhang, Daniella DiPaola, and Cynthia Breazeal. 2021. Developing Middle School Students' AI Literacy. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (Virtual Event, USA) *(SIGCSE '21)*. Association for Computing Machinery, New York, NY, USA, 191–197. https://doi.org/10.1145/3408877.3432513
[37] Christophe Leys and Sandy Schumann. 2010. A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology* 46, 4 (2010), 684–688.
[38] Leib Litman, Jonathan Robinson, and Cheskie Rosenzweig. 2015. The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior research methods* 47, 2 (2015), 519–528.
[39] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *arXiv preprint arXiv:2101.05303* (2021).
[40] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
[41] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.
[42] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.
[43] Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, et al. 2020. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Medical Image Analysis* 66 (2020), 101714.
[44] Brad Mehlenbacher, Michael S Wogalter, and Kenneth R Laughery. 2002. On the reading of product owner's manuals: Perceptions and product complexity. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 46. SAGE Publications Sage CA: Los Angeles, CA, 730–734.
[45] Luana Micallef, Pierre Dragicevic, and Jean-Daniel Fekete. 2012. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE transactions on visualization and computer graphics* 18, 12 (2012), 2536–2545.

[46] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[47] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[48] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *Proceedings of the 2022 chi conference on human factors in computing systems*.

[49] 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.

[50] Gordon W. Romney and Brady R. Stevenson. 2004. An Isolated, Multi-Platform Network Sandbox for Teaching IT Security System Engineers. In *Proceedings of the 5th Conference on Information Technology Education* (Salt Lake City, UT, USA) *(CITC5 '04)*. Association for Computing Machinery, New York, NY, USA, 19–23. https://doi.org/10.1145/1029533.1029539

[51] Julian Sanchez, Wendy A Rogers, Arthur D Fisk, and Ericka Rovira. 2014. Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical issues in ergonomics science* 15, 2 (2014), 134–160.

[52] Thomas Schultze, Anne-Fernandine Rakotoarisoa, and Stefan Schulz-Hardt. 2015. Effects of distance between initial estimates and advice on advice utilization. *Judgment & Decision Making* 10, 2 (2015).

[53] Jennifer Skeem, Nicholas Scurich, and John Monahan. 2020. Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and human behavior* 44, 1 (2020), 51.

[54] Jack B Soll and Richard P Larrick. 2009. Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of experimental psychology: Learning, memory, and cognition* 35, 3 (2009), 780.

[55] Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *12th ACM Conference on Web Science*. 315–324.

[56] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2021)*.

[57] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. 2019. Envisioning AI for K-12: What should every child know about AI?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9795–9799.

[58] David S Touretzky. 2017. Computational thinking and mental models: From Kodu to Calypso. In *2017 IEEE Blocks and Beyond Workshop (B&B)*. IEEE, 71–78.

[59] Jessica Van Brummelen, Tommy Heng, and Viktoriya Tabunshchyk. 2021. Teaching Tech to Talk: K-12 Conversational Artificial Intelligence Literacy Curriculum and Development Tools. In *2021 AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI)*.

[60] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. 2019. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE International Conference on Computer Vision*. 692–702.

[61] 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.

[62] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[63] 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.

[64] Ilan Yaniv. 2004. Receiving other people's advice: Influence and benefit. *Organizational behavior and human decision processes* 93, 1 (2004), 1–13.

[65] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. Making sense of recommendations. *Journal of Behavioral Decision Making* 32, 4 (2019), 403–414.

[66] 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.

[67] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1245–1257.

[68] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate? An Investigation from Perception to Decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277

[69] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* 15, 11 (2018), e1002683.

[70] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

[71] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[72] Michelle Renée Zimmerman. 2018. *Teaching AI: exploring new frontiers for learning*. International Society for Technology in Education.

[73] Abigail Zimmermann-Niefield, Makenna Turner, Bridget Murphy, Shaun K Kane, and R Benjamin Shapiro. 2019. Youth learning machine learning through building models of athletic moves. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. 121–132.