Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks

Zhuoran Lu
Purdue University
lu800@purdue.edu

Ming Yin
Purdue University
mingyin@purdue.edu

ABSTRACT

This paper addresses an under-explored problem of AI-assisted decision-making: when objective performance information of the machine learning model underlying a decision aid is absent or scarce, how do people decide their reliance on the model? Through three randomized experiments, we explore the heuristics people may use to adjust their reliance on machine learning models when performance feedback is limited. We find that the level of agreement between people and a model on decision-making tasks that people have high confidence in significantly affects reliance on the model if people receive no information about the model's performance, but this impact will change after aggregate-level model performance information becomes available. Furthermore, the influence of high confidence human-model agreement on people's reliance on a model is moderated by people's confidence in cases where they disagree with the model. We discuss potential risks of these heuristics, and provide design implications on promoting appropriate reliance on AI.

CCS CONCEPTS

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

KEYWORDS

Machine learning, appropriate reliance, human-AI interaction

ACM Reference Format:
Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3411764.3445562

1 INTRODUCTION

AI-assisted decision-making has become increasingly ubiquitous over the past few years. From music recommendation [56, 60] to financial risk assessment [18, 28] to medical diagnosis [15, 25], various AI-driven decision aids, which are often powered by machine learning (ML) models, have been built and applied in a wide range of domains to aid humans in making better decisions. As a result, an increasingly large number of people are interacting with and potentially influenced by these ML models in their decision-making.

A critical step in advancing our knowledge of people's usage of AI-driven decision aids is to understand how do people determine how much they could rely on the ML models underlying these decision aids and adopt the recommendations they provide. Perhaps intuitively, an important kind of information that people utilize to calibrate their reliance on an ML model is the objective feedback on the model's performance. Indeed, previous research has shown that people adjust their reliance on an ML model based on its accuracy [30, 64], and correctness feedback on the level of individual decision-making task further helps people build "mental models" of the ML model by understanding when the model is likely to err [3, 4]. Information on model's confidence, which is the model's own estimate of its likelihood of being correct, is also shown to significantly affect how much people would rely on the model [65].

However, there are cases where external performance feedback of an ML model is not readily available. For example, model designers may fail to transparently communicate the model's performance to its end-users. Significant time delays may exist before one can meaningfully evaluate the performance of a model (e.g., ML models assisting college admission, long-term investment, and matchmaking). Sometimes, it is even impossible to fully observe the model's performance due to the decisions made (e.g., decide not to admit a student into college, thus observing how this student would do if admitted becomes impossible). An interesting question, then, is when salient performance feedback of an ML model is *absent* or *scarce* (e.g., performance information is not available or only available on the aggregate level) during people's interaction with an ML model, how would people decide how much to rely on the model? With limited performance feedback, are there any *heuristics* that people utilize to adjust their reliance on an ML model?

While few research has systematically examined how people would rely on an ML model when they have little information about how well the model performs, decades of research in social psychology on how people take the advice from other people into consideration reveals some consistent behavior pattern [33, 39, 46, 50, 63]—for example, the phenomenon of "*naive realism*" [61] suggests that people often consider their own judgment to be objective reflections of reality and tend to discount advice that is more different from their own opinion, while the "*agreement-in-confidence heuristic*" [46] further proposes that people may also utilize their own internal decision confidence in deciding how much to down-weight the opinion from a disagreeing advisor. Inspired by these findings, we hypothesize that people may adopt a similar strategy in determining how much to rely on an ML model when there is limited performance feedback. For example, people may believe their own

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '21, May 8–13, 2021, Yokohama, Japan
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445562

decisions on those tasks that they feel confident about to be "correct." As a result, they may form a subjective evaluation of an ML model's performance based on how often the model agrees with themselves on tasks that they are confident—the more often the model matches people's decisions, the more accurate people feel the model is, and the more they would be willing to rely on it.

To validate this hypothesis and thoroughly examine how people adjust their reliance on an ML model when they have limited access to performance feedback of the model, in this study, we conducted a series of three human-subject randomized experiments, focusing on answering the following questions:

- When people receive no information about an ML model's performance, does the level of agreement between people and the model on tasks that people have high confidence in affect people's reliance on the model?
- If so, does it continue to do so after people have had the opportunity to obtain some aggregate information about the model's performance (e.g., the model's overall accuracy on a set of decision-making tasks) in practice?
- In the real world, people may encounter both cases that they feel confident and cases that they are not confident when interacting with an ML model. How does the people's own confidence in those cases that they agree or disagree with the model change their reliance on the model?

In our experiments, subjects were recruited from Amazon Mechanical Turk to make a sequence of predictions on the outcomes of speed dating events with the help from ML models. The prediction tasks were divided into two phases. In both our first and second experiment, subjects in different experimental treatments worked on exactly the same tasks, and the first phase was composed of only tasks that people would feel confident about their own decisions according to a pilot study we conducted. Then, in the first experiment, subjects were randomly assigned to one of the three treatments, and the ML model used in different treatments made different predictions in the first phase, leading to varying levels of agreement between subjects and the model on tasks that subjects had high confidence in. In addition, subjects received no performance feedback about the ML model at any point throughout the experiment. Our second experiment was completely analogous to the first experiment, except that we revealed the model's overall accuracy in Phase 1—which was designed to be either relatively low (50%) or high (80%)—to subjects between the two phases. Finally, in our third experiment, in order to vary people's confidence in their agreement or disagreement with the model, we used different tasks in the first phase of the experiment across the 4 experimental treatments, which were arranged in a 2×2 design along two factors—people's confidence in their own decisions when they agree with the model (high vs. low), and people's confidence in their own decisions when they disagree with the model (high vs. low). In all three experiments, we measured subject's reliance on the model through their willingness to follow the model's recommendations in Phase 2, and we also collected information on subject's perceptions of the model on a variety of reliance-related factors through surveys.

Our results show that when there is limited performance feedback about an ML model, people adjust their reliance on the model in a sophisticated way—When they have no information about a model's performance at all, people's reliance on the model is significantly influenced by the level of agreement between the model and themselves on tasks that they have high confidence in. However, once people have obtained some aggregate-level performance information about the model, people's reliance on the model is mostly affected by the model's observed performance, but not the level of high confidence agreement with the model any more. Finally, we detect an interaction effect between people's confidence in their agreement with a model and people's confidence in their disagreement with the model in influencing people's reliance on the model. For example, having high confidence disagreement with an ML model would reduce people's reliance on the model if people agree with this model mostly on tasks that they are confident, but it brings about no impact on reliance if people agree with this model mostly on tasks that they have low confidence in.

Taken together, our study shows that with limited performance feedback of an ML model, people indeed adopt some heuristics to adjust their reliance on the model based on the level of agreement between the model and themselves. Moreover, how the human-model agreement/disagreement affects reliance on the model varies with people's confidence in their agreement and disagreement with the model, and it may also change after people have obtained some aggregate information about the model's performance in practice. We highlight that the usage of these heuristics have both benefits and risks. Notably, when people are not aware of their own limitations in decision making (e.g., biases or poor degree of calibration), they may show over-reliance on models sharing the same biases as themselves and show under-reliance on complementary models. On the other hand, by overly relying on a model's observed performance in practice to gauge how much to rely on the model, people might miss the opportunity to leverage useful information carried in their agreement and disagreement with the model to further differentiate reliable models from the unreliable ones. We conclude by discussing the design implications and limitations of our work.

2 RELATED WORK

The rapid development of automation over the past decades has inspired active research in understanding reliance on automation [10, 44, 49]. For example, researchers have identified that reliance on automated systems could be affected by factors related to humans (e.g. trust disposition, affective process) [38], factors related to the automation (e.g. error types) [52], as well as factors related to the context (e.g. time pressure, risk) [8]. Theoretical models and frameworks have been proposed to explain human's reliance on the automation in various situations [12, 19, 20, 31, 57]. In practice, however, collaborations between humans and automation often fail to achieve optimal outcomes due to human's inappropriate reliance on the automation [13, 43], including both over-reliance on the automation when it performs poorly (i.e., misuse) and under-reliance on the automation when it is well-functioning (i.e., disuse).

More recently, with the widespread usage of machine learning (ML) based technologies in aiding human decision-making, a new line of research examining human's reliance on ML emerges [22, 55, 64]. Compared to traditional automated systems, ML models are built on a massive amount of data. As a result, they often become

highly sophisticated and sometimes even opaque in order to capture the complex nonlinearity in real-world data, and uncertainty is baked into every prediction of ML models. Correspondingly, studies on how people rely on ML models often emphasize on people's capability of utilizing external, explicit information about a model, such as explanations and uncertainty qualifications of the model's decisions to adjust their reliance on the model. For example, it was shown that people's reliance on an ML model is affected by the model's stated accuracy on a set of held-out data [30], the model's observed accuracy on real-world trials [64], and the model's confidence associated with each individual recommendation [65]. On the other hand, while a number of studies have shown the promise that explanations of an ML model enhance people's understanding of the model and therefore lead to more appropriate reliance on the model [30, 48, 62], the impact of model intelligibility and transparency on reliance also seems to vary across decision-making tasks of different properties, explanations of different types, and people with different characteristics [7, 9, 47, 53, 65].

