You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift

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

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

ABSTRACT

Decision-making aids powered by machine learning models become increasingly prevalent on the web today. However, when applied to a new distribution of data that is different from the training data (i.e., when covariate shift occurs), machine learning models often suffer from performance degradation and may provide misleading recommendations to human decision-makers. In this paper, we conduct a randomized experiment to investigate how people rely on machine learning models to make decisions under covariate shift. Surprisingly, we find that people rely on machine learning models more when making decisions on out-of-distribution data than in-distribution data. Moreover, while increasing people's awareness of the machine learning model's possible performance disparity on different data helps decrease people's over-reliance on the model under covariate shift, enabling people to visualize the data distributions and the model's performance does not seem to help. We conclude by discussing the implication of our results.

CCS CONCEPTS

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

KEYWORDS

Machine Learning, covariate shift, human-AI interaction, appropriate reliance

ACM Reference Format:
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 (WebSci '21), June 21–25, 2021, Virtual Event, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3447535.3462487

1 INTRODUCTION

Internet users today are increasingly assisted by recommendations supplied by machine learning (ML) models to make better decisions online in diverse domains from entertainment to investment. Achieving the optimal human-machine partnership, however, requires humans to rely upon the model recommendations *appropriately*, that is, rely on the model when its recommendation is right and override it when it is wrong [6]. A typical scenario for ML models to provide unreliable recommendations is when the distribution of data on which a model is trained is different from that to which the model is applied, leading to what is known as *covariate shift* [21]. Indeed, many ML models are developed based on the assumption that the training data is drawn from an identical distribution as the test data. Yet, this assumption often does not hold in reality due to practical issues like sampling biases in the training data collection process [4] and the constant evolvement of the deployment environment. Unfortunately, when covariate shift occurs, the performance of ML models may significantly deteriorate [18], implying that the use of machine assistance in these scenarios potentially poses risks to effective human decision-making.

A critical but currently under-explored question, thus, is how humans rely on ML models when making decisions under covariate shift: Can people recognize the changes in data distributions? How would they adjust their reliance on ML models on out-of-distribution data? And what can be done to help people rely on ML models more appropriately under covariate shift? In this work, we provide some initial answers to these questions to understand how *laypeople*—who are increasingly the end-users of ML-powered decision aids—rely on ML models when covariate shift occurs.

Specifically, we conducted a randomized controlled experiment with 549 human subjects recruited from Amazon Mechanical Turk. Subjects were asked to predict house sale price with the assistance of an ML model in a sequence of 20 tasks, which were divided into two phases of 10 tasks each. In Phase 1, subjects interacted with the ML model and observed its performance on some houses drawn from the in-distribution held-out validation dataset. Then, in Phase 2, subjects needed to decide whether to delegate the decision making right to the model for predicting the price for some unseen houses. Subjects were randomized into treatments where houses they saw in Phase 2 came from either the same distribution as the training data of the ML model or a different distribution. Moreover, to overcome people's possible inability to recognize data distribution changes and/or their possible tendency to generalize an ML model's performance from one data distribution to another, we designed two types of external interventions that aimed at helping people address these limitations. Subjects in our experiment were randomly assigned to receive one of these two interventions or receive no intervention at all.

Our experimental results show that, surprisingly, laypeople tend to rely on an ML model *more* when covariate shift occurs, effectively resulting in over-reliance on an ML model when its performance is poor. A closer look into the data suggests that people have some capability in detecting the change of data distribution. However, they actively choose to rely on the model more under covariate

WebSci '21, June 21–25, 2021, Virtual Event, United Kingdom

ACM ISBN 978-1-4503-8330-1/21/06.
https://doi.org/10.1145/3447535.3462487

shift because they expect the model to maintain its performance on out-of-distribution data, while they believe their own decision-making performance would decrease on those data. Besides, we find that providing people with a brief education session about the possible performance disparity of an ML model on different data can effectively reduce people's over-reliance on ML models under covariate shift. In contrast, equipping people with an interactive tool to visualize the distribution of data and the model's performance on different data is ineffective in helping people rely on an ML model more appropriately on out-of-distribution data.

Taken together, our results reveal a concerning finding that laypeople may have misbelief about an ML model's capacity, and thus overly rely on an erroneous model when the distribution of data changes. These results highlight the importance of clearly and transparently communicating to people the scope of application and potential limitation of an ML model, as well as actively assisting people in understanding the range of cases that they can generalize a model's observed performance to. There is also a pressing need to increase people's AI literacy, such as raising people's awareness of the possible performance degradation of ML models on out-of-distribution data. We conclude by discussing the implications of our study on promoting appropriate reliance on AI.

2 RELATED WORK

The increasing prevalence of ML-aided decision making has inspired a growing number of empirical studies that aim at understanding how people interact with, trust, and rely upon ML models. For example, previous research has shown that while laypeople tend to adopt recommendations supplied by ML models over human suggestions in an objective and unfamiliar domain [13], they are in general unwilling to rely on algorithmic models in highly subjective domains [30], or after witnessing the model makes errors [8]. Researchers have also identified a variety of factors that could influence people's reliance on ML models. For example, people are shown to increase their reliance on models with higher levels of accuracy [31]. In addition, people's first impression and mental model of an ML model [2, 26], the model's confidence and interpretability [22, 29, 32], and the consistency between the model and humans in both their decisions and rationales [15, 33] are all shown to impact people's reliance on the model.

A major risk in ML-aided decision making is that people may rely on a model inappropriately, and such risk is elevated when ML models are operated on out-of-distribution data. Indeed, the phenomenon of the distribution of input variables (i.e., features) changes between the data of training and deployment stages is known as "*covariate shift*" [21, 24]. Many ML models are known to be not good at adapting to new and unfamiliar data [18, 25], which raises the question of how people would rely on ML models when covariate shift occurs.

