When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models

Amy Rechkemmer
Purdue University
West Lafayette, Indiana, USA
arechke@purdue.edu

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

ABSTRACT

Previous research shows that laypeople's trust in a machine learning model can be affected by both performance measurements of the model on the aggregate level and performance estimates on individual predictions. However, it is unclear how people would trust the model when multiple performance indicators are presented at the same time. We conduct an exploratory human-subject experiment to answer this question. We find that while the level of model confidence significantly affects people's belief in model accuracy, both the model's stated and observed accuracy generally have a larger impact on people's willingness to follow the model's predictions as well as their self-reported levels of trust in the model, especially after observing the model's performance in practice. We hope the empirical evidence reported in this work could open doors to further studies to advance understanding of how people perceive, process, and react to performance-related information of machine learning.

CCS CONCEPTS

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

KEYWORDS

Machine learning, confidence, accuracy, trust, human-subject experiments

ACM Reference Format:
Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3491102.3501967

1 INTRODUCTION

Today, numerous innovative machine learning (ML) models have been rapidly developed and applied to a wide range of application scenarios to assist people, from decision-making in everyday life to problem-solving in critical societal challenges. For example, neural networks have been used to forecast traffic speed in large-scale transportation networks and recommend optimal routes [16, 40]. Researchers have trained ML models to predict poverty in developing countries and help local governments better allocate their scarce resources [27]. ML has also shown potential in accurately predicting the household re-entry to homeless system, which can further inform the design of more effective intervention strategies [23, 32].

With the rapid growth of user-facing systems that are built on top of ML models, a growing line of research on understanding whether and how do end-users trust these models has recently emerged [52, 57, 58]. Many different factors have been identified as influencing people's trust in ML, such as people's understanding of how the model works [19, 37] and people's perceptions on whether the model is biased [9, 62]. Perhaps more intuitively, people's trust in an ML model is also highly dependent on how well the model can perform. For example, it was found that people's trust in an ML model is significantly affected by the model's performance, as measured by both the model's stated accuracy on some held-out data and its observed accuracy in practice [35, 58].

While a performance metric like accuracy may provide useful summary information for people to evaluate the *overall* reliability of an ML model across a set of predictions, it contains little insight into how likely each *individual* prediction that the model makes is correct. On the other hand, an ML model can often quantify its uncertainty on each individual prediction using a *confidence* score, which represents the model's accuracy estimate on that particular prediction [25, 43, 46, 60]. More recently, researchers have found that such model confidence also influences how much people would be willing to trust the model [61]—people trust the ML model more in cases when the model has higher confidence.

Our knowledge of how different kinds of performance indicators of an ML model *alone* affects people's trust in the model continues to grow. However, there remains a key, but currently under-explored, aspect in further advancing our understanding of people's trust in ML—that is, how would people trust an ML model in the presence of *multiple* performance indicators of it. For example, when information on both the model's accuracy measurements (on some held-out data and/or on a number of real-world trials) and the model's confidence estimates on individual predictions are provided, which one(s) would people choose to rely upon in deciding how much to trust the model? And what's the role of one performance indicator—say model confidence—in moderating or even changing the effects of other performance indicators (e.g., model accuracy) on people's trust in the model?

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501967

Specifically, we are interested in finding out the answers to these questions for those users of ML models who are *laypeople*. This is because laypeople's interactions with ML models are becoming highly prevalent today, especially in various *low-stakes* decision making settings such as optimal route planning [16, 40] and entertainment selection [6]. This implies that, due to the legal requirement of transparency and design guidelines of best practices [1], laypeople are increasingly exposed to a variety of performance information of an ML model during their usage of the model. However, our empirical understanding of how laypeople would interpret and act upon this rich set of performance information of the ML model and whether their reactions to this information leads to appropriate trust in the model is still largely lacking. In fact, prior research has indicated that laypeople may experience difficulty in making use of numerical information [44, 48], and they may follow an incorrect ML model recommendation even when performance information about the recommendation suggesting it as less trustworthy is given [51]. Thus, such understanding is critical for us to analyze how effective the current ways of ML model performance communication are for a target user population of laypeople, and it can also inform us of how to make such communication truly useful to laypeople.

Therefore, in this paper, we conduct an exploratory study to experimentally examine how multiple performance indicators of an ML model, together, affect laypeople's trust in the model in a low-stakes decision-making task. Specifically, we ask:

- When people receive information on both the model's stated accuracy on held-out data and the model's confidence on individual predictions, but have not observed the model's accuracy in practice yet, does the stated accuracy (alternatively, the model confidence) affect people's trust in the model?
- Does the answer for the question above change after people have observed the model's accuracy in practice?
- Does a model's observed accuracy in practice affect people's trust in the model, in the presence of model confidence?
- How do model confidence and model accuracy interact with each other to influence people's trust in the model?