In comparison to the growing literature on how people's reliance on an ML model is affected by external information about the model, much less attention has been paid to understand how people's reliance on an ML model is determined when people have limited access to such information. In a different context of human advice-taking, however, psychologists have revealed rich insights into how people decide whether to rely on another human advisor's advice during their decision making without objective feedback on the quality of the advice. For example, research has identified that people exhibit a degree of "*naive realism*" [23, 61], that is, they tend to believe they perceive the world objectively and people who disagree with themselves are uninformed or biased. As a result, people often discount or even ignore the advice provided by advisors who disagree with them [14, 33, 39, 63]. In practice, it was shown that using the frequency of agreement between oneself and an advisor as a proxy can help reliably estimate the quality of the advisor's advice when judgements between the person and the advisor are independent, but could also cause systematic errors in decision-making when the person suffers from the "*false-consensus effect*" [36, 50, 63] and shows more trust to advisors who simply share the same biases as themselves [46]. Moreover, as humans often use their own confidence as a way to express their estimated likelihood of making correct decisions [17, 45], it was found that both the advisor and advisee's confidence in their own decisions, as well as the advisee's confidence in their agreement with the advisor, affects the likelihood of the advice being taken [46, 54, 58]. In particular, previous study showed that when the agreement between advisor and advisee mostly occurs on cases that the advisee is confident, the advisee shows higher levels of trust towards the advisor, and also gets influenced by the advisor more [46].

Inspired by these findings, we hypothesize that when interacting with an ML model, people may also use their agreement with the model as a heuristic to gauge how much they could rely on the model when performance feedback of the model is limited. To this end, while a few recent studies have examined how the alignment between a model's explanations and human logic affects people's reliance on the model [41, 66], there is no systematic investigation into how the agreement between people and a model on decisions they make affects reliance, thus our study fills in this gap.

3 EXPERIMENT 1

In Experiment 1, we set out to understand how do people adjust their reliance on a machine learning (ML) model when they have *not* obtained any feedback on how well the model performs. We speculate one heuristic that people may adopt in such a scenario is to use the level of agreement between the ML model and themselves on tasks that *they are highly confident about their own predictions* as a proxy to estimate the performance of the model and adjust their levels of reliance accordingly. Specifically, we hypothesize that:

- **[H1]** When people receive no information about an ML model's performance, the level of agreement between people and the model on tasks that people have high confidence in significantly affects people's reliance on the model; the higher the level of agreement, the more people rely on the model.

To formally test this hypothesis, we conducted a randomized experiment where human subjects were recruited from Amazon Mechanical Turk (MTurk) to complete a sequence of 30 decision-making tasks, with the help of different ML models, which showed varying levels of agreement with subjects on tasks that subjects had high confidence in.

3.1 Experimental Task

The decision-making task that we asked subjects to complete in this experiment is to predict the outcome of speed dating events. This task is suitable for our experimental study because human subjects do not need to possess specialized knowledge to make predictions on such a task, and subjects may feel confident about their own predictions on some of these tasks. Moreover, it also represents a realistic scenario where people can be assisted by an ML model to make better decisions. Similar tasks have been used in previous research to understand how performance information of an ML model affects people's trust and reliance on ML models [64].

Specifically, in each task, the human subject was presented with a set of information about one participant in a speed dating event and his/her date, including a total of 22 features on (1) basic demographics of the participant and the date (e.g., gender, age, race), (2) the participant's preferences in romantic partners (e.g., out of 100 points, how many points does the participant assign to attractiveness when evaluating whether a person is an ideal romantic partner or not?), and (3) the participant's impression of the date (e.g., the participant's rating of the attractiveness of the date). After the subject reviewed these information, she was asked to make a binary prediction on whether the participant would like to meet with the date again in the future, following the steps as listed below:

- First, the subject was asked to make an initial prediction *on her own* about whether the participant in this event wanted to see his or her date again.
- Then, the subject was presented with the information on an ML model's binary prediction on this task, and accordingly, whether the model agreed with her or not.
- Finally, the subject was asked to submit a *final* prediction.

Figure 1 shows an example of the interface of the prediction task in our experiment. The information on speed dating participants and their dates that we showed to subjects in each task were taken

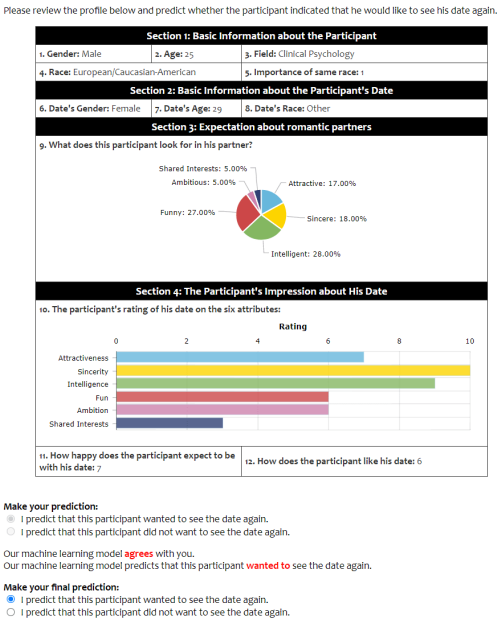


Figure 1: Interface of the prediction task.

from real speed dating events that were conducted in the experimental study of Fisman et al. [16], which also provided us with the ground truth answer of whether participants in each speed dating event were willing to see each other again.

Task Instance Categorization. To enable our experimental manipulation (i.e., varying the level of agreement between an ML model and subjects on tasks that subjects are highly confident; will be detailed in Section 3.2), we first conducted a pilot study on a set of 214 speed dating outcome prediction tasks to collect necessary information on them, including what the subjects' majority prediction on a task is, whether the majority prediction is correct, and how confident subjects are about their predictions. A total of 315 subjects were recruited from MTurk to take our pilot study, in which they needed to complete a random sample of 20 prediction tasks, and for each task, they were also asked to indicate their confidence level in their prediction. Based on these data, for each task, we determined subjects' majority prediction on it, the accuracy of the majority prediction, as well as the average confidence of subjects' predictions on the task. For more details of the pilot study, see the supplementary materials (SM) Section 2.

Using the first quartile and third quartile of the average confidence on a task as the thresholds, we identified a set of 54 tasks that subjects are highly confident about their predictions (i.e., subjects' average confidence is above Q3) and 52 tasks that subjects have low confidence in their predictions (i.e., subjects' average confidence is below Q1). Note that when subjects are highly confident about their own predictions, they are *not* necessarily correct. This could be caused by a number of reasons, including subjects overestimating or underestimating the importance of certain features in influencing the speed dating outcome, subjects' over-confidence in their own predictions, and a degree of uncertainty inherent in predicting human behavior.

3.2 Experimental Design

Each subject completed exactly the *same* 30 prediction tasks in our experiment, which were divided into two phases. Phase 1 consisted of the first 20 tasks in the sequence, and we varied the level of agreement between people and the ML model on these tasks to create our experimental treatments. Phase 2 included the rest 10 tasks, and was designed for measuring how much people would be willing to rely on the ML model in their decision-making.

Experimental Treatments. Three treatments were created in this experiment. Each treatment was associated with a unique ML model, and the three models differed on how often their predictions agreed with those of subjects' on *Phase 1* tasks. Importantly, the 20 prediction tasks of Phase 1 were all selected from those *high* confidence tasks that we identified from our pilot study. To make it easy to control the level of agreement between the ML models and the human subjects, the 20 tasks we included in Phase 1 further satisfied an additional criterion that for each task, our pilot study had shown that at least 80% of subjects made the *same* prediction on it. Thus, depending on whether people's majority prediction on a task was correct or not, Phase 1 included both tasks on which most subjects would make **c**orrect predictions with **h**igh confidence ("CH" tasks) and tasks on which most subjects would make **i**ncorrect predictions with **h**igh confidence ("IH" tasks).

Given our knowledge of what the majority of people would predict for each task, we then created three experimental treatments where the ML model used in these treatments was designed to make the *same* predictions as the majority of people on 40% (low agreement), 70% (medium agreement), or 100% (high agreement) of the 20 Phase 1 tasks, respectively[1]. To single out the effect of the human-model agreement on reliance, we minimized the differences across treatments by ensuring that the three ML models had exactly the *same accuracy* and the *same positive prediction rate*[2] on these 20 tasks (both accuracy and positive prediction rate were set at 70%). Practically, we did this by first setting the ML model's prediction to be the same as people's majority prediction on all Phase 1 tasks in order to create the "high agreement" treatment. The other two treatments were then created by "flipping" the model's predictions on some tasks, hence decrease the level of agreement between people and the model. Importantly, whenever we had the model disagree with people's majority prediction on a task that the majority prediction is correct, we also needed to have the model disagree with people's majority prediction on another task that the majority prediction is wrong (hence the model accuracy was kept at the same level across treatments). Similarly, changing the model's prediction on a task where the majority of people made a positive prediction was always accompanied by changing the model's prediction on another task for which people's majority prediction was negative (hence the model's positive prediction rate was kept at the same level across treatments).

[1]Predictions of ML models on each task were artificially set based on the designed level of human-model agreement in different treatments. However, we note that in reality, different ML models can be trained to match such predictions, as sophisticated models are shown to be able to approximate any continuous function [32]. Empirical studies also showed that real-world prediction problems often admit competing models that make wildly conflicting predictions [37].

[2]That is, the fraction of tasks where the model predicted "the participant wants to see the date" was kept the same across treatments.

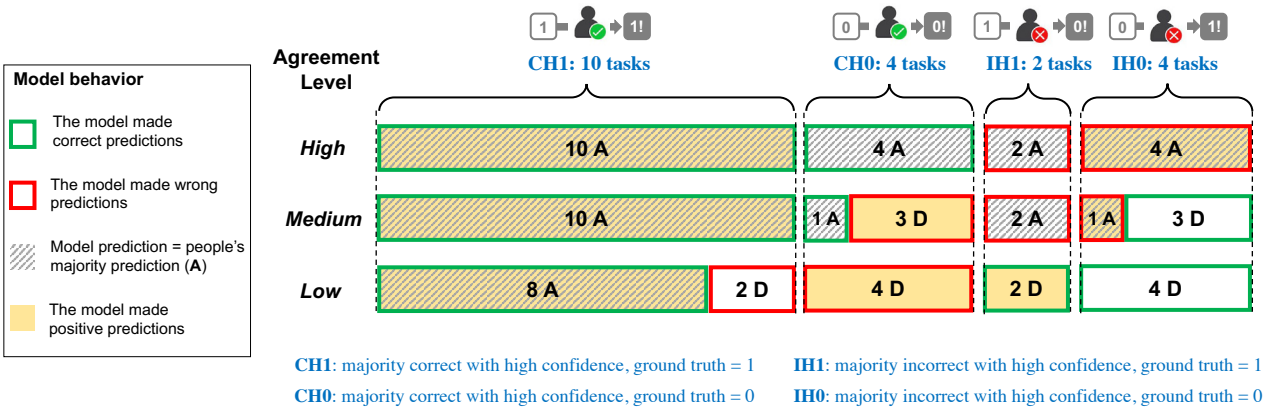


Figure 2: Predictions of the three ML models on Phase 1 tasks of Experiment 1. Each row represents the predictions made by the ML model in one treatment: "A" means the model agreed with the predictions made by the majority of people on the tasks, and "D" means the model disagreed with the majority predictions. Bars with stripes represent tasks on which the model's prediction agreed with people's majority prediction. Bars with the green (red) outline are tasks where the model made correct (wrong) predictions. Bars with the yellow background are tasks where the model made a positive prediction (i.e., predicted 1, the participant wants to see the date again). On the 20 tasks in Phase 1, the three models have the same accuracy of 70%, the same positive prediction rate of 70%, but different levels of agreement with people's majority predictions.

Figure 2 illustrates the details for the 20 task instances that we selected to use in Phase 1 of this experiment, as well as the ML models' predictions on them in the three treatments. With such design, we expect that human subjects who are assigned to the high agreement treatment find that on Phase 1 tasks, there is a higher level of agreement between the predictions made by the model and the *initial* predictions made by themselves, compared to subjects who are assigned to the medium agreement treatment, who in turn have a higher level of agreement with the model than subjects in the low agreement treatment.

Lastly, Phase 2 was designed to measure whether the perceived level of agreement between people and an ML model (in Phase 1) affects people's reliance on the model. Thus, ML models in the three treatments were designed to make exactly the *same* predictions on the 10 tasks in Phase 2. To maximize the possible effect, all Phase 2 tasks were selected from those task instances that people had *low* confidence in their own predictions according to the pilot study. To minimize the difference between the two phases, we made sure that the model's accuracy and positive prediction rate in Phase 2 were still 70%, which was the same as that in Phase 1.

Experimental Procedure. Our experiment was posted as a HIT on MTurk that was open to U.S. workers only, and each worker was allowed to take the HIT only once. Each HIT contained exactly the *same* 30 prediction tasks separated in two phases, and the order of tasks was randomized within each phase. Besides, an attention check question was added at a random position in the task sequence, and the subject was asked to select some pre-specified options in the attention check question.

Upon arrival at the HIT, each subject was first presented with an instruction of the task interface. Then, the subject needed to answer three qualification questions to show that she understood the meanings of various information presented to her on the task interface, and she could only proceed on after correctly answering all these questions. Once qualified, the subject would be randomly assigned to one of the three experimental treatments to work on the sequence of 30 prediction tasks. On each task, the subject was asked to make initial and final predictions following the three-step procedure described above. On tasks in Phase 1, subjects in different treatments saw the ML model making *different* predictions, leading to different perceived levels of agreement with the model. The subject did *not* receive any feedback on whether her own prediction or the model's prediction was correct on each of the tasks. After completing all 20 tasks in Phase 1, the subject was asked to take a pause and reflect on her experience in the first phase of the experiment by filling in a few survey questions. In particular, the subject was asked to first make a guess about the model's accuracy in Phase 1. Then, the subject rated on a 7-point Likert scale about how much she agrees with each of the following statements from 1 (strongly disagree) to 7 (strongly agree):

- **(Competence)** The recommendation that the machine learning model provides to me is as good as that which a highly competent person could provide.
- **(Reliability)** The machine learning model provides the reliable recommendation to me in each task.
- **(Understandability)** I understand how the machine learning model will assist me with predictions I have to make.
- **(Faith)** If I am not sure about my prediction in a task, I have faith that the machine learning model will provide the best solution.

After completing the survey, the subject proceeded to Phase 2 to work on the rest of 10 prediction tasks. Again, the subject was asked to make initial and final predictions on each of the Phase 2 tasks, *without* receiving any feedback on their own accuracy or the model's accuracy on any task. Finally, before submitting the HIT, the subject was asked to answer the following question to indicate her overall trust in the ML model on a 7-point Likert scale, with

1 representing "do not trust the model at all" and 7 representing "fully trust the model.":

- **(Overall trust)** Overall, how much do you trust the predictions of our machine learning model?

The base payment of our HIT was $1.5. To motivate subjects to carefully consider whether and how much to rely on the ML model when making their predictions, we also provided a performance-based bonus to subjects: after the subject completed the HIT, we randomly selected one prediction task in the sequence to check whether the subject's *final* prediction on that task was correct. If so, the subject would receive a $1 bonus on top of the base payment. Note that in this experiment, we *never* provided any feedback to subjects on the accuracy of the ML model on any of the tasks.

3.3 Experimental Data and Analysis Methods

In total, 301 subjects passed the attention check and completed our HIT. For each subject, we recorded her initial prediction and final prediction on each of the 30 prediction tasks. Besides, we also recorded the subject's responses to our survey questions both between the two phases and at the end of the HIT.

Similar to previous studies [30, 64, 65], based on the information we collected in our experiment, we operationalized the measurement of reliance on ML models using two metrics which evaluate subject's willingness to follow the model's predictions in Phase 2:

- **Agreement fraction**: in Phase 2, the fraction of tasks on which subject's *final* prediction agreed with the model's prediction[3].
- **Switch fraction**: in Phase 2, the fraction of tasks on which subject's final prediction agreed with the model's prediction, among all tasks that subject's initial prediction was different from that of the model's.

Intuitively, the higher agreement fraction and switch fraction are, the more subjects rely on the ML model in their decision-making. Since Phase 2 tasks were exactly the same across the three treatments and the model in different treatment also made exactly the same predictions on these tasks, differences in agreement fraction and switch fraction across treatments can be causally attributed to the subject's perceived level of agreement with the model in Phase 1. To formally examine whether the level of agreement between people and an ML model in Phase 1 affects people's reliance on the model, we conducted the one-way analysis of variance (ANOVA) on the two measures of reliance (i.e., agreement fraction, switch fraction) across the three treatments[4].

In addition, to explore the mechanisms through which the human-model agreement on tasks that people have high confidence in affects people's reliance on the model, we also analyzed the data on subject's perceived accuracy, technical competence, reliability, and understandability of the model, as well as subject's faith and overall trust in the model, all of which are previously identified as key constructs that would affect people's reliance in an automated system [8, 31, 35, 51]. In particular, we conducted one-way ANOVA on subject's perceived model accuracy and overall trust on the model. We also conducted proportion tests on the proportion of subjects who agreed/disagreed with the four statements regarding the competence, reliability, understandability of the model, and their faith in the model to see whether subject's perceptions on these aspects differ significantly across treatments[5]. Following all these tests, post-hoc Tukey HSD tests or pairwise proportion tests were then used to identify pairs of treatments in which subjects exhibited significant differences in their reliance on the model or their reliance-related perceptions in the model (we adjusted p-values in post-hoc analyses to control for a family-wise error rate of 0.05).

3.4 Experimental Results

To begin with, we checked whether we successfully created different levels of agreement between subjects and the ML model in different treatments. The average fraction of Phase 1 tasks on which subjects' initial predictions were the same as the model's predictions was 0.446, 0.673, and 0.902 for low, medium, and high agreement treatment, respectively, and one-way ANOVA test confirmed that the difference was statistically significant ($p < 0.001$). This means that the agreement level between people and the model was successfully varied across the three treatments in this experiment.

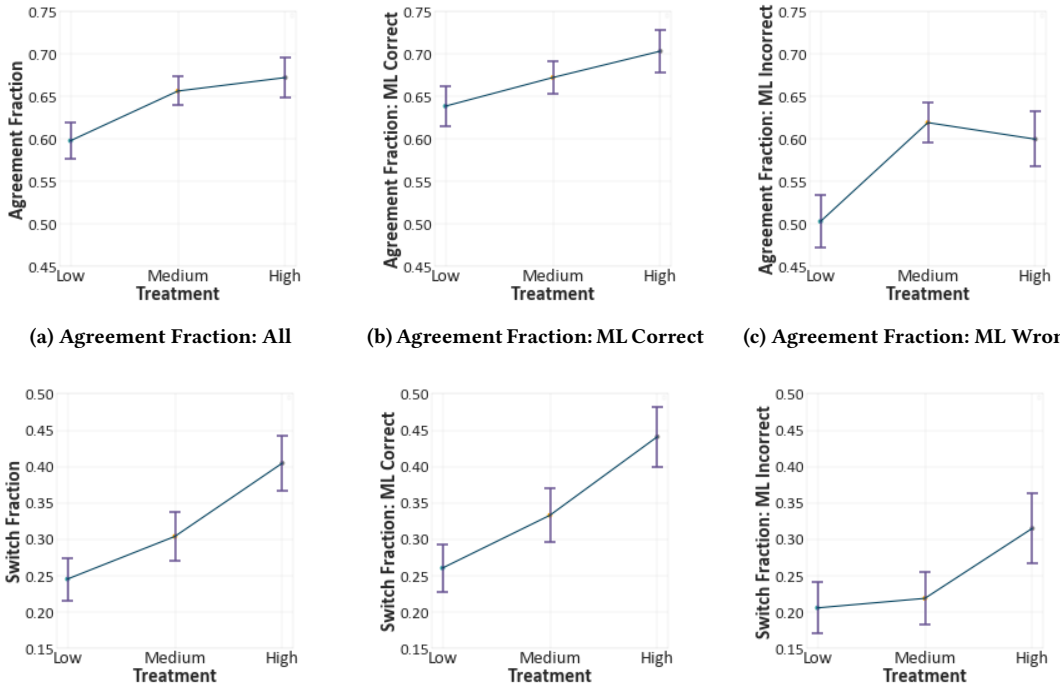
Does High-Confidence Agreement Affects Reliance When Performance Feedback is Absent? We plot the comparisons of subjects' average agreement fractions and average switch fractions on all Phase 2 tasks across the three treatments in Figures 3a and 3d, respectively. Visually, it is clear that when subjects did not know how well a model performs, they were more likely to follow the predictions of the model if there was a higher level of agreement between the model and themselves on tasks that they are highly confident about their own predictions. Our one-way ANOVA test results confirmed that the differences in subject's reliance on the model across treatments were statistically significant (agreement fraction: $F(2, 298) = 3.67$, $p = 0.027$; switch fraction: $F(2, 298) = 5.24$, $p = 0.006$). Post-hoc pairwise comparisons further suggested that the differences in subjects' reliance on the ML model were significant between the low agreement and high agreement treatments ($p = 0.037$, Cohen's $d = 0.34$ for agreement fraction; $p = 0.004$, $d = 0.49$ for switch fraction).

A closer look into the data further suggests that higher levels of high-confidence agreement between an ML model and people lead to increased reliance on the model *regardless of* whether the model's predictions are correct or not. For example, Figures 3b and 3e show subject's reliance on the ML model on Phase 2 tasks where the model made *correct* predictions, while Figures 3c and 3f compare subject's reliance on the ML model across the three treatments on Phase 2 tasks where the model was *wrong*. Although it seems that subjects tended to rely on the model more when the model was correct, we observed the same increasing trend with respect to how the level of human-model agreement on high confidence tasks affects people's reliance on the ML model, both when the model

[3] Agreement fraction concerns how often a subject's *final* prediction son *Phase 2* tasks agreed with the model, thus it is a measure of reliance. This should not be confused with the level of agreement between a subject and the ML model, which examines how often a subject's *initial* predictions agree with the model on *Phase 1* tasks.

[4] Additional non-parametric tests and regression analyses show consistent results. See details in SM Section 3.

[5] On a 7-point Likert scale, we considered subjects who provided a rating of 1–3 as disagreeing with the statement, and subjects who provided a rating of 5–7 as agreeing with the statement. Due to a mistake in Experiments 1 and 2, for subject's evaluations on the competence, reliability, understandability, and faith statements, ratings of 6 and 7 were all recorded as 5. Thus, we were not able to conduct more fine-grained analyses on these data for Experiments 1 and 2.



(a) Agreement Fraction: All (b) Agreement Fraction: ML Correct (c) Agreement Fraction: ML Wrong

(d) Switch Fraction: All (e) Switch Fraction: ML Correct (f) Switch Fraction: ML Wrong

Figure 3: The average values of agreement fraction and switch fraction in Phase 2 across three treatments in Experiment 1, on all Phase 2 tasks (Fig. 3a, 3d), on the subset of Phase 2 tasks where the ML model was correct (Fig. 3b, 3e), and on the subset of Phase 2 tasks where the ML model was incorrect (Fig. 3c, 3f). Error bars represent the standard errors of the mean. Upward trends are observed in all plots, indicating H1 is supported.

was correct and when the model was wrong. Indeed, on Phase 2 tasks where the model made correct predictions, the differences in subject's switch fraction across the three treatments were found to be statistically significant ($F(2, 295) = 5.56, p = 0.005$), so do the differences in subject's agreement fraction on Phase 2 tasks where the model made wrong predictions ($F(2, 298) = 4.94, p = 0.008$).

Exploring Why High-Confidence Agreement Affects Reliance. To gain some insights into why agreement level between a subject and the model in Phase 1 significantly affects subject's reliance on the model, we proceed to examine how such agreement influenced subject's perceptions of the model on a variety of aspects that relate to reliance behaviors.

We found that, in fact, without observing a model's performance, subjects who experienced a higher level of agreement with the model on tasks that they had high confidence in tended to consider the model to be *more* accurate, *more* competent, *more* reliable, more understandable and they also have *more* faith and trust in the model. One-way ANOVA and proportion test results again showed that the differences in subject's perceptions on all these reliance-related aspects were consistently and significantly different across treatments ($p < 0.001$). This provides a solid explanation for why subjects in different treatments show different levels of reliance on the models in Phase 2, despite all models made exactly the same predictions on Phase 2.

Overall, the results we obtained from our Experiment 1 support **H1**. That is, we indeed found that people tend to rely on an ML model more if the model has a higher level of agreement with them on cases that they feel confident about. Our data suggest that people may have used the level of such high confidence agreement as a proxy to gauge the key properties of the ML model, such as the model's accuracy, competence and reliability, which then guide people to adjust their reliance on the model accordingly.

4 EXPERIMENT 2

Experiment 1 shows that the agreement level between people and the model on decision-making tasks that people are highly confident about their own decisions significantly affects people's reliance on the model when they receive no information about the ML model's performance at all. However, such agreement may not necessarily be an accurate approximation of the performance of the ML model. Naturally, one may wonder whether the level of high-confidence agreement between people and a model still affects people's reliance on the model *after* people obtaining some feedback on how well the model performs, such as the aggregate information of the model's accuracy on a set of decision-making tasks.

To answer this question, we conducted our second experiment: Subjects were again asked to complete the same sequence of 30 prediction tasks as those used in Experiment 1, but this time, between the two phases, subjects received feedback on the model's

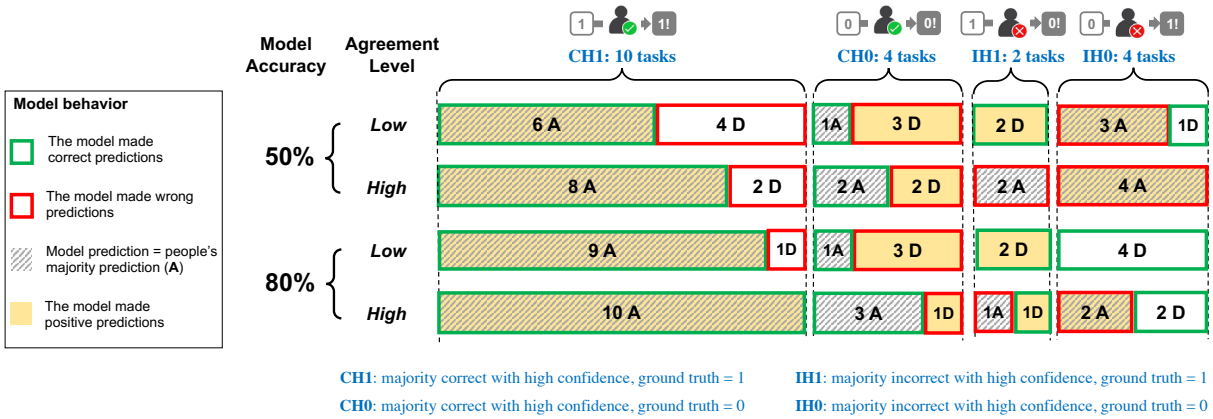


Figure 4: Predictions of the ML models on Phase 1 tasks in the four treatments of Experiment 2. Each row represents a unique model used in one treatment. "A" means the model agreed with the predictions made by the majority of people on the tasks, and "D" means the model disagreed with the majority predictions. Bars with stripes represent tasks on which the model's prediction agreed with people's majority prediction. Bars with the green (red) outline are tasks where the model made correct (wrong) predictions. Bars with the yellow background are tasks where the model made a positive prediction (i.e., predicted 1, the participant wants to see the date again).

overall accuracy in Phase 1. In addition to varying the level of human-model agreement on Phase 1 tasks, we also controlled the model's accuracy in Phase 1 to be either relatively low or high to understand whether and how the impact of high confidence human-model agreement is moderated by the level of model performance in influencing people's reliance on the model. We hypothesize that:

- [**H2**] The level of agreement between people and an ML model on tasks that people have high confidence in still significantly affects people's reliance on the model after the model's accuracy is observed in practice, regardless of the level of the observed accuracy; the higher the level of agreement, the more people rely on the model.

4.1 Experimental Design

Experimental Treatments. Subjects were randomly assigned to one of the four treatments in Experiment 2, which were arranged in a 2×2 factorial design differing along two dimensions:

- *the level of high confidence human-model agreement in Phase 1*: the model agreed with the majority of people on 50% of the 20 Phase 1 tasks (low agreement) or 80% of the 20 Phase 1 tasks (high agreement).
- *the model's overall accuracy in Phase 1*: either 50% (low accuracy) or 80% (high accuracy.)

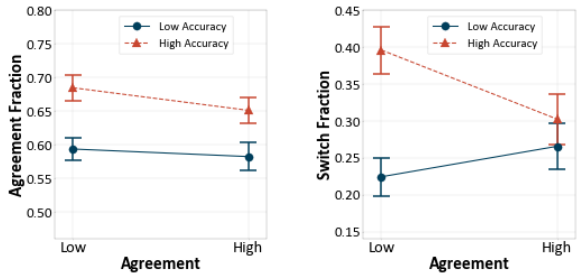
We associated each treatment with a unique ML model. To minimize the differences across treatments, we made sure that the four ML models had the *same* positive prediction rate of 70%. Figure 4 illustrates the details on the ML models' predictions on the 20 tasks of Phase 1 across the four treatments. Similar to that in Experiment 1, when the model's accuracy in Phase 1 was fixed, we varied the agreement level between people and the model in Phase 1 by starting from the high agreement treatment and then coupling the addition of disagreement on tasks where the majority is correct with the addition of disagreement on tasks where the majority is wrong. In addition, we were able to obtain treatments with the same level of agreement between people and the model but different levels of model accuracy by controlling how often the agreement occurs on tasks where the majority of people are correct. Finally, on each Phase 2 task, all ML models in different treatments made the same prediction.

Experimental Procedure. We again posted this experiment as a HIT on MTurk to U.S. workers only. It had an identical procedure as that of Experiment 1 except for the following differences: (1) Workers who had participated in Experiment 1 were not allowed to participate in this experiment; (2) After completing the 20 tasks in Phase 1, the subject was first shown a feedback screen with information about the ML model's overall accuracy on the tasks in Phase 1 (by design, it was either 50% or 80%) before she was asked to make an assessment on the four statements on the competence, reliability, understandability of the model as well as her faith in the model. In addition, the subject was *not* asked to guess the model's accuracy in Phase 1 this time since the model's accuracy was already revealed to the subject.

4.2 Experimental Results

In total, we collected valid data from 466 subjects who completed our second experiment and answered the attention check question correctly. We used the same metrics and followed the same methods as discussed in Section 3.3 to conduct our statistical analysis, except for that all one-way ANOVAs were replaced by two-way ANOVAs now given the two-factor design of this experiment. Again, as a validity check of our design, we confirmed that subjects in the high agreement treatments found the ML model's predictions agreed with their own predictions in Phase 1 tasks significantly more than subjects in the low agreement treatments ($p < 0.001$).

Does High-Confidence Agreement Still Affect Reliance After Some Performance Feedback is Obtained? Figures 5a and 5b compare subject's agreement fractions and switch fractions on all



(a) Agreement Fraction **(b) Switch Fraction**

Figure 5: The average values of agreement fraction and switch fraction in Phase 2 across four treatments in Experiment 2, on all Phase 2 tasks. Error bars represent the standard errors of the mean. Without observing a clear and consistent upward trend in either plot, H2 is not supported.

Phase 2 tasks across the 4 treatments, respectively. Using two-way ANOVA tests, we found that after observing the model's performance in practice, only the model's observed accuracy has a significant impact on people's reliance on the model in Phase 2 (agreement fraction: $F(1, 462) = 17.69, p < 0.001$, Cohen's $d = 0.39$; switch fraction: $F(1, 462) = 12.28, p < 0.001$, Cohen's $d = 0.32$), but *not* the level of agreement between a subject and the model in Phase 1. We also detected a significant interaction between the level of high confidence human-model agreement and the model's accuracy on influencing subject's switch fraction in Phase 2 ($F(1, 462) = 4.80, p = 0.029$). Looking into this interaction in more depth, we further found a surprising trend—while a higher level of human-model agreement in Phase 1 seems to have little effect on subject's reliance on the model after the model's low performance is observed, it actually leads to a *decrease* in subject's reliance on the model in Phase 2 after they observe the model's accuracy is relatively high, though post-hoc pairwise comparison suggests the decrease is not significant ($p = 0.166$). Similar patterns have been consistently observed when we looked into subject's reliance on the ML model on Phase 2 tasks where the model was correct and on Phase 2 tasks where the model was incorrect separately (see SM Section 4.1 for additional analyses).

Exploring Why High-Confidence Agreement No Longer Affects Reliance. We now turn to see how people's reliance on the model might have been influenced by their perceptions of the model (i.e., model competence, reliability, understandability, faith and trust in the model) after they obtain some feedback about the model performance. Detailed results are included in SM Section 4.2.

Overall, we found that after subjects observed the model's accuracy in Phase 1 was 50%, the fraction of subjects who agreed/disagreed that the model was competent, reliable, understandable, and they have faith in the model was *not* affected by the level of agreement that subjects had with the model in Phase 1. Further, subject's self-reported overall trust in the model was not affected by the level of human-model agreement in Phase 1 either. In fact, after a low performance of the ML model is observed, people tend to believe the model's competence, reliability, and understandability are at relatively low levels, and their faith and trust in the model are also low, which may all have led to their limited reliance on the model regardless of how often the model agrees with them previously on cases that they feel confident about.

On the other hand, after subjects observed the model's accuracy in Phase 1 was 80%, subjects still perceived the model that has a higher level of agreement with themselves in Phase 1 to be more competent, more reliable, and more understandable, they had a marginally increased trust in the model, though they did not show a higher level of faith in the model. Interestingly, such perceptions are not reflected in people's reliance behavior—in fact, we even find higher levels of high confidence human-model agreement results in a non-significant, but seemly consistent, decrease in people's reliance on the model when the model's performance is high. A possible explanation for this phenomenon is that people's reliance behavior may be not only affected by people's perceptions of the model but also be influenced by people's belief about *their own capability* on the decision-making tasks. In our experiment, subjects only obtained the *aggregate* information of the model's overall accuracy in Phase 1, but not the fine-grained performance feedback on the level of individual tasks, making it impossible for them to reason about their own performance in Phase 1 precisely[6]. Thus, when a model has a low level of agreement with a subject but still turns out to be highly accurate, the subject may suspect that she was *not* competent at the prediction tasks herself. As a result, the subject might have increased her reliance on the model as compared to the case when the model has a high level of agreement with her.

In sum, our results in Experiment 2 do not support **H2**. We did not find sufficient evidence that the level of agreement between people and an ML model on tasks that people have high confidence in still influences people's reliance on the model after people have obtained some aggregate performance information of the model.

5 EXPERIMENT 3

In the previous two experiments, we examine the effects of the human-model agreement on people's reliance on the model by varying the level of agreement between people and an ML model on tasks that people are *highly confident* about their own predictions. As a result, in these experiments, both agreement and disagreement between subjects and the ML model always occurred on high confidence tasks. However, when people interact with an ML model in the real world, they may be highly confident about their own predictions on some tasks, while not confident on the others, thus they may agree or disagree with an ML model both on high confidence tasks and low confidence tasks. An interesting question, then, is when people receive no information about an ML model's performance, whether and how people's confidence in the tasks that they agree or disagree with the model changes their reliance on the model. For example, when people agree with the model on tasks that they are not confident about their own predictions, do they have a similar level of reliance on the model as when they agree with the model on tasks that they are highly confident? Similarly, does disagreement with a model on low confidence tasks make people rely on the model to a similar degree as when they disagree with the model on high confidence tasks?

[6]It is possible in real life that only aggregate-level information of model performance is available due to various reasons like privacy concerns.

To answer these questions, we conducted a third experiment in which we controlled the level of agreement between people and an ML model to be *similar* across treatments, while varying whether the agreement/disagreement occurs on tasks that people have high confidence or not. We hypothesize that:

- [**H3**]: When people receive no information about an ML model's performance, people's confidence in their own predictions on tasks that they *agree* with the model significantly affects people's reliance on the model; the higher the confidence, the more people rely on the model.
- [**H4**]: When people receive no information about an ML model's performance, people's confidence in their own predictions on tasks that they *disagree* with the model significantly affects people's reliance on the model; the higher the confidence, the less people rely on the model.

5.1 Experimental Design

Experimental Treatments. Same as before, each subject was still asked to complete a sequence of prediction tasks divided into two phases, where Phase 1 was used for creating the treatment and Phase 2 was designed for measuring reliance.

We had two goals in mind when designing experimental treatments for Experiment 3: First, we hope to maintain *a similar level of agreement* (and therefore disagreement) between people and the ML model in Phase 1 for subjects in different treatments—without such control, it would be difficult to tell whether differences in people's reliance on the model across different treatments are due to differences in people's confidence in the tasks where agreement or disagreement occurs, or due to the varying levels of agreement between people and the model. Second, we aim to understand the impact of people's confidence in their own predictions during an *agreement* with the model on their reliance on the model, *separately* from that impact during a *disagreement*. This implies that we can *not* design our experimental treatments by having subjects in different treatments work on the same sequence of Phase 1 tasks, which are composed of both low confidence and high confidence tasks—in that case, since both the task instances in Phase 1 and the designed agreement level between people and the model need to be fixed across treatments, we would have to couple the decrease of human-model agreement on high confidence tasks with the increase of agreement on low confidence tasks. As a result, we would not be able to tell whether changes in people's reliance on the model, if any, are resulted from the increase of high confidence disagreement or the increase of low confidence agreement.

With these goals in mind, we used *different* tasks in Phase 1 for different treatments. We created a set of 4 experimental treatments arranged in a 2×2 design along two factors:

- *human's confidence on agreement with the model in Phase 1*: We included 6 tasks in Phase 1 on which the ML model would make the same predictions as the majority of people, and we varied whether on these tasks, people are highly confident about their own predictions (i.e., ***h**igh confidence **a**greement, "HA"*) or not confident about their own predictions (i.e., ***l**ow confidence **a**greement, "LA"*). Practically, we selected 6 high confidence tasks and 6 low confidence tasks from the pilot study. In high (low) confidence agreement treatments, the 6 high (low) confidence tasks were included in Phase 1, and the model's predictions on these tasks always *agreed* with the prediction given by the majority of people.
- *human's confidence on disagreement with the model in Phase 1*: We included another 4 tasks in Phase 1 on which the ML model would make different predictions than the majority of people, and we varied whether on these tasks, people are highly confident about their own predictions (i.e., ***h**igh confidence **d**isagreement, "HD"*) or not confident about their own predictions (i.e., ***l**ow confidence **d**isagreement, "LD"*). Similar as before, we selected another set of 4 high confidence tasks and 4 low confidence tasks from the pilot study. In high (low) confidence disagreement treatments, the 4 high (low) confidence tasks were included in Phase 1, and the model's predictions on these tasks always *disagreed* with the prediction given by the majority of people.

Again, we controlled the accuracy and positive prediction rate for ML models in different treatments to be the same to minimize the differences across treatments. After adding this constraint, we found that the feasible combination of tasks we could select for Phase 1 turned out to only include task instances where the majority of people would make a positive prediction according to our pilot study. To minimize subject's tendency to make some unnecessary negative predictions in the experiment—which may cause varying levels of agreement between subjects and the model across treatments—we further added the same set of two task instances to all four treatments, and on these two tasks, the majority of people would correctly make negative predictions. Together with the previous 10 tasks, subjects in each treatment would see a total of 12 tasks in Phase 1. Figure 6 shows the details on the task instances we selected for Phase 1 in each of the 4 treatments, as well as the ML models' predictions on them. We expect following such design, subjects in different treatments have roughly similar levels of agreement with the ML model in Phase 1, but their confidence in the tasks where they agree or disagree with the model differs.

Finally, we included in Phase 2 the same 10 low confidence tasks as those used in Experiments 1 and 2, and the ML models in different treatments made the same prediction on each of these tasks.

Experimental Procedure. This experiment had an identical procedure as that of Experiment 1 except for the following changes: (1) Workers who had participated in the previous two experiments were not allowed to participate in this experiment; (2) After introducing the task interface to subjects, we gave subjects a preview of possible prediction tasks that they might work on by showing each subject the same set of 4 profiles of speed dating events which represented task instances that people would likely be confident or not confident about their own predictions; this was intended to give subjects a sense of the range of confidence they might have in their own predictions on different tasks. Subjects were asked to review these profiles for at least 30 seconds before proceeding on to answer the qualification questions; (3) After completing Phase 1, in addition to report her perceptions of the model (e.g., guess model's accuracy, evaluate statement on model's competence, etc.), the subject was also asked to make a guess about the accuracy of her *own* independent predictions in Phase 1 (i.e., the predictions that she made in each task before seeing the model's predictions); (4) The base payment of the HIT was $1.2.

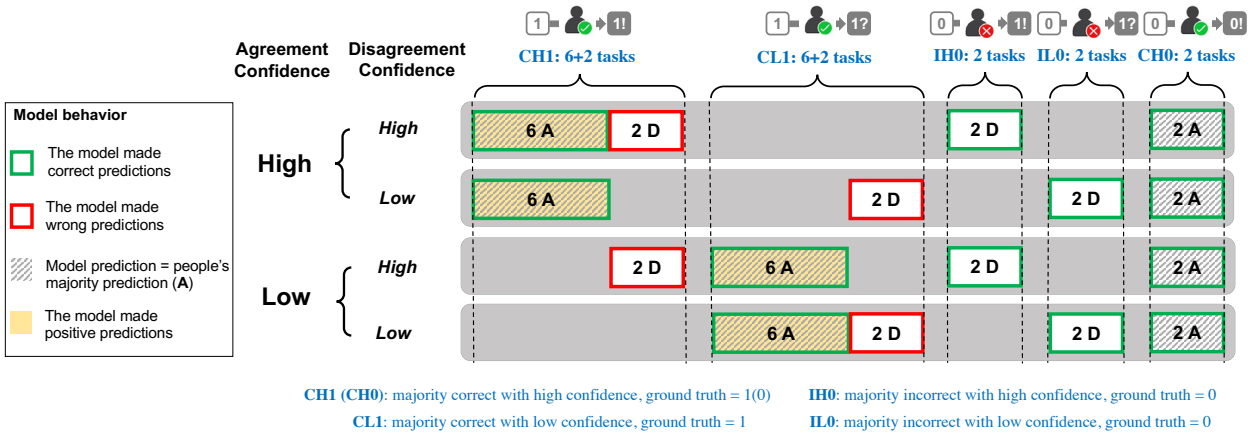


Figure 6: Predictions of the 4 ML models on Phase 1 tasks of Experiment 3. Each row represents the predictions made by one ML model (each model was used in an experimental treatment): "A" means the model agreed with the predictions made by the majority of people on the tasks, and "D" means the model disagreed with the majority prediction. Bars with stripes represent tasks on which the model's prediction agreed with people's majority prediction. Bars with the green (red) outline are tasks where the model made correct (wrong) predictions. Bars with the yellow background are tasks where the model made a positive prediction (i.e., predicted 1, the participant wants to see the date again). On the 12 tasks in Phase 1, the four models have the same accuracy of 83.3%, the same positive prediction rate of 50%, and the same level of agreement with people's majority predictions (i.e., agree on 66.7% of the tasks), though agreement/disagreement occurs on tasks where people have different levels of confidence in their own predictions.

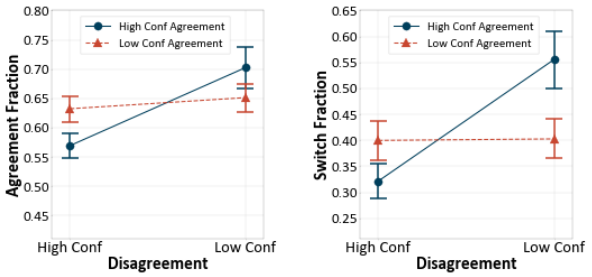
5.2 Experimental Results

We collected valid data from 402 subjects who completed this experiment and answered the attention check question correctly.

Despite our best effort, we found that there was still a small but significant difference in the actual level of agreement between subjects and the model in Phase 1 across the 4 experimental treatments. The average number of Phase 1 tasks on which a subject's initial predictions were the same as those of the model's was 7.46, 8.27, 6.61, and 7.33 for the HA–HD, HA–LD, LA–HD, and LA–LD treatments, respectively, and one-way ANOVA test indicates that the difference is statistically significant ($p < 0.001$). To enable a valid comparison across our experimental treatments, we then searched for a *range* such that when limiting our analyses only to those subjects whose actual agreement with the model in Phase 1 (i.e., the number of Phase 1 tasks that their initial predictions were the same as those of the model's) fall into this range, the actual level of human-model agreement in Phase 1 across the 4 treatments was statistically the *same*. The maximum such range that we could find was $[0, 8]$. That is, when focusing on the subset of subjects who agreed with the ML model on at most 8 tasks in Phase 1, we did not find a statistically significant difference in the actual level of agreement between subjects and the model across the 4 experimental treatments ($p > 0.05$). In total, 308 subjects (76% of all subjects) belong to this subset. Therefore, in the following, we restrict our analyses to this subset of subjects. We confirmed that, within this subset, subjects' confidence in the tasks that they agreed or disagreed with the model across different treatments aligned well with our expectations (see SM Section 5.1 for more details).

Exploring How Confidence on Agreement/Disagreement Affects Reliance. We start by examining how subject's confidence in their own predictions when they agree or disagree with the model affects their reliance on the model. Figures 7a and 7b compare the average values of agreement fraction and switch fraction on all Phase 2 tasks for subjects across the 4 treatments, respectively. A first glance at these figures suggests that there seem to be *interaction* effects between the two factors—subject's confidence on agreement and subject's confidence on disagreement—in influencing subject's reliance on the model. That is, whether high confidence human-model agreement increases people's reliance on the model, as compared to low confidence agreement, seems to be dependent on whether people disagree with the model on tasks that they are confident about their own predictions or not. Indeed, two-way ANOVA test results confirm that the interactions between the two factors are statistically significant for both two measures of reliance (agreement fraction: $F(1, 304) = 4.93, p = 0.027$; switch fraction: $F(1, 304) = 7.66, p = 0.006$).

More specifically, in Figure 7, we observed a trend that when having low confidence disagreement with the model, subject's high confidence in their agreement with the model leads to higher reliance on the model (i.e., subjects in the HA–LD treatment relied on the model more than those in the LA–LD treatment), but this trend was reversed when subjects had high confidence disagreement with the model. Post-hoc Tukey HSD tests, however, suggest that subject's confidence in their agreement with the model only had a marginal impact on switch fraction when subjects had low confidence disagreement with the model ($p = 0.094$). Meanwhile, with respect to how people's confidence in their *disagreement* with a model affects their reliance on the model, we found that subjects



(a) Agreement Fraction **(b) Switch Fraction**

Figure 7: The average values of agreement fraction and switch fraction in Phase 2 across four treatments in Experiment 3, on all Phase 2 tasks. Error bars represent the standard errors of the mean. Without observing blue lines are consistently above red lines, H3 is not supported. Without observing consistent upward trends, H4 is not supported.

in the HA–HD treatment relied on the model significantly less than subjects in the HA–LD treatment (agreement fraction: $p = 0.005$, Cohen's $d = 0.28$; switch fraction: $p = 0.002$, Cohen's $d = 0.27$), but there was no significant difference in subject's reliance on the model between the LA–HD and LA–LD treatments.

Again, when we looked into subject's reliance on the model within Phase 2 tasks where the model made correct/wrong predictions separately, we observed the same interaction patterns. For more details, see SM Section 5.2.

Understanding the Interaction. Next, we turn our attention to data on the subject's perceptions of the model and themselves in different treatments to explore why people's confidence in their agreement or disagreement with an ML model affects their reliance on the model in the way that we have observed above.

To begin with, we plot the average values of the subject's perceived model accuracy in Phase 1 across the 4 treatments in Figure 8a. To give a more complete picture, we also plot subject's perceived accuracy of their own in Phase 1 in Figure 8b, and Figure 8c presents the comparison on subject's perceived difference between the model's accuracy and their own accuracy across treatments. Interestingly, it appears that compared to subjects in the HA treatments, subjects in the LA treatments tended to believe the model to be less accurate, but they also perceived themselves as less accurate on the prediction tasks. In addition, compared to subjects in the HD treatments, subjects in the LD treatments seemed to believe the model to be more accurate while perceiving themselves as less accurate. Therefore, when we put both perceptions together and examine subjects' perceived accuracy difference between the model and themselves in Figure 8c, we found a similar cross-over interaction pattern as that in subjects' reliance on the model.

Moreover, we also found the similar pattern that people's confidence in their agreement and disagreement with the model interact with each other in influencing people's perceptions of the model's competence, the model's understandability, as well as people's overall trust in the model (see SM Section 5.3 for detailed figures and analyses). Notably, though, the cross-over interaction pattern was not observed in subject's perception in the reliability of a model or their faith in a model. To the contrary, we found that compared to low confidence disagreement, high confidence disagreement consistently results in a significantly lower perceived level of model reliability ($p < 0.001$), while low confidence human-model agreement leads to higher faith in the model compared to high confidence human-model agreement ($p = 0.042$).

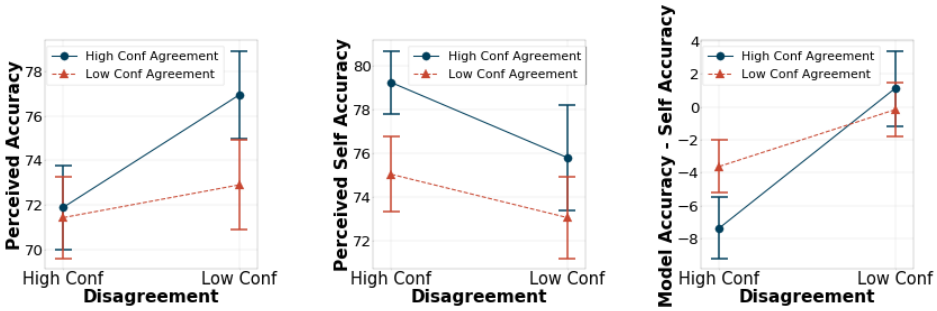
As an attempt to connect our observations on how people's confidence in their agreement and disagreement with the model affects their perceptions and reliance on the model together, we conjecture that human-model agreement/disagreement with varying levels of human confidence may influence people's reliance on an ML model from different perspectives. For example, people might use the number of high confidence disagreement they have experienced with a model to determine whether the model makes any "obvious mistakes" (as people may tend to believe they are correct on tasks that they feel confident about), which largely dominates their perceptions of the reliability of the model. For those models that do not make obvious mistakes, a larger number of high confidence human-model agreement possibly serve as a piece of evidence for people to believe the model has the capability to provide correct recommendations, again based on the assumption that people are themselves correct on tasks that they have high confidence in. Moreover, although low confidence human-model agreement may provide limited clues for people to gauge the correctness of the model as people are not sure whether their own predictions are correct or not, it may make people feel the model shares a similar reasoning process as themselves, thus increase their faith in the model based on such similarity. People's final level of reliance on the model, then, is jointly determined by multiple factors including these reliance-related perceptions, and future studies are needed to rigorously study the relationships between them.

To summarize, our results in Experiment 3 do not support **H3** or **H4**. Rather, the experimental data we obtained indicate that when people decide how much to rely on an ML model without knowing any information about the model's performance, they take an *integrated* view to examine their confidence in both their agreement with the model and their disagreement with the model together. As a result, we observed that how people's confidence in their agreement (disagreement) with an ML model affects their reliance on the model is dependent on their confidence in those cases where they disagree (agree) with that model.

6 DISCUSSION

Via three experiments, we examined the possible heuristics that people may use to adjust their reliance on an ML model when performance feedback about the model is limited. Table 1 summarizes the design and findings of the three experiments. In this section, we provide discussions on the potential benefits and risks brought by people's usage of these heuristics, and steps that can be taken to reduce the risks. We also provide design implications for enhancing human-AI collaboration and discuss the limitations of our work.

Benefits and risks of the heuristics. We find that without any information about an ML model's performance, people seem to use the level of high confidence agreement between the model and themselves to decide how much to rely on the model. Such heuristics can bring about both benefits and risks. On the one hand, if



(a) Model Accuracy (b) Self Accuracy (c) Accuracy Difference

Figure 8: Average values of subjects' perception of model accuracy, their own accuracy, and the difference between the accuracy of the model's and their own, in Phase 1 of Experiment 3 (constrained to subjects who agreed with the ML model's predictions on at most 8 tasks in Phase 1). Error bars represent the standard errors of the mean.

	Experiment 1	**Experiment 2**	**Experiment 3**
Hypothesis to test	**H1**: High-confidence human-model agreement affects reliance when performance is unknown.	**H2**: High-confidence human-model agreement affects reliance after performance feedback is obtained.	**H3/H4**: Human's confidence on their agreement/disagreement with the model affects reliance when performance is unknown.
Independent variables	*Human-model agreement level in Phase 1*: low, medium, high	*Agreement level in Phase 1*: low, high; *Model accuracy in Phase 1*: 50%, 80%	*Confidence on agreement in Phase 1*: low, high; *Confidence on disagreement in Phase 1*: low, high
Phase 1	20 high-confidence tasks same tasks, but different ML predictions across treatments	20 high-confidence tasks same tasks, but different ML predictions across treatments	12 high-confidence and low-confidence tasks different tasks for different treatments
Phase 2	10 low-confidence tasks same tasks and same ML predictions for all treatments	10 low-confidence tasks same tasks and same ML predictions for all treatments	10 low-confidence tasks same tasks and same ML predictions for all treatments
Phase 1 accuracy revealed?	No	Yes	No
Findings	**H1** supported	**H2** not supported Model accuracy affects reliance	**H3**, **H4** not supported Interactions between the two factors detected

Table 1: Summary of the design and findings of the three experiments.

people are highly calibrated in their own predictions—for example, if people are almost always correct when they are confident—and the ML model does not have substantial performance disparity on tasks with different properties (e.g., tasks that humans are confident vs. not confident), such heuristic indeed allows people to estimate the performance of an ML model accurately and thus establish an appropriate level of reliance on the model, even without access to any objective information about the model performance. On the other hand, there also exist conditions where such heuristic fails to help people rely on an ML model appropriately. For example, people may overestimate their own accuracy on tasks that they are confident due to factors like their inaccurate interpretation of the decision-making task and their cognitive biases (e.g., the Dunning–Kruger effect [29]). As a result, by comparing the model's predictions to their own on tasks that they have high confidence in, people may systematically underestimate an ML model's accuracy, which potentially leads to underreliance. Furthermore, when people's predictions are *not independent* of those of the model's and they still use the level of high confidence agreement between the model and themselves to calibrate their reliance on the model, people may overly rely on models that share the same biases as themselves while showing a degree of underreliance on models that are complementary to themselves.

The usage of such heuristics may raise additional risks if the ML model's performance has systematic discrepancy on different data, which in fact is not rare in real-world applications of ML models [6, 42, 59]. In such scenario, by adjusting the reliance on an ML model simply based on the level of human-model agreement on high confidence tasks, people may inappropriately over-generalize their perceived model performance on tasks with certain characteristics to other tasks with different characteristics. In fact, this risk has been partly reflected in our experiment as we observed that once seeing a high level of agreement between the model and themselves on tasks that they are confident, people tend to rely on the ML model more no matter whether its prediction is correct or not.

Interestingly, we find that once observing an ML model's performance, people tend to use this observed performance rather than their perceived level of high confidence agreement with the model to decide their reliance on the model. This behavior also poses both benefits and risks. On the positive side, when the observed model performance is an accurate reflection of the model's overall performance in the real world, it enables people to correct their over-reliance on models that share the same biases as themselves while overcoming their under-reliance on models that are complementary to themselves. However, the model's observed performance on a small set of decision-making tasks does not necessarily accurately

reflect the model's average performance in practice. Thus, simply adjusting one's reliance based on the observed model performance on a limited number of real-world trials makes people overlook useful cues about model performance that is carried in their agreement and disagreement with the model, and may again result in inappropriate reliance on the model.

Finally, our observations on how people utilize their own confidence in cases where they agree and disagree with an ML model to adjust their reliance on the model, again, reveal some potential risks. Notably, when people and an ML model mostly agree with each other on cases that people lack confidence, people's confidence in their disagreement with the model no longer influences their reliance on the model. This indicates the possibility that for people who feel they lack expertise in a decision-making task, they might not be sufficiently alarmed by the high confidence disagreement between the ML model and themselves, and may therefore show inappropriate over-reliance on the model.

The need of enabling people to understand their own decision-making performance. Using high confidence human-model agreement to estimate the performance of an ML model would not be reliable if people are not calibrated in their own decisions (i.e., if people's confidence in their decisions does not accurately reflect their accuracy). A critical step towards helping people to make reliable estimation, thus, is to facilitate people's understanding of their own decision-making performance. This includes raising people's awareness of biases in their own decision-making [2, 24], designing methods to allow people assess their own independent decision-making performance (e.g., ask people to make decisions on historical data and provide performance feedback), and encouraging people to serve as their own devil's advocate by challenging themselves to probe why they make certain decisions and how they might possibly be wrong [11]. Only if people have a solid understanding of when their own decision-making accuracy is high and when not, obtaining an accurate estimate of the ML model's performance without performance feedback would become possible.

The need of helping people better understand ML model's performance. Our experimental results suggest that people may have a tendency to over-generalize the performance of an ML model, which is either estimated or observed by themselves, between tasks with different characteristics. To overcome this problem, efforts should be made to improve people's capability of interpreting the ML model's performance. For example, tutorials can be provided to end-users of ML models to increase their general knowledge of ML, including that ML models are often learned from data, and depending on the quality of the data, ML models may exhibit different performance on different data [34]. Designers of ML models should transparently report properties of a model to people, including its intended uses and potential limitations [40]. The uncertainty inherent in the process of model performance estimation process should also be properly communicated to people, so that people can better understand how much they can generalize the model performance they have observed to other contexts [27].

Implications for enhancing human-AI collaborations. Our work also provides implications for designing better ML models to enhance the partnership between humans and machines. In the context of AI-assisted decision-making, the ultimate goal is to optimize the human-AI joint team decision-making performance [5, 26], and achieving this goal requires humans and AI to complement each other. However, our results in this work indicate when performance feedback of the AI is limited, people might be unwilling to rely on the complementary AI partner, due to the limited agreement between them. Our findings suggest one possible way to minimize such undesirable human reaction to AI is to maximize the agreement between humans and AI on those cases that humans are confident and *actually correct*, for example, by integrating expert knowledge into the development of ML models [21]. Model designers can also provide explanation of the model decision when high confidence human-model disagreement occurs, to help people better understand why the disagreement exists and mitigate the potential negative impact such disagreement brings about.

On the conceptual level, our findings echos that of other work in human-AI collaboration [1, 3, 4, 26] and reiterates that the design and development of ML-based decision aids should *not* be isolated from the people who will use them. Only by taking people's perceptions and reactions to the ML model into consideration can the decision aids release their full potential to improve the joint performance of the human-AI team. However, critically, we emphasize that the consideration of human behavior in the AI development process should *not* be used for nudging people into blindly relying on the AI, which would be unethical and even dangerous.

Limitations and future work. Our study was conducted with laypeople (i.e., subjects recruited from Amazon Mechanical Turk) on one specific type of prediction task. Cautions should be used when generalizing results in this work to different settings, such as how the agreement level between an ML model and an expert would affect the expert's reliance on the model on some tasks involving significant higher levels of stakes (e.g., ML assists doctors in medical decision-making). More experimental studies should be carried out in the future with different populations on different types of tasks to understand to what extent the results reported here can be generalized. In addition, people's perception of their agreement with a model may involve not only how frequently the model agrees with themselves on decisions for specific cases, but also how consistent the model's predicted orderings of different cases are compared to their own's, which may relate to the perceived internal consistency of the model. Separating the impact of the frequency of human-model agreement on reliance from the impact of human-model agreement on the relative ordering of different cases on reliance would be a challenging but exciting future work.

7 CONCLUSION

In this paper, we present our initial attempt to uncover the heuristics that people adopt to adjust their reliance on machine learning models in AI-assisted decision making, when objective performance feedback of the models is limited. Via three randomized human-subject experiments, we show that people tend to use their levels of agreement with a model on cases that they are highly confident as a proxy to estimate the model performance and adjust their reliance on the model accordingly, but such agreement shows limited impact in influencing reliance after people have observed the model's performance in practice. Moreover, holding the level of human-model

agreement constant, how people's confidence in their agreement with a model affects their reliance on the model depends on their confidence in those cases where they disagree with the model. We highlight people's usage of these heuristics may raise risks of inappropriate reliance on ML models, and we discuss both actions that can be taken to reduce these risks and possible directions to improve human-AI collaboration by taking human's heuristics into account.

8 ACKNOWLEDGEMENTS

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] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[3] 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.

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhu, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S Weld. 2020. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *arXiv preprint arXiv:2006.14779* (2020).

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

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

[8] Eric T Chancey, James P Bliss, Yusuke Yamani, and Holly AH Handley. 2017. Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human factors* 59, 3 (2017), 333–345.

[9] 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.

[10] Stephen R Dixon, Christopher D Wickens, and Jason S McCarley. 2007. On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human factors* 49, 4 (2007), 564–572.

[11] David Dunning. 2014. We are all confident idiots. *Pacific Standard* 7 (2014), 46–54.

[12] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718.

[13] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. 1999. Misuse and disuse of automated aids. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 43. SAGE Publications Sage CA: Los Angeles, CA, 339–339.

[14] Philipp Ecken and Richard Pibernik. 2016. Hit or miss: what leads experts to take advice for long-term judgments? *Management Science* 62, 7 (2016), 2002–2021.

[15] 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.

[16] Raymond Fisman, Sheena S Iyengar, Emir Kamenica, and Itamar Simonson. 2006. Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics* 121, 2 (2006), 673–697.

[17] SM Fleming and ND Daw. 2016. Self-evaluation of decision performance: A general Bayesian framework for metacognitive computation. *Psychol Rev* 124 (2016), 1–59.

[18] Jorge Galindo and Pablo Tamayo. 2000. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational Economics* 15, 1-2 (2000), 107–143.

[19] Ji Gao and John D Lee. 2006. Extending the decision field theory to model operators' reliance on automation in supervisory control situations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 36, 5 (2006), 943–959.

[20] Ji Gao, John D Lee, and Yi Zhang. 2006. A dynamic model of interaction between reliance on automation and cooperation in multi-operator multi-automation situations. *International Journal of Industrial Ergonomics* 36, 5 (2006), 511–526.

[21] Efstathios D Gennatas, Jerome H Friedman, Lyle H Ungar, Romain Pirracchio, Eric Eaton, Lara G Reichmann, Yannet Interian, José Marcio Luna, Charles B Simone, Andrew Auerbach, et al. 2020. Expert-augmented machine learning. *Proceedings of the National Academy of Sciences* 117, 9 (2020), 4571–4577.

[22] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[23] Dale W Griffin and Lee Ross. 1991. Subjective construal, social inference, and human misunderstanding. In *Advances in experimental social psychology*. Vol. 24. Elsevier, 319–359.

[24] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[25] Sarfaraz Hussein, Kunlin Cao, Qi Song, and Ulas Bagci. 2017. Risk Stratification of Lung Nodules Using 3D CNN-Based Multi-task Learning. In *Information Processing in Medical Imaging*, Marc Niethammer, Martin Styner, Stephen Aylward, Hongtu Zhu, Ipek Oguz, Pew-Thian Yap, and Dinggang Shen (Eds.). Springer International Publishing, Cham, 249–260.

[26] H Kaur, A Williams, and WS Lasecki. 2019. Building shared mental models between humans and ai for effective collaboration. (2019).

[27] Yea-Seul Kim, Paula Kayongo, Madeleine Grunde-McLaughlin, and Jessica Hullman. 2020. Bayesian-Assisted Inference from Visualized Data. *arXiv preprint arXiv:2008.00142* (2020).

[28] Gang Kou, Yi Peng, and Guoxun Wang. 2014. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences* 275 (2014), 1–12.

[29] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[30] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[31] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[32] Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. 1993. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks* 6, 6 (1993), 861–867.

[33] Varda Liberman, Julia A Minson, Christopher J Bryan, and Lee Ross. 2012. Naïve realism and capturing the "wisdom of dyads". *Journal of Experimental Social Psychology* 48, 2 (2012), 507–512.

[34] 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.

[35] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

[36] Gary Marks and Norman Miller. 1987. Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological bulletin* 102, 1 (1987), 72.

[37] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*. PMLR, 6765–6774.

[38] Stephanie M Merritt. 2011. Affective processes in human–automation interactions. *Human Factors* 53, 4 (2011), 356–370.

[39] Julia A Minson, Varda Liberman, and Lee Ross. 2011. Two to tango: Effects of collaboration and disagreement on dyadic judgment. *Personality and Social Psychology Bulletin* 37, 10 (2011), 1325–1338.

[40] 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.

[41] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105.

[42] Besmira Nushi, Ece Kamar, and E. Horvitz. 2018. Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. In *HCOMP*.

[43] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[44] Raja Parasuraman and Christopher D Wickens. 2008. Humans: Still vital after all these years of automation. *Human factors* 50, 3 (2008), 511–520.

[45] Niccolo Pescetelli, Geraint Rees, and Bahador Bahrami. 2016. The perceptual and social components of metacognition. *Journal of Experimental Psychology: General* 145, 8 (2016), 949.

[46] Niccolo Pescetelli and Nick Yeung. 2018. The role of decision confidence in advice-taking and trust formation. *arXiv preprint arXiv:1809.10453* (2018).

[47] 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).

[48] 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.

[49] Victor Riley. 1996. Operator reliance on automation: Theory and data. *Automation and human performance: Theory and applications* (1996), 19–35.

[50] Lee Ross, David Greene, and Pamela House. 1977. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology* 13, 3 (1977), 279–301.

[51] Julian Sanchez, Arthur D Fisk, and Wendy A Rogers. 2004. Reliability and age-related effects on trust and reliance of a decision support aid. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48. Sage Publications Sage CA: Los Angeles, CA, 586–589.

[52] 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.

[53] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.

[54] Kelly E See, Elizabeth W Morrison, Naomi B Rothman, and Jack B Soll. 2011. The detrimental effects of power on confidence, advice taking, and accuracy. *Organizational behavior and human decision processes* 116, 2 (2011), 272–285.

[55] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.

[56] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. In *Advances in neural information processing systems*. 2643–2651.

[57] Kees Van Dongen and Peter-Paul Van Maanen. 2013. A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies* 71, 4 (2013), 410–424.

[58] Lyn M Van Swol and Janet A Sniezek. 2005. Factors affecting the acceptance of expert advice. *British journal of social psychology* 44, 3 (2005), 443–461.

[59] 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.

[60] Xinxi Wang and Ye Wang. 2014. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*. 627–636.

[61] Andrew Ward, L Ross, E Reed, E Turiel, and T Brown. 1997. Naive realism in everyday life: Implications for social conflict and misunderstanding. *Values and knowledge* (1997), 103–135.

[62] 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.

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

[64] 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.

[65] 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. *arXiv preprint arXiv:2001.02114* (2020).

[66] 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.