In this paper, we focus on understanding, under covariate shift, whether people rely on ML models appropriately and how to promote appropriate reliance. Previously, researchers mainly attempt to enhance people's appropriate reliance on ML models through calibrated model confidence scores [32] or carefully designed model explanations [22, 32]. These approaches have mixed success when being evaluated on in-distribution data, and their effectiveness in promoting appropriate reliance on ML models on out-of-distribution data is unclear. For example, it is shown that the state-of-the-art ML models that produce calibrated confidence scores on in-distribution data often come with uncalibrated confidence scores on out-of-distribution data [19], while increasing an ML model's transparency actually decreases people's capability in detecting obvious model mistakes on out-of-distribution data [20].

In light of this, here, we design two alternative interventions, specifically for improving people's appropriate reliance on ML under covariate shift. In the first intervention, similar to the general user education used in other domains like automated-driving [10], we provide people with information that increases their understandings of the performance of ML models, especially on ML models' possible performance disparity on different data. Our second intervention involves a visualization tool that helps people explore both the data distribution and the model's performance on different data; this is inspired by previous efforts that use interactive visualizations to explain the behavior of ML models [11].

3 STUDY DESIGN

To understand people's reliance on ML models under covariate shift, we conducted a randomized behavioral experiment[1], in which human subjects were recruited from Amazon Mechanical Turk (MTurk) to complete some decision-making tasks with assistance from an ML model. Our main research questions are:

- **RQ1**: When covariate shift occurs, how will people adjust their levels of reliance on ML models?
- **RQ2**: Can external interventions, such as educating people about the performance of ML models and enabling people to visualize the distributions of decision-making tasks as well as the model's performance on different tasks, help people rely on ML models more appropriately when covariate shift happens?

3.1 Experimental Task

The decision-making task that subjects worked on in our experiment was to predict the sale prices of houses. In each task, subjects were presented with information about a house on eight features (e.g., living area size, quality, year built), and were asked to make a prediction of the sale price of the house. The housing data we used came from a public dataset [7] containing houses sold in Iowa, United States, from 2006 to 2010.

We chose the task of house price prediction for several reasons. First, this task characterizes a kind of decision-making activity in people's daily life; thus, it is easily understandable by our human subjects. Second, it represents a realistic scenario where ML models are developed to assist human decision-making. Another critical reason for us to select this task in our experiment is that the housing dataset we used allowed us to simulate changes in the data distribution and build real ML models whose performance would decrease when applied to a new distribution of data. In particular, by applying the K-means clustering algorithm on the entire set of houses, we obtained two distinctive clusters of houses—Cluster 1 mostly consisted of houses with small living areas and low quality, while Cluster 2 mostly contained houses that were bigger and of

[1] Our experiment was approved by the Purdue IRB.

Pre-experiment Survey

S1 How much expertise do you have in estimating house price?
S2 How much knowledge do you have in machine learning?

Post-experiment Survey

S3 Based on your observations, are the houses you saw in phase 2 similar to those that you saw in phase 1?
S4 Do you think the model's performance in phase 2 would be better than its performance in phase 1?
S5 Do you think your performance in phase 2 would be better than your performance in phase 1?
S6 [checkbox] What are the factors that make you stop using the model in phase 2? (*for subjects who stopped using the model in phase 2*)
S7 [checkbox] What are the factors that make you use the model for all the tasks in phase 2? (*for subjects who always used the model in phase 2*)

Table 1: Survey questions we asked in our experiment. For S1–S5, subjects answered each question using a 5-point Likert scale. For S6 and S7, checkbox options are determined via a pilot study in which subjects provided free-form answers to the same questions.

high-quality. Moreover, we found that the linear regression model **M** that was trained using houses from Cluster 1 performed much better on Cluster 1 than on Cluster 2 (e.g., the R^2 of **M** on Cluster 1 and Cluster 2 are 0.47 and 0.17, respectively). As a result, in our experiment, we used **M** as our ML model and presented the predictions of **M** to subjects in each task as the model's recommendations, making houses in Cluster 2 (Cluster 1) effectively the out-of-distribution (in-distribution) data.

3.2 Experimental Procedure

The subject started our experiment by reporting her expertise in predicting house price and in ML on a five-point Likert scale. Then, she performed a sequence of 20 house price prediction tasks, divided into two phases of 10 tasks each, with the help of a pre-trained ML model. Phase 1 was designed to help subjects understand their ability as well as the ML model's ability in accurately predicting house prices. In particular, on each task of phase 1, we showed to the subject the information of a house that was drawn from the held-out validation dataset of **M**, which belonged to Cluster 1. After reviewing the house's information, the subject was asked first to forecast its sale price by herself. Then, the model's prediction, produced by **M**, and the house's actual sale price would be revealed to her. All subjects saw the *same* 10 tasks in phase 1, though the order was randomized. Upon completing all tasks in phase 1, the subject received a mid-point feedback page, summarizing in a table her own prediction accuracy as well as the model's accuracy in phase 1, in terms of both the absolute percentage error (APE) on each of the 10 tasks and the average APE across all 10 tasks.

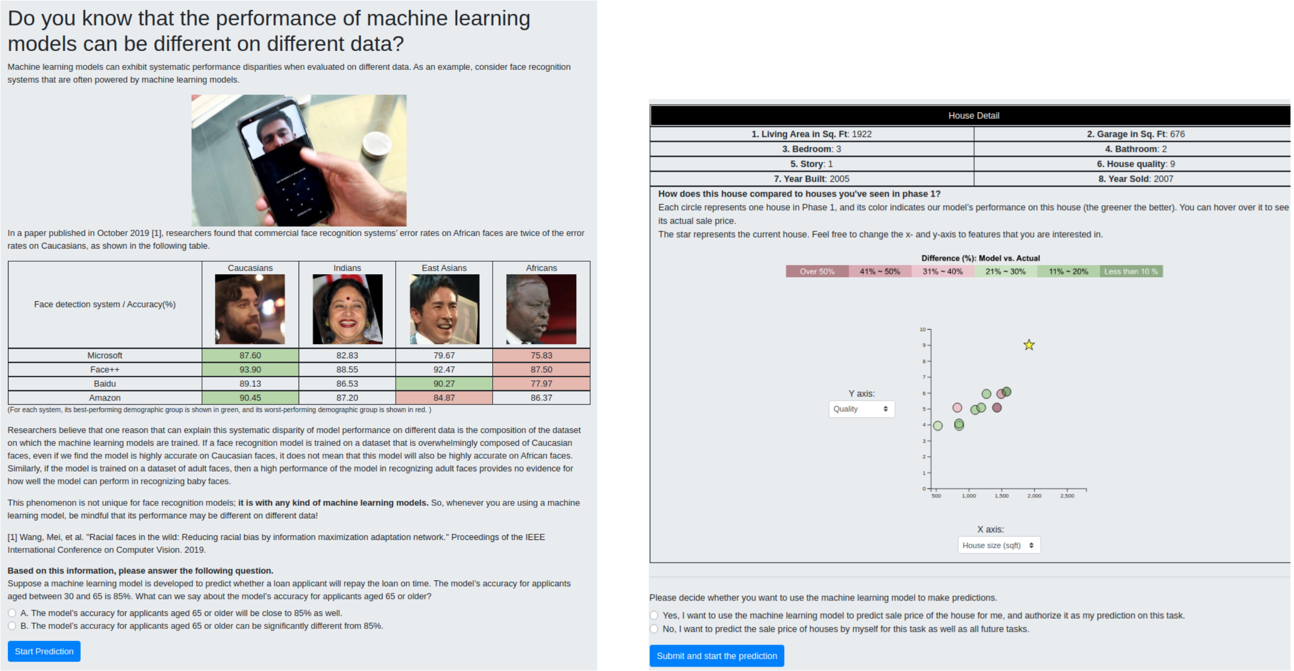
Next, in phase 2, the subject was asked to predict prices for 10 additional houses for real. On these 10 tasks, the subject would *not* receive the immediate feedback about the actual sale price of the house. Specifically, on each task, after viewing the house's information, the subject needed to decide whether to delegate the decision-making right to the ML model—if yes, the model's prediction on this task would be used as the subject's prediction; otherwise, the subject needed to make her own prediction on this task *as well as in all future tasks*. This experimental setup was designed to reflect the real-life scenarios that people may abandon an ML model once they find it untrustworthy [12]—people could choose to rely on an ML model by authorizing the model to make decisions on behalf of themselves (e.g., use an auto-trading program to trade), but they could also override such authorization anytime later by opting out of the usage of the model when they lose faith in it (e.g., stop paying for the auto-trading program thus lose access to it). Depending on the treatment a subject was assigned, the 10 houses she saw in phase 2 could come from Cluster 1 (small and low-quality houses) or Cluster 2 (large and high-quality houses), but the model prediction the subject saw on each house in phase 2 was always generated by the model **M**, which was trained using data from Cluster 1 (see more details in Section 3.3).

After completing all the prediction tasks, the subject was asked to complete an exit survey to report her perceptions of the tasks, her belief of the model's performance and her own performance in the tasks, the factors that influence her usage of the model in phase 2, as well as some demographic information. Table 1 shows the list of questions we asked in our surveys. In the end, we revealed to the subject the actual sale prices for the 10 houses in phase 2, together with the subject's prediction accuracy on these houses.

We opened the experiment only to U.S. workers on MTurk, and each worker can participate at most once. The base payment of this experiment was \$0.5. In addition, to encourage subjects to carefully consider whether to rely on the ML model in *phase 2*, we informed each subject at the beginning of the experiment that for each phase 2 task, if the APE of her prediction is less than 30%, she could earn additional bonuses (APE<10%: \$0.30 bonus, 10%≤APE<20%: \$0.20 bonus, 20%≤APE<30%: \$0.10 bonus). This bonus scheme leads to a maximum bonus amount of \$3, which could only be earned if subjects made accurate predictions in phase 2. We also carefully selected the bonus threshold (i.e., APE<30%) given the set of prediction tasks we used in phase 1—the model **M** had an average APE of 28.3% in phase 1, and for 7 out of the 10 tasks in phase 1, the model's APE was less than 30%. Meanwhile, we found via a pilot study that on average, a subject's own predictions could achieve an APE that was less than 30% on 5.7 out of the 10 tasks in phase 1. In other words, the bonus threshold was selected to ensure that after completing phase 1, an average subject would feel her own prediction performance was worse than the model, but it's still possible for her to earn some bonuses by herself without relying on the model.

3.3 Experimental Design

Subjects in our experiment were randomly assigned to one of the six experimental treatments that were arranged in a 2 × 3 design. The treatments differed along two dimensions: the type of *task distribution* in phase 2, and the existence and type of *external interventions* that subjects received to help them appropriately rely on ML models when covariate shift occurs. With respect to the task distribution, we randomized subjects into one of the two levels:



Do you know that the performance of machine learning models can be different on different data?

Machine learning models can exhibit systematic performance disparities when evaluated on different data. As an example, consider face recognition systems that are often powered by machine learning models.

In a paper published in October 2019 [1], researchers found that commercial face recognition systems' error rates on African faces are twice of the error rates on Caucasians, as shown in the following table.

Face detection system / Accuracy(%)	Caucasians	Indians	East Asians	Africans
Microsoft	87.60	82.83	79.67	75.83
Face++	93.90	88.55	92.47	87.50
Baidu	89.13	86.53	90.27	77.97
Amazon	90.45	87.20	84.87	86.37

(For each system, its best-performing demographic group is shown in green, and its worst-performing demographic group is shown in red.)

Researchers believe that one reason that can explain this systematic disparity of model performance on different data is the composition of the dataset on which the machine learning models are trained. If a face recognition model is trained on a dataset that is overwhelmingly composed of Caucasian faces, even if we find the model is highly accurate on Caucasian faces, it does not mean that this model will also be highly accurate on African faces. Similarly, if the model is trained on a dataset of adult faces, then a high performance of the model in recognizing adult faces provides no evidence for how well the model can perform in recognizing baby faces.

This phenomenon is not unique for face recognition models; **it is with any kind of machine learning models.** So, whenever you are using a machine learning model, be mindful that its performance may be different on different data!

[1] Wang, Mei, et al. "Racial faces in the wild: Reducing racial bias by information maximization adaptation network." Proceedings of the IEEE International Conference on Computer Vision. 2019.

Based on this information, please answer the following question.
Suppose a machine learning model is developed to predict whether a loan applicant will repay the loan on time. The model's accuracy for applicants aged between 30 and 65 is 85%. What can we say about the model's accuracy for applicants aged 65 or older?

○ A. The model's accuracy for applicants aged 65 or older will be close to 85% as well.
○ B. The model's accuracy for applicants aged 65 or older can be significantly different from 85%.

(a) Education intervention

House Detail	
1. Living Area in Sq. Ft: 1922	**2. Garage in Sq. Ft**: 676
3. Bedroom: 3	**4. Bathroom**: 2
5. Story: 1	**6. House quality**: 9
7. Year Built: 2005	**8. Year Sold**: 2007

How does this house compared to houses you've seen in phase 1?
Each circle represents one house in Phase 1, and its color indicates our model's performance on this house (the greener the better). You can hover over it to see its actual sale price.
The star represents the current house. Feel free to change the x- and y-axis to features that you are interested in.

Difference (%): Model vs. Actual

Over 50%	41% ~ 50%	31% ~ 40%	21% ~ 30%	11% ~ 20%	Less than 10 %

Y axis: Quality
X axis: House size (sqft)

Please decide whether you want to use the machine learning model to make predictions.

○ Yes, I want to use the machine learning model to predict sale price of the house for me, and authorize it as my prediction on this task.
○ No, I want to predict the sale price of houses by myself for this task as well as all future tasks.

(b) Visualization intervention (phase 2)

Figure 1: Two external interventions that are designed to promote appropriate reliance on ML models under covariate shift. In Figure 1b, the star on the coordinate plane represents the house for which the subject needs to predict sale price in the current task. The circles on the coordinate plane are the houses in phase 1, with the color representing the model accuracy on them (green means higher accuracy, and red means lower accuracy).

- **Stationary distribution**: The houses in phase 2 were all selected from *Cluster 1*. That is, the prediction tasks subjects needed to work on in phase 2 came from the *same* distribution as the prediction tasks that subjects had worked on in phase 1. Model predictions on these phase 2 tasks were produced by **M**, which has an APE of 30% or less on all 10 tasks, and the average APE of **M** across these 10 tasks is 16.9%.
- **Shifted distribution**: The houses in phase 2 were all selected from *Cluster 2*. That is, the prediction tasks subjects needed to work on in phase 2 came from a *different* distribution compared to the prediction tasks that subjects had worked on in phase 1. Model predictions on these phase 2 tasks were again produced by **M**, which *systematically underestimated* the price of these houses. Specifically, **M** has an APE of 30% or less on *none* of these 10 tasks, and the average APE of **M** is 39.1% in phase 2. This is designed to reflect the realistic real-world scenario that model performance degrades in a novel environment with different data distribution, making blindly relying on the ML model a suboptimal choice under covariate shift.

In addition, we speculated that people might inappropriately rely on ML models when covariate shift occurs if (1) they are not able to recognize the data distribution has changed, or (2) they are not aware that an ML model's performance can be different on different data and over-generalize the model's observed performance on one data distribution to another. To help people better determine whether to rely on ML models when covariate shift occurs, we designed two interventions to assist people in addressing these possible limitations. Specifically, we randomized subjects into one of the three levels of interventions:

- **None (Control)**: Subjects did not receive any external interventions to help them appropriately rely on ML models.
- **Education on Performance of ML Models**: We provided written materials to educate subjects about the possible performance disparity of ML models on different data at the beginning of the experiment before subjects started to work on any decision-making tasks in phase 1 (see Figure 1a). More specifically, we presented subjects with a recent research finding which showed that commercial face recognition systems have different error rates on faces from different demographic groups [28]. We explained to subjects that one possible reason underlying such performance disparity could be the unbalanced training dataset and that when the face recognition system was trained mostly using or only using faces of a certain subgroup of people, its high performance on this subgroup of people does not mean its performance on faces from other subgroups of people would be equally high. We emphasized that this phenomenon is not unique for face recognition systems but is with any kind of ML models. We asked subjects to be aware of the possible performance difference on different data when using ML models. At the end of the page, subjects were asked to answer an understanding question regarding the information they just read. They could only proceed to work on decision-making tasks in phase 1 after they answered the question correctly.

Task distributions	Stationary distribution			Shifted distribution		
Interventions	**Control**	**Education**	**Visualization**	**Control**	**Education**	**Visualization**
N	95	96	93	88	85	92
Expertise in house price prediction (mean)	3.0	3.3	3.3	3.2	3.3	3.2
Expertise in machine learning (mean)	3.3	3.4	3.4	3.3	3.4	3.3
Phase 1 average APE (median)	0.35	0.34	0.35	0.34	0.33	0.32

Table 2: Subjects' expertise and prediction performance in phase 1. We excluded the first task when computing each subject's phase 1 average APE, since subjects may need to use the feedback of the house's actual price they received from the first task to calibrate their predictions.

- **Visualization of Task Properties and Model Performance**: We provided subjects with an interactive tool to visualize data distributions as well as the ML model's performance on different data. Specifically, on the mid-point feedback page, along with the table summarizing the APE of the model and the subject in phase 1, an interactive 2D plot was also presented. In this plot, the subject could freely determine what x- and y-axis of the plot represents by choosing from the list of house features. Each of the 10 houses in phase 1 was shown as a circle on this plot based on its values on the chosen features on the axes, and its color reflected the model's performance on it. The default features on the axes were "living area size" and "quality of the house," the two features that can well separate the two clusters of houses. We encouraged subjects to change the axes to different features to "explore when the model performs well or poorly, and when you have little evidence about how well the model performs." In addition, in each of the tasks in phase 2, besides presenting the feature information of the house to subjects, we also used a similar interactive 2D plot to help subjects compare the house in the current task with the houses that she had predicted prices for in phase 1—houses in phase 1 were again shown as circles with their colors representing the model's performance, while the house in the current task was shown as a star on the plot (see Figure 1b).

4 RESULT

In total, 600 unique subjects participated in our experiment. We considered subjects whose predictions were less than $1,000 for more than 5 out of the last 9 tasks in phase 1 as not paying attention[2]. After filtering out the inattentive subjects, we were left with the data from 549 unique subjects, and our analyses were conducted on these data.

On average, a subject in our experiment spent 15.6 seconds on each prediction task. Table 2 compares subject's self-reported expertise in house price prediction, expertise in machine learning as well as their prediction performance in phase 1 across treatments. We did not find significant differences across treatments on subjects' demographics, self-reported expertise in house price prediction, expertise in ML, as well as their prediction performance in phase 1. Moreover, across all treatments, the median value for a subject's average APE in phase 1 was 34%, which was not far away from either the model's average APE in phase 1 (28.3%) or the bonus threshold we set in the experiment (30%). This confirms that subjects in our experiment would likely perceive themselves as having some degree of capability in making accurate house sale price predictions after completing phase 1.

4.1 RQ1: How Do People Rely On ML Models under Covariate Shift?

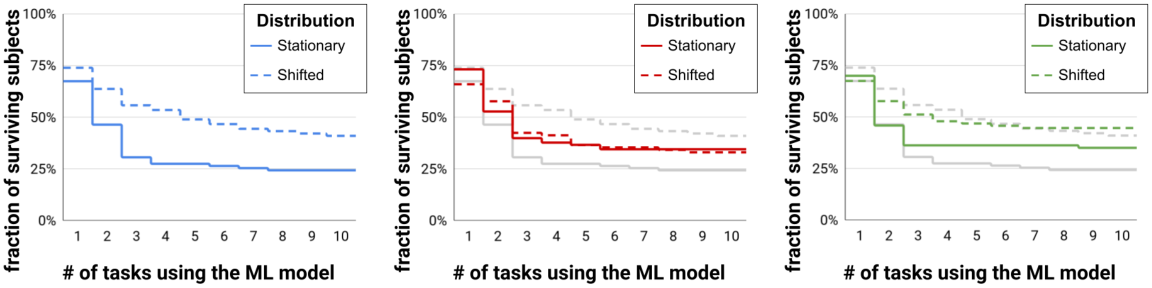
We start by examining that without any external interventions, how subjects would adjust their reliance on ML models under covariate shift (**RQ1**). Specifically, we quantified a subject's reliance on the ML model using the number of tasks that the subject authorized the ML model to make predictions on behalf of her in phase 2—intuitively, the larger the number, the more the subject relied on the model. Figure 2a shows the survival curves—the proportion of subjects whose usage of the ML model in phase 2 would reach a certain number of tasks—for the two control treatments (i.e., the treatments with no external interventions). Surprisingly, we found that subjects seemed to rely on the ML model *more* when the prediction tasks they needed to work on in phase 2 came from a different distribution than those in phase 1, that is, subjects *increased* their reliance on the ML model under covariate shift. A Wilcoxon rank-sum test further confirmed that this increase was significant ($Z = -2.50, p = 0.012$). Recall that the design of our experiment implies that the ML model would have a poor performance in phase 2 when tasks in phase 2 come from a different distribution compared to tasks in phase 1. That is to say, by increasing their reliance on the ML model on the shifted distribution of tasks, subjects actually exhibited a degree of *over-reliance* on the model under covariate shift.

4.2 RQ2: Can External Interventions Promote Appropriate Reliance under Covariate Shift?

Next, we move on to examine whether the use of external interventions could promote appropriate reliance on ML models under covariate shift (**RQ2**). Comparing subjects' reliance on the ML model on out-of-distribution tasks between the treatments with or without external interventions in Figures 2b and 2c, it seems that educating subjects about ML models' possible performance disparity on different data could decrease subjects' over-reliance on ML models on out-of-distribution data to some degree, but the impact of the visualization intervention was not obvious.

To rigorously test the effects of external interventions on influencing subjects' reliance on the ML models under covariate shift, we fitted a regression model to predict a subject's reliance on the ML model in phase 2, considering the main effects of task distribution and intervention, as well as the interaction effects between these

[2] The actual sale prices for houses in phase 1 were all above $60,000. Since the actual prices were all revealed to subjects as feedback in phase 1, subjects should be able to learn the magnitude of the house prices after completing the first task if they were paying attention.



(a) Control (b) Education intervention (red) (c) Visualization intervention (green)

Figure 2: Survival curves showing the fraction of subjects in each treatment who still used the ML model to make predictions after completing X tasks in phase 2. In Figures 2b and 2c, survival curves for the control treatments are shown in grey lines as references.

	Model Parameters				
Variable	$\hat{\beta}$	Std. Error	z value	p value	
Intercept	3.232	0.447	7.231	< 0.001	***
Education	0.904	0.630	1.434	0.152	
Visualization	0.715	0.635	1.125	0.261	
Shifted Distribution	1.893	0.645	2.939	0.003	**
Education × Shifted	-1.888	0.915	-2.064	0.039	*
Visualization × Shifted	-0.894	0.907	-0.984	0.326	

Table 3: Regression models for predicting the number of tasks subjects would delegate the decision-making right to the ML model in phase 2 in different treatments; the treatment with stationary task distribution and no intervention was used as the reference. *, **, and * represents the statistical significance level of 0.05, 0.01, and 0.001, respectively.**

two factors. Results of this model are reported in Table 3. We first noticed that subjects who worked on a shifted distribution of tasks in phase 2, in general, chose to rely on the ML model significantly more ($p = 0.003$). However, we found a significant *negative* interaction term between the education intervention and the shifted data distribution. This means that when making predictions on out-of-distribution tasks, compared to subjects who did not receive any intervention, subjects who were told the possible performance disparity of an ML model significantly decreased their reliance on the model thus showed more appropriate reliance on the model ($p = 0.039$). In contrast, enabling subjects to visualize the distributions of decision-making tasks and the model's performance on different tasks did not significantly influence subjects' reliance on the model on out-of-distribution tasks ($p = 0.326$).

4.3 Exploratory Analyses

So far, we have learned that without any external interventions, laypeople tend to rely on an ML model more when covariate shift occurs, despite the model's poor performance on out-of-distribution data. Meanwhile, the education intervention could promote laypeople's appropriate reliance on the ML model under covariate shift, but the visualization intervention could not. To understand *why* laypeople behave in this way, we conducted a set of exploratory analyses.

4.3.1 Why do people overly rely on the ML model under covariate shift? To gain a better understanding of why subjects relied on the ML model more under covariate shift when no external intervention was provided, we restricted our attention to subjects in the two control treatments, and we analyzed these subjects' responses in the exit survey on their perceptions of the prediction tasks as well as their belief of the model's and their own prediction performance.

First, we noticed that even without any external intervention, people have some capability in detecting covariate shift as subjects in the shifted distribution treatment perceived the houses in the two phases to be more different than subjects in the stationary distribution treatment. Specifically, in the exit survey, subjects were asked to respond to our survey question **S3** ("are the houses you saw in phase 2 similar to those that you saw in phase 1?") using a 5-point Likert scale from 1 (totally different) to 5 (totally same). On average, subjects in the shifted distribution treatment reported the phase 2 houses to be more different than those in phase 1 ($M = 2.63, SD = 1.16$) compared to subjects in the stationary distribution treatment ($M = 3.09, SD = 0.81$), and a two-sample t-test suggested that the difference was significant ($t(155) = 3.149, p = 0.002$).

In addition, in the exit survey, subjects were also asked to compare their perceived model's prediction performance in phase 2 and phase 1 (i.e., survey question **S4**), as well as their perceived prediction performance of themselves in the two phases (i.e., survey question **S5**). We next analyzed subjects' responses to these two comparison questions, considering both the actual task distribution that subjects worked on in phase 2 (stationary vs. shifted), and subjects' perceptions on whether the task distribution in phase 2 had changed from that in phase 1 (yes vs. no)[3].

Interestingly, as shown in Figure 3 (top panel), subjects' belief of the ML model's performance in phase 2 was *not* significantly influenced by either the actual task distribution that they worked on in phase 2, or their perceptions of whether the task distribution they worked on in phase 2 was different from that in phase 1 or not. However, we got a different story when examining subjects' belief of their own prediction performance in phase 2 (Figure 3, bottom panel). Using a two-way ANOVA, we found that although

[3] When a subject's response to the phase 1 vs. phase 2 house similarity question (i.e., **S3**) in the exit survey was below 3 on a 5-point scale, we categorized this subject as believing the task distribution has changed; otherwise, we categorized the subject as believing the task distribution has not changed.

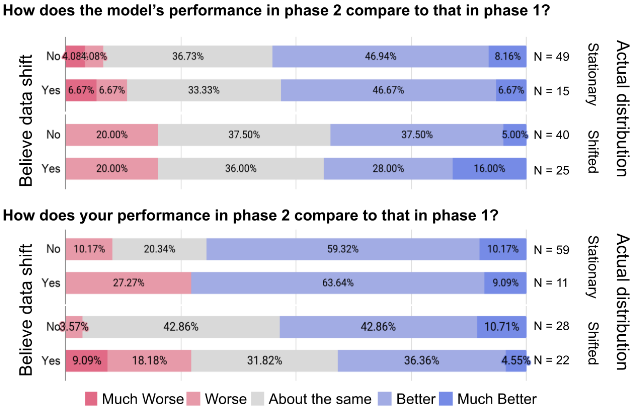


Figure 3: How actual task distribution of phase 2 (stationary vs. shifted) and subjects' perceptions on whether task distribution has changed across the two phases affect subjects' belief in the model's prediction performance (top panel) and their own performance (bottom panel) in phase 2.

the actual task distribution subjects worked on in phase 2 did not significantly change subjects' belief of their own prediction performance in phase 2 ($F(1, 116) = 1.29, p = 0.259$), subjects' perceptions of the task distribution in phase 2 do ($F(1, 116) = 5.83, p = 0.017$)—when a subject felt the task distribution had changed between the two phases, she was significantly more likely to believe her own prediction performance in phase 2 was *worse* than that in phase 1. Lastly, Figure 4 shows the survival curves for subjects in the control treatments who believed the task distribution had or had not changed between the two phases, separately. Clearly, subjects who believed houses in phase 2 were different than houses in phase 1 tended to rely on the ML model more in phase 2, compared to subjects who believed houses in the two phases were similar.

Putting these analyses together, we find a plausible explanation for why laypeople overly rely on the ML model under covariate shift. When covariate shift occurs, a significant portion of laypeople can recognize the change in data distributions without external interventions. However, when people feel the data distribution has changed, they do not expect the ML model's performance on out-of-distribution data to be much different than that on in-distribution data, but they feel their own prediction performance would *decrease* on out-of-distribution data. As a result, those who detect the change of data distributions actually decide to rely on the model *more*, leading to over-reliance on the model under covariate shift.

4.3.2 Why the education intervention could promote appropriate reliance? To answer this question, we conducted the analogous analyses as those in Section 4.3.1 on the data that we obtained from the two treatments with the education intervention. Consistent with the findings in Section 4.3.1, we found that subjects who received the education intervention were able to detect changes in the task distribution, and subjects who believed the data distribution had changed across the two phases also perceived the model's performance to be unchanged on out-of-distribution data compared to that on in-distribution data, but they perceived their own performance on out-of-distribution data to decrease. However, when we looked into how a subject's reliance on the model in phase 2 was affected by the subject's perception of task distribution change, we found the effect was minimal. That is, for subjects who received the education intervention, after recognizing the change of task distributions, their perceived self-performance decrease on out-of-distribution data *no longer* translates into higher levels of reliance on the ML model under covariate shift.

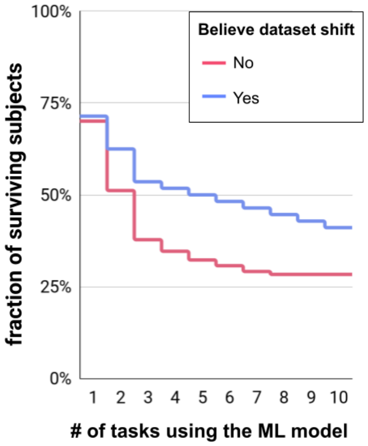


Figure 4: Survival curves showing the fraction of subjects who still used the model to make predictions after completing X tasks in phase 2 for subjects who believed the task distribution had/had not changed across the two phases.

A closer look into subjects' responses to survey question **S6** (i.e., "what are the factors that make you stop using the model in phase 2") provided us with some explanations for this observation—56% of the subjects who received the education intervention and worked on the shifted task distribution in phase 2 indicated that one of the reasons for them to stop using the ML model in phase 2 was that they worried the model's performance would get worse on a new distribution of data; this was a 47.4% increase from the average fraction of subjects in other treatments choosing this option as their reason for stopping using the model. In other words, the education intervention may have decreased people's over-reliance on the ML model under covariate shift mainly by *raising people's awareness* of ML model's possible performance degradation on the out-of-distribution data.

4.3.3 Why the visualization intervention is not effective in promoting appropriate reliance? Finally, we turned our attention to subjects who received the visualization intervention to understand why the provision of the interactive visualization tool did not effectively promote subjects' appropriate reliance on the ML model under covariate shift.

To our surprise, we found no significant difference in the perceptions on whether phase 2 houses were similar to phase 1 houses or not between subjects who worked on the stationary or shifted distribution of tasks in phase 2. This suggests that subjects who received the visualization intervention seemed to have limited capability in detecting covariate shift. We suspect this may be caused by subjects' insufficient engagement with the visualization tool, as more than half of the subjects had never changed the axis combinations of the interactive 2D plot throughout the experiment. We thus analyzed the data again by taking subjects' engagement with the visualization tool into consideration.

Specifically, using the average number of times that subjects in the visualization intervention treatments changed the axis combination in the 2D plot (i.e., 6.13) as a threshold, we separated subjects into two groups whose engagement level with the visualization tool was high or low. We then used a two-way ANOVA test to examine how a subject's perception of phase 2 tasks was influenced by the actual task distribution the subject worked on in phase 2 and the subject's engagement level with the visualization tool. Interestingly, we detected a significant interaction effect ($F(1, 174) = 4.076, p = 0.045$): When working on the stationary distribution of tasks in phase 2, high engagement subjects had similar perceptions on the similarity of phase 1 and phase 2 houses as low engagement subjects. However, when subjects worked on a shifted distribution of tasks in phase 2, high engagement subjects were more likely to perceive a change in the task distribution than low engagement subjects. This means that subjects who had a sufficient amount of interactions with the visualization tool were still able to detect the covariate shift. Unfortunately, the high engagement subjects' reaction to the detected covariate shift was to increase their reliance on the ML model—we again found a significant interaction between the actual task distribution a subject worked on in phase 2 and the subject's engagement with the visualization tool in affecting the subject's reliance on the model in phase 2 ($F(1, 181) = 8.888, p = 0.003$). In particular, high engagement subjects chose to rely on the ML model for significantly more tasks only when a shift of task distribution happens ($p = 0.015$).

5 DISCUSSION

In this section, we relate our findings to prior work, provide design recommendations for promoting laypeople's appropriate reliance on ML models under covariate shift, and highlight opportunities for future research. We also discuss limitations of our work, cautioning readers to generalize our results to other settings.

Clearly and transparently communicate the intended use cases of a ML model to end-users. Previous research has identified that a key principle for promoting appropriate reliance in AI is to ensure people are well informed of the capabilities and limitations of an AI system [1]. Detailed recommendations have been provided on the kind of information that needs to be communicated to end-users in a transparent model reporting process, including the use cases of the model that were envisioned during model development, benchmarked evaluation of the model in a variety of conditions, and relevant details of training data of the model [17]. Our findings in this work warn about the risk of laypeople overly relying on poorly-performed ML models under covariate shift if these recommendations are not followed. In other words, one of the most straightforward implications that we have learned from our study is that designers of ML models should clearly and transparently communicate the scope of application and potential limitation of their model to the end-users, as this information would likely be critical for helping laypeople form right expectations about the model's performance on different data and therefore avoid inappropriate reliance.

The need of increasing people's understanding of performance of ML models. Our results that providing people with a brief education session on ML model's possible performance disparity on different data reduces people's over-reliance on the ML model under covariate shift highlights the importance of user education. Indeed, without a solid understanding of ML model's performance, people may still exhibit inappropriate reliance on ML models under covariate shift even when they manage to identify the changes in data distributions, because they might have mistakenly generalized the model's observed performance from one data distribution to another. On the high level, this is related to increase the general public's AI literacy [9, 14], especially on enabling people to recognize that ML models learn from data, and the data used to train the models largely influences the results of the model. While in our study, we used simple written material to help people better understand the performance of ML models, future work could explore the effectiveness of different methods (e.g., interactive tutorial [10]) for educating end-users on both general knowledge of AI and specific information related to properties of ML models under covariate shift. An alternative approach for increasing people's understanding of performance of ML models is to design better ways to communicate an ML model's performance to end-users that discourage over-generalization of model performance into novel data distributions, which provides rich opportunities for future research in uncertainty quantification and communication [27].

Towards more effective visualizations. To encourage people to trust/rely on ML models appropriately, previous research has explored the usage of interactive visualization to explain the model predictions [5, 11]. Different from these studies, our interactive visualization is designed to explain the data (i.e., the decision-making tasks) as well as the model's performance on different data. Our observation that the presence of this visualization tool hampers people's capability of detecting changes in task distributions is deviated from what we have expected. This might be partly caused by laypeople's limited data visualization literacy [3]. In fact, previous literature has shown that even experienced users like data scientists or students with advanced coursework in science and mathematics sometimes can not precisely interpret the outputs of data visualization [16]. On the other hand, our results show that subjects who are highly engaged in interacting with the visualization tool are more likely to detect changes in the task distributions. Future research could consider designs that encourage people's interactions with the visualization as a possible direction for improving the effectiveness of the visualization. Finally, we note that our visualization does not help those people who successfully identify the changes in the task distributions to reduce their reliance on the ML model when covariate shift occurs. This means that when people see a data point lying far away from the set of data points on which they have observed the model's performance, they might fail to recognize it as an indication of lacking evidence for the model's performance on that data point. Aside from educating people about an ML model's potential performance drop on novel data distribution, this also highlights the need of actively assisting people to interpret the information carried in data visualization.

Limitations. We conducted our experiment on a crowdsourcing platform (MTurk) on one particular type of task (i.e., house price prediction). While the worker population on MTurk well represents the laypeople population, which serves our purpose of studying how laypeople would rely on ML models under covariate shift well, cautions should be used when generalizing our findings to a different population, such as domain experts. In fact, we have observed that a small proportion of subjects in our experiment largely outperform the model in phase 1, and most of these subjects choose to not rely on the model at all in phase 2 regardless of whether the task distribution has changed, indicating that experts' reliance on ML models under covariate shift could be fundamentally different from that of laypeople's. The particular nature of crowdsourcing platforms might have also shaped the results we've obtained to some extent. For example, one factor that subjects reported as the reason for them to always use the model in the experiment was that they hope to "save time and effort" [23]. Finally, we note that how people rely on an ML model under covariate shift might also depend on people's perceptions of the difficulty of the prediction task and the level of model performance that people have observed on the in-distribution data; thus, we caution the readers from generalizing our results to different settings.

6 CONCLUSION

In this paper, we present an experimental study to understand how laypeople adjust their reliance on machine learning models when covariate shift occurs. We find that people increase their reliance on the model on out-of-distribution data compared to that on in-distribution data, both because people are not able to accurately detect the model's performance deterioration on out-of-distribution data, and because people perceive their own decision-making performance on out-of-distribution data to be decreasing. We design two interventions aiming at improving people's appropriate reliance on machine learning models under covariate shift, one focuses on educating people on the possible performance discrepancy of a machine learning model on different data, and another focuses on enabling people to visualize the properties of decision-making tasks and the model's performance on different tasks. We find that people who have learned about machine learning model's possible performance drop on a novel distribution of data are able to decrease their over-reliance on machine learning models when covariate shift occurs. Our work provides important implications for enhancing human-AI partnership in a dynamically changing world, 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] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[2] 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.

[3] 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.

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[5] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[6] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[7] 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).

[8] 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.

[9] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 43–52.

[10] 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.

[11] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[12] Spencer C Kohn, Daniel Quinn, Richard Pak, Ewart J de Visser, and Tyler H Shaw. 2018. Trust repair strategies with self-driving vehicles: An exploratory study. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 1108–1112.

[13] 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.

[14] 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.

[15] 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*.

[16] Adam V Maltese, Joseph A Harsh, and Dubravka Svetina. 2015. Data visualization literacy: Investigating data interpretation along the novice—expert continuum. *Journal of College Science Teaching* 45, 1 (2015), 84–90.

[17] 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.

[18] Jose G Moreno-Torres, Troy Raeder, RocíO Alaiz-RodríGuez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern recognition* 45, 1 (2012), 521–530.

[19] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*. 13991–14002.

[20] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[21] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2009. *Dataset shift in machine learning*. The MIT Press.

[22] 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.

[23] Susumu Saito, Chun-Wei Chiang, Saiph Savage, Teppei Nakano, Tetsunori Kobayashi, and Jeffrey Bigham. 2019. Predicting the Working Time of Microtasks Based on Workers' Perception of Prediction Errors. *Human Computation* 6, 1 (2019), 192–219.

[24] Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* 90, 2 (2000), 227–244.

[25] Masashi Sugiyama and Motoaki Kawanabe. 2012. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation.* MIT press.

[26] 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)*.

[27] Anne Marthe van der Bles, Sander van der Linden, Alexandra LJ Freeman, James Mitchell, Ana B Galvao, Lisa Zaval, and David J Spiegelhalter. 2019. Communicating uncertainty about facts, numbers and science. *Royal Society open science* 6, 5 (2019), 181870.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.