Conjecturing the answer to any of these questions turns out to be quite challenging. On the one hand, compared to a model's stated accuracy, model confidence is provided on the level of individual prediction and appears more directly relevant for people to evaluate whether each of the model's predictions is trustworthy. After all, the model's stated accuracy is obtained on held-out data which may be fundamentally different from the current use cases at hand. Following this line of thinking, one may expect to see a significant impact of model confidence on people's trust in the model, and perhaps the effect of the model's stated accuracy on trust is minimal, if any. On the other hand, a model's confidence on a prediction is only the model's "estimate" on how likely the prediction would be correct, and it has been shown in many studies in the machine learning community that the raw confidence scores produced by an ML model can be poorly calibrated [25, 33, 42]. That is, the confidence estimate an ML model associates to a prediction does *not* necessarily reflect the true correctness likelihood of that prediction. In this case, it is reasonable to hypothesize that people's trust in an ML model would still be affected by the model's stated accuracy, but not so much by model confidence.

Even more complicated, after people get the chance to interact with the ML model and observe its accuracy in practice, people may make additional inference on how calibrated the model's confidence and how reliable the model's stated accuracy is through their limited interactions with the model. Thus, in deciding how much to trust the model after observing its performance in practice, people may need to consider both relevance and reliability for each of the three pieces of information that can be used to gauge the trustworthiness of the ML model—model confidence, stated accuracy, and observed accuracy. It is unclear how these three factors would affect trust, separately and collectively.

To answer these questions, we designed and conducted a randomized behavioral experiment in which we recruited human subjects from Amazon Mechancial Turk to complete a sequence of decision making tasks (i.e., predict speed dating outcome) with the help of an ML model. Our experiment consisted of a total of eight treatments, arranged in a 2×2×2 design, and ML models used in different treatments differed along three factors—the ML model's *confidence level*, *stated accuracy*, and *observed accuracy*. Due to the multidimensional nature of trust, we used a variety of measures to quantify subjects' trust in the ML model in our experiment, including subjective measures focusing on trust perceptions (e.g., subject's belief in model competence and self-reported trust level) and more objective measures characterizing trusting behavior (e.g., frequency for a subject to "follow" a model's prediction).

Our experimental results show that, overall, model confidence affects people's belief in model competence but has no reliable impact on their self-reported levels of trust in the model or their willingness to follow the model's predictions, both before and after they have observed the model's performance through real-world trials. In contrast, the model's accuracy, including both the stated accuracy and observed accuracy, consistently and significantly affects people's trust in the model in all dimensions. Comparing the magnitude of the effects of different performance indicators on various measures of trust, we found that model accuracy—especially the model's observed accuracy after it has been obtained from real-world trials—has a larger effect on all measures of trust except for people's belief in model competence. Further analyses also reveal that there exist some interactions between model confidence and model accuracy in influencing people's trust in an ML model.

Taken together, our results provide exploratory evidence that model confidence and model accuracy play different roles in influencing people's trust in an ML model. They also highlight behavior of people when they react to multiple performance indicators that can potentially be irrational, such as their over-reliance on a model's observed accuracy despite it having been obtained through a small number of real-world trials. We conclude by discussing the design implications and limitations of our work. In particular, proper cautions should be used when generalizing our results to other settings. More confirmatory studies should be conducted to examine to what extent our exploratory findings still hold for different populations, on different types of tasks, and using different ways to communicate model performance. We hope these findings will inspire more theoretical investigations into the mechanisms of how people make sense of and make use of various performance-related information of machine learning.

2 RELATED WORK

Research on understanding whether, when, and how do people trust the outputs of an ML system has received increased attention over the years. It was shown that, for example, people accept ML model recommendations over human suggestions or their own judgment in an unfamiliar domain [11, 38], but they quickly lose trust in an algorithm after seeing it err [17]. Such decrease in trust is reduced if people are given the control to adjust the algorithm predictions rather than simply accepting them as is [18]. However, loss of trust due to incorrect ML system decisions cannot be fully gained back by the same amount of correct decisions made by the system [59].

More recently, a growing number of experimental studies have been conducted to examine what the key influencing factors are in determining people's trust in an ML model and how. One such factor is the ML model's intelligibility [19], though different studies showed inconsistent results to this end. For instance, Lai and Tan [35] found that in the context of deception detection, showing explanations of an ML model significantly increases people's trust in the model. In contrast, the transparency of an ML model, in terms of whether the model is a clear or black-box model, was shown to not affect people's trust in the model [47], while providing explanations for a text emotion predicting system only affected trust before people experienced the system [50]. Moreover, it was found that the type of explanation presented to users impacted their ability to appropriately calibrate trust in a model [55]. Other than the model's intelligibility, researchers have also demonstrated that the level of agreement between humans and the ML model [39], a human's mental model of the ML model's error boundary [4], and the compatibility of an ML model update with humans' prior experience of the model [5], can all affect how much people will trust the model.

Another natural influencing factor of trust in an ML model is the model's performance. Indeed, Yin et al. [58] ran a sequence of human-subject experiments and found that both an ML model's stated accuracy on held-out data and its observed accuracy in practice significantly affect laypeople's trust in the model. Yet, even for models with the same level of accuracy, the ways that expectations on model performance are set prior to the use of them were also found to affect people's perceptions and acceptance of the model [31]. Beyond the accuracy measurements, a model's confidence on each individual prediction is also relevant in influencing end-user's trust. For example, a few previous studies have found that for context-aware systems and autonomous systems, displaying system confidence on the quality of its decision aids (e.g., memory aids or location predictions) would both improve user's trust in the system [2, 54] and increase user performance in the task [3, 15]. Lim and Dey [36], however, showed that displaying low confidence levels of a context-aware system allows users to realize the limited capabilities of the system and thus leads to decreased trust. Most recently, Zhang et al. [61] conducted an experimental study and showed that confidence scores of an ML model help people to calibrate their trust with confidence levels, such that they trust an ML model more when its confidence is high. Suresh et al. [51], however, found that a person's capability of calibrating trust in an ML model based on its confidence may vary with the person's characteristics, such as the person's math and logic skills.

Differing from previous research, we look into how people trust in an ML model when they have *multiple* types of performance information that they can leverage to infer the trustworthiness of the model. Earlier literature has explored how humans would utilize multiple, possibly probabilistic, pieces of information during their decision-making when machines are *not* in the loop. For example, in processing multiple pieces of performance information of stocks in a sequence, a recency effect was observed among investors showing that they heavily relied upon the most recent information in their final decisions [45]. Various mechanisms have also been studied on how decision-makers aggregate a probabilistic forecast of multiple experts with different levels of confidence and bias [8, 53, 56]. However, to the best of our knowledge, there are no experimental studies or theoretical models on how humans interpret multiple performance-related information of ML-based decision aids, especially when different pieces of information are of differing nature and provided at different granularity (i.e., local accuracy estimate vs. aggregate accuracy measurements).

3 EXPERIMENTAL DESIGN

To understand how laypeople's trust in an ML model for making low-stakes decisions is affected by both model confidence and model accuracy, we designed and conducted a randomized behavioral experiment with human subjects recruited from Amazon Mechanical Turk (MTurk).

3.1 Experimental Tasks: Predicting Speed Dating Outcome

Each subject in our experiment was asked to complete a sequence of 40 tasks on predicting the outcome of speed dating events with the help from a pre-trained ML model. Specifically, in each prediction task the subject was presented with a profile of one participant in a speed dating event along with some information about his or her date, including:

- *Basic demographics of the participant and the date*, e.g., gender, age, field of study, and race.
- *The participant's dating preferences*, which consisted of the participant's allocation of 100 points to six attributes (e.g., attractiveness, sincerity, intelligence) to show the relative importance of them in relation to one another when it comes to romantic attraction, as well as the level of importance for the participant to date someone of the same race.
- *The participant's rating of the date*, with respect to the six attributes using a scale from one to ten, and two additional self-reported scores on how happy the participant expected to be with the date and how much the participant liked the date, again in the range of one to ten.

Figure 1 shows an example of the task interface. Profiles that we showed to subjects (e.g., Figure 1A) were taken from participants of real-life speed dating events in an experimental study [22]. At the end of each speed dating event, the participant was asked to indicate if he/she would want to see the date again in the future, and this is what we asked the subject to predict. In particular, the subject followed a 4-step procedure in each task to make the prediction:

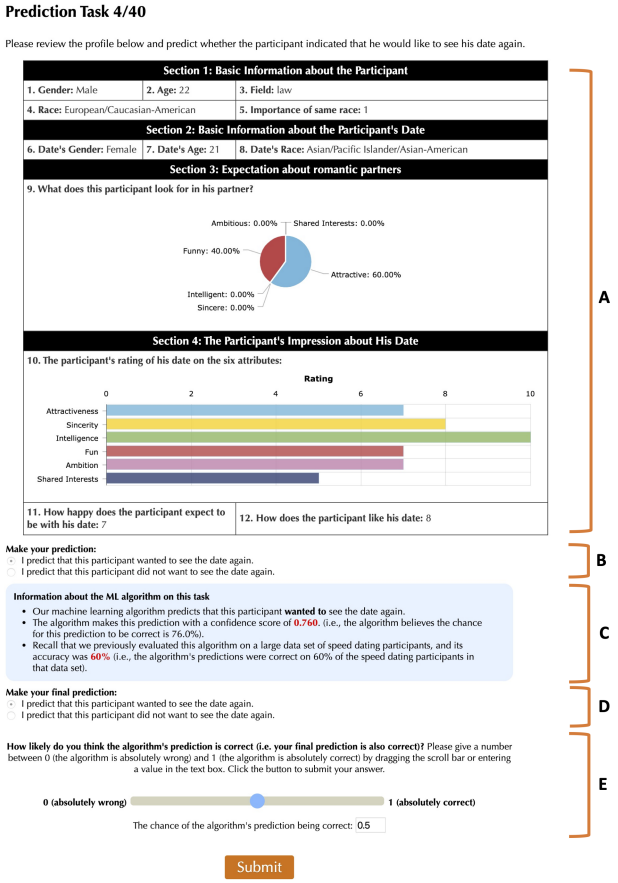


Figure 1: Interface of the experimental task in Phase 1. A: subjects are shown a profile of one participant in a speed dating event along with some information about his or her date; B: based on this profile, subjects are asked to make an initial binary prediction of the speed dating event outcome; C: the ML model's binary prediction and confidence score are then shown to subjects; D: subjects are asked to make a final prediction of the speed dating event outcome; E: subjects are asked to indicate their belief in the model's prediction being correct; depending on whether subjects' final prediction agreed with the model's prediction or not, we rephrased this question as equivalent to reporting subjects' belief in their own final prediction being correct (if agreed) or wrong (if disagreed).

- *Step 1*: the subject needed to carefully review the profile and make her own binary prediction about whether or not the participant in the profile would want to see the date again (Figure 1B).
- *Step 2*: then, the subject was shown an ML model's binary prediction on the current case as well as a confidence score between 0 and 1 representing how confidently the model believed its prediction was correct; the higher the score, the more certain the model was that it made a correct prediction (Figure 1C).
- *Step 3*: after considering both her own prediction and the model's prediction, the subject was asked to make a final binary prediction on whether or not the participant would want to see the date again (Figure 1D).
- *Step 4*: finally, before moving on to the next task, the subject needed to indicate her belief in the ML model's prediction being correct as a number between 0 ("the model's prediction is absolutely wrong") and 1 ("the model's prediction is absolutely correct")[1]. To help the subject contextualize this question, depending on whether the subject's final prediction in the task agreed with the model's prediction, we further indicated to subjects that this question is effectively the same as reporting how much they believe their final prediction was "also correct" (if agreed) or "wrong" (if disagreed) (Figure 1E).

The task of predicting romantic relationship is suitable for our study for two main reasons. First, such a task does not require special expertise or domain knowledge and involves relatively limited risks, so it can be easily understood by our laypeople subjects (i.e., MTurk workers) while still representing realistic *low-stakes decision-making* tasks that laypeople undertake in their day-to-day life well. Second, ML models have been developed to make predictions in romantic attraction and compatibility [24, 28, 41], which makes the experimental setting sufficiently credible and ensures the ecological validity of our study. Note that similar prediction tasks were previously used in [58] to explore the effects of model accuracy *alone* on people's trust in ML models.

3.2 Experimental Procedure

Figure 2 shows a flowchart of our experiment, in which each subject completed a sequence of 40 prediction tasks that were divided into two phases. Specifically, before a subject started to work on any prediction tasks, we explained the prediction tasks to the subject and walked through an example of the task interface with step-by-step instructions detailing each component of the speed dating profile (Figure 1A). Subjects were also given basic written instructions on how to complete the task and explaining the meaning of a confidence score assigned to a prediction by the model. We also revealed the ML model's stated accuracy to the subject by stating "We previously evaluated this model on a large data set of speed dating participants and its accuracy was x%, i.e., the model's predictions were correct on x% of the speed dating participants in this data set." After reading the instruction, the subject started Phase 1 of the experiment to work on a set of 20 prediction tasks in succession, and in each task, the subject followed the 4-step procedure as described above. In particular, when the model's binary prediction was shown to the subject in Step 2 of each task, we communicated the model's confidence about this prediction to the subject through the sentence "The model makes this prediction with a confidence score of y, (i.e., the model believes the chance for this prediction to be correct is $100\times y$%)," and we also reminded the subject of the model's stated accuracy (Figure 1C).

The subject received *no* feedback on whether her prediction or the model's prediction was correct on each of the individual tasks in Phase 1. However, after all 20 tasks in Phase 1 were completed,

[1]Both a slider and a text box were provided for subjects to indicate their belief in the model's correctness, and the number could be reported to the hundredths place.

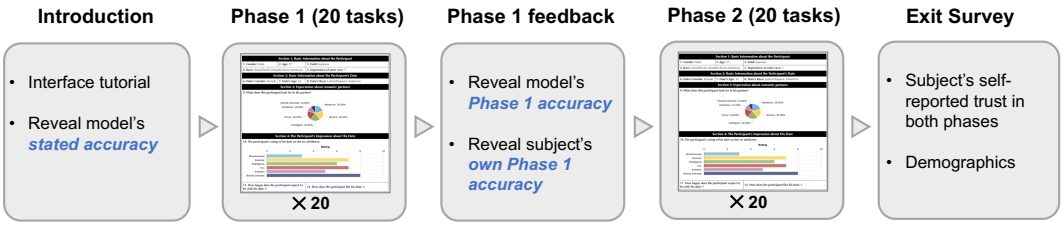


Figure 2: Flowchart of the experiment.

we provided a brief summary to the subject reporting the model's overall accuracy on the 20 tasks of Phase 1 (i.e., the model's "observed accuracy") along with the subject's own accuracy on those 20 tasks before seeing the model predictions (accuracy feedback on individual tasks was not given though). Then, the subject went on to complete another set of 20 prediction tasks in Phase 2 following a similar procedure. The only difference was that whenever the model prediction was shown in each of the tasks in Phase 2, in addition to displaying the model's confidence, we reminded the subject of not only the model's stated accuracy, but also its observed accuracy on the 20 tasks in Phase 1.

Finally, after completing all 40 prediction tasks, the subject was asked to report her level of trust in the ML model in each phase of the experiment by answering two exit-survey questions:

- How much did you trust our machine learning model's predictions on the *first* twenty speed dating participants (that is, *before* you saw any feedback on your performance and the model's performance)?
- How much did you trust our machine learning model's predictions on the *last* twenty speed dating participants (that is, *after* you saw any feedback on your performance and the model's performance)?

The subject answered these questions using a scale from 1 ("I didn't trust it at all") to 10 ("I fully trusted it"). We also collected some basic demographic information (e.g., gender, age, and highest level of education) from the subject through the exit survey.

3.3 Experimental Treatments

We considered a total of 8 experimental treatments with a 2×2×2 design along three factors:

- *Confidence level*: the level of confidence scores that the ML model associates to its predictions, which has two levels — low and high. In the low confidence treatments, the confidence scores of the ML model on all 40 tasks were between 0.5 and 0.8, while the confidence scores of the ML model for the high confidence treatments were between 0.8 and 1 on all tasks.
- *Stated accuracy*: the ML model's stated accuracy on the held-out data, which has two levels — 60% and 90%.
- *Observed accuracy*: the ML model's accuracy on the Phase 1 tasks (i.e., the first 20 tasks), which also has two levels — 55% and 95%.

Figure 3 illustrates the design of our experimental treatments. To minimize the differences across treatments as much as possible, we had subjects in different treatments see exactly the *same* 40 prediction tasks. The predictions of ML models shown to subjects in different treatments, including both the binary predicted labels and the confidence scores, were produced by real ML models that we developed prior to the experiment deployment. We trained 4 ML models for this experiment, including a neural network, a random forest, a naive Bayes classifier, and a support vector machine (SVM)[2]. For the neural network, random forest, and naive Bayes models, we directly used the conditional probability of the predicted label as the confidence score. For the SVM model, we adopted Platt scaling [46] to compute the probability of the predicted label given the binary prediction and used that as the confidence score. On the 40 tasks in our experiment, the Pearson correlation between the model's confidence on a task and the model's accuracy on that task was found to be positive for all 4 models, though with different levels of significance (neural network: r=0.643, $p < 0.001$; random forest: r=0.281, p=0.079; naive Bayes: r=0.175, p=0.281; SVM: r=0.254, p=0.114).

Note that in our experiment, for any two treatments with the same level of model confidence and observed accuracy, but different stated accuracies, model predictions shown to the subject were taken from the same pre-trained ML model. For example, as shown in Figure 3, T1 (i.e., "stated-60%, observed-55%, low confidence") and T2 (i.e., "stated-90%, observed-55%, low confidence") shared exactly the same model predictions (including both binary predicted labels and confidence scores) on all 40 tasks, which were generated from a single SVM model[3]. As such, the only difference between these two treatments is the level of stated accuracy for the ML model.

3.4 Other Experimental Control

Our experiment was implemented as a Human Intelligence Task (HIT) on Amazon Mechanical Turk (MTurk), and we limited the subjects of our experiments to be U.S. workers on MTurk only. Upon arrival, each subject was *randomly* assigned to one of the eight treatments.

The 20 prediction tasks in Phase 1 were carefully selected such that ML models for the 4 treatments with a 55% model's observed accuracy (i.e., T1–T4) always had the *same* binary predicted labels on each of these 20 tasks, and 11 out of these 20 labels were correct

[2]These models were chosen because they naturally tend to produce different extremes of observed accuracies and confidence scores. For instance, neural networks can be trained to have high observed accuracy, but they are often overconfident in their incorrect predictions, leading to inflated confidence scores.

[3]We note that in reality, different levels of stated accuracy can be claimed for a single ML model when the set of held-out data on which the model is evaluated is different.

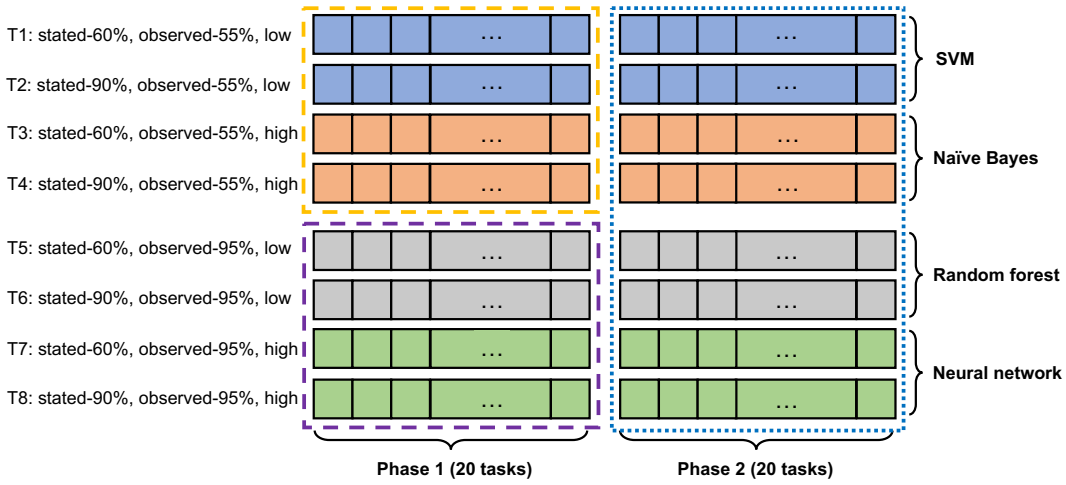


Figure 3: Design of experimental treatments. Model predictions presented to subjects in treatments with the same level of model confidence and observed accuracy (e.g., T1 and T2, or any two treatments shown in the same color) were taken from the same pre-trained ML model. In Phase 1, treatments with the same level of observed accuracy (i.e., T1–T4 as highlighted in the yellow dashed box, or T5–T8 as highlighted in the purple dashed box) had the same binary predicted labels on each of the 20 tasks (but model confidence differed between low confidence treatments and high confidence treatments). In Phase 2, all treatments (i.e., T1–T8 as highlighted in the blue dotted box) had the same binary predicted labels on each of the 20 tasks (but again, model confidence may differ).

(see the yellow dashed box in Figure 3). Similarly, ML models for the other 4 treatments with a 95% model's observed accuracy (i.e., T5–T8) also had the *same* binary predicted labels on each of these 20 tasks, while 19 out of these 20 labels were correct (see the purple dashed box in Figure 3). So, within each set of 4 treatments that shared the same level of observed accuracy, the only differences among these treatments in Phase 1 were the model's stated accuracy and its confidence score for each prediction. Importantly, since there was *no* overlap in the range of confidence scores between the low confidence treatments and the high confidence treatments, on *any* of the 20 prediction tasks in Phase 1, the confidence score produced by models of low confidence treatments was lower than the score produced by models of high confidence treatments.

Moreover, the 20 prediction tasks in Phase 2 were also carefully selected such that ML models for all 8 treatments made exactly the *same* binary predictions on each of them (see the blue dashed box in Figure 3). 16 out of these 20 predictions were correct, although the subject received no feedback on the model's accuracy during Phase 2. So, the only differences across all eight treatments in Phase 2 were the model's stated accuracy, observed accuracy in Phase 1, and confidence scores that the model associated with its predictions. Again, on *any* of the 20 tasks in Phase 2, models of high confidence treatments were more confident about their prediction than models of low confidence treatments.

The order of prediction tasks was randomized within each phase. To incentivize high-quality predictions from subjects, in addition to the $1.5 base payment that each subject was guaranteed to receive once they submitted the experiment HIT, we also provided performance-contingent bonuses—we told the subject that we would randomly select one task from the 40 prediction tasks, and we would pay a $1 bonus to her if her *final* prediction on that task was correct. As the median amount of time subjects spent on our HIT was about 17 minutes, the bonus payment we provided was roughly equivalent to an additional hourly wage of $3.5/hour; this is 75% higher than workers' median hourly wage on MTurk (about $2/hour, [26]) and likely provided considerable motivation for workers to carefully decide whether to trust the ML model in their decision making.

4 DATA

After removing workers who had accidentally completed our HIT more than once, we were left with a total of 1,224 unique subjects who participated in our experiment[4]. Among these subjects, 42.3% of them were female, and their average age was 35. When each subject worked on a prediction task, we recorded her initial prediction on whether the participant in that task would want to see the date again before seeing the model's prediction, her final prediction after seeing the model's prediction, and her reported belief on how likely the model's prediction would be correct. At the end of the experiment, we also recorded the subject's responses to the survey questions, which asked her to report the level of trust she had on our model in each phase. Based on these data, we used four different measures to quantify the subject's trust in the ML model, and we computed the values of these measures separately for each phase:

- *Subject's belief in model accuracy (belief)*: the average value of the number that the subject gave in Step 4 of each prediction task indicating how much they believed the model's prediction would be correct.

[4]The number of subjects in T1–T8 were 163, 156, 138, 150, 145, 159, 154, and 159.

- *Agreement fraction (agreement)*: the number of tasks that the subject's *final* prediction agreed with the model's prediction, divided by the total number of tasks.
- *Switch fraction (switch)*: the number of tasks for which the subject initially made an opposite prediction as compared to the model, but after seeing the model's prediction she decided to *switch* her final prediction to agree with the model, divided by the number of tasks that the subject initially disagreed with the model.
- *Self-reported trust level (self-report)*: the level of trust that the subject reported to have in the ML model in the exit survey.

We note that measures we adopted here were among the most common ones that have been used in previous literature to quantify trust in ML models. For example, subject's belief in model accuracy was used in [31, 50], agreement or switch fraction was used in [35, 58, 61], and subject's self-reported level of trust in the model was used in [10, 58]. However, subtle differences also exist between these trust measures. While subject's belief in model accuracy and self-reported trust level tend to characterize how much subjects "think" they trust the ML model, the other two metrics—agreement fraction and switch fraction—focus more on measuring trust by examining how subjects behave (e.g., how often did a subject "follow" a model's prediction?). Intuitively, for all 4 trust measures, larger values imply higher levels of trust.

Our data suggests that human subjects are, in general, not very good at making accurate predictions for speed dating outcomes by themselves. Specifically, the average accuracy of a subject's initial predictions before seeing the model's predictions was 53.6% in Phase 1 and 62.8% in Phase 2, and it was not significantly different across treatments (in contrast, depending on the experimental treatment the subject was in, the model's accuracy was 55% or 95% in Phase 1, and 80% in Phase 2).

5 RESULTS

In this section, we report our findings on how laypeople are affected by multiple performance indicators of an ML model when this model is provided to assist them in low-stakes decision making.

5.1 Analyzing Trust in Phase 1

We start by analyzing the experimental data that we obtained in Phase 1 to understand overall how laypeople's trust in an ML model is affected by both the model's stated accuracy and its confidence *before* they observe the model's performance in practice (i.e., view the model's actual accuracy in Phase 1). Before people have had the chance to observe the ML model's performance in practice, research questions we are interested in examining include:

- **RQ1**: Does a higher level of confidence that the ML model associates with a prediction make people trust the prediction more?
- **RQ2**: Is people's trust in a model prediction still affected by the model's stated accuracy, even though the model's confidence on that prediction is provided?
- **RQ3**: How does the effect of model confidence on trust vary between models with different levels of stated accuracy?

To answer these questions, we used the model's stated accuracy and confidence level as our independent variables, and the four trust measures as described above (i.e., belief, agreement, switch, and self-report) were used as the dependent variables. Recall that in Phase 1, models in T1–T4 (i.e., the 4 treatments with a 55% model accuracy in Phase 1) made different predictions on the tasks than models in T5–T8 (i.e., the 4 treatments with a 95% model accuracy in Phase 1). As a result, to separate *only* the effects of model's stated accuracy and confidence on trust for **RQ1**–**RQ3**, we analyzed the Phase 1 data by comparing trust *within* each set of 4 treatments with the same Phase 1 model predictions.

To emphasize the exploratory nature of this study, in addition to controlling for false discovery and avoiding the issues with multiple comparisons, we conducted our analysis using the interval estimate method [14, 20]. To analyze the effects of our independent variables on the four trust measures, we plotted effect sizes for a given trust measure between subjects assigned to different levels of a given independent variable (e.g., levels 60% and 90% for stated accuracy). Effect sizes were measured using Cohen's d[5], and we also plotted their 95% bootstrap confidence intervals ($R = 5000$). To indicate the size of an effect, we also followed Cumming (2013) in considering a confidence interval's range in relation to zero [13]. We conducted this analysis separately for both confidence and stated accuracy (answering **RQ1** and **RQ2**, respectively), and this was repeated for both T1–T4 and T5–T8[6].

Meaningful interaction effects were identified by conducting a two-way ANOVA (stated accuracy × confidence) for each of the 4 dependent variables *within* T1–T4 (and *within* T5–T8), and we followed the same interval estimate method approach for interactions worth noting (answering **RQ3**)[7]. Unlike in previous analysis using the interval analysis method to illustrate the main effect of a single independent variable, we calculate it for interaction effects using a *difference in difference*: Consider the interaction effect between two independent variables A (with two levels A_1 and A_2) and B (with two levels B_1 and B_2) on a measure of trust Y. The difference in difference is then defined as the difference in Y between two treatments that both belong to the higher level of A (i.e., A_2) and differ on the level of B, minus the difference in Y between two treatments that both belong to the lower level of A (i.e., A_1) and differ on the level of B.

5.1.1 RQ1: The Main Effect of Confidence in Phase 1. We first analyzed the data obtained in the 4 treatments with a 55% model's observed accuracy (i.e., T1–T4). As shown in Figure 4a, when the ML model associates a higher confidence score to its prediction, people believe that the prediction (*belief*) is more likely to be correct (Cohen's d=0.72 [0.55, 0.89]). On average, subjects in the high confidence treatments considered the model predictions to be 7.2% more accurate compared to the same model predictions that are shown in the low confidence treatments (i.e., ΔM=0.072). Beyond that, it also appears that higher model confidence on a prediction *may* nudge people into following the prediction more often, both in terms of how often people agree with the model's predictions (*agreement*) and in how often they will switch to agree with the model's predictions (*switch*). However, we find that the evidence

[5] In computing Cohen's d, we always treat the level with the lowest value(s) as the baseline group.
[6] For figures of raw data distributions and a summary of all estimated effect sizes as well as their 95% bootstrap confidence intervals, see the supplementary materials.
[7] For the ANOVA test results, see the supplementary materials.

(a) T1–T4 (observed-55%) **(b) T5–T8 (observed-95%)**

Figure 4: The difference in trust between subjects with a high confidence treatment and a low confidence treatment in Phase 1. Cohen's *d* values are plotted and listed above each point, and error bars represent 95% bootstrap confidence intervals. An interval to the right (or left) of 0 represents a higher (or lower) mean for the subjects in a high confidence treatment. Measures of trust that see a difference between treatments are also shown with a red and bolded Cohen's *d* value.

of this impact on trust is not as compelling for agreement fraction (Cohen's d=0.12 [−0.04, 0.28]) or for switch fraction (Cohen's d=0.15 [−0.01, 0.31]) as it was for belief in the model's predictions. In terms of self-reported trust in the model (*self-report*) in Phase 1, we found no evidence that model confidence has an impact.

We next looked at the other 4 treatments in Phase 1 with a 95% model's observed accuracy (i.e., T5–T8). Similar to T1–T4, we found that subjects given high confidence predictions believe that the prediction is more likely to be correct (Cohen's d=0.47 [0.31, 0.64]), as seen in Figure 4b. We also continue to see that a higher model confidence *may* influence people to agree with the model more frequently (Cohen's d=0.16 [−0.01, 0.32]), but we find no impact of model confidence on how often people will switch to a model's predictions. Again, model confidence is not found to have an effect on people's self-reported trust in Phase 1.

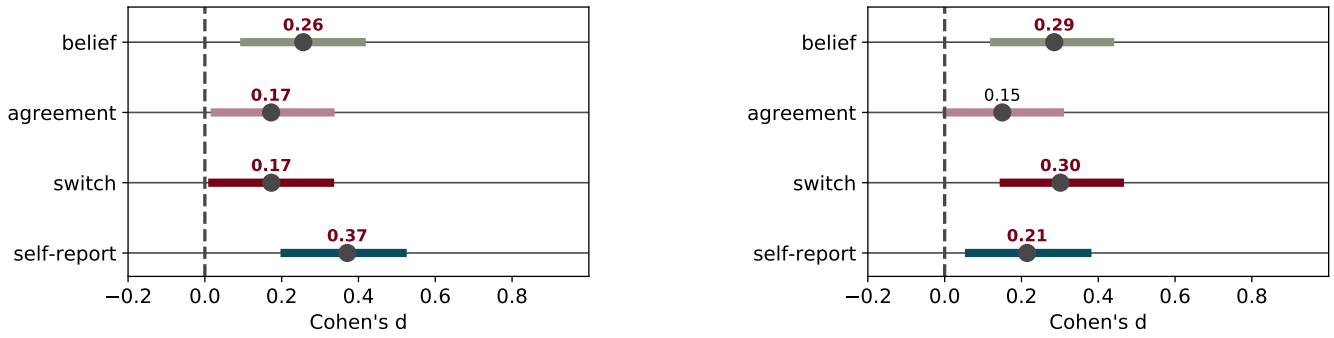
Putting it all together, to answer RQ1, our results suggest that before observing an ML model's accuracy in practice, people believe a model with high confidence scores to be more accurate, but do not self-report to trust it more. In terms of following the model's predictions, we find some evidence that a higher model confidence can have greater influence, but the evidence is not reliable.

5.1.2 RQ2: The Main Effect of Stated Accuracy in Phase 1. We now look into whether the stated accuracy of an ML model can still influence people's trust in the model in the presence of model confidence, before the model's accuracy is observed in practice.

A visual inspection of Figures 5a and 5b indicates a positive answer. For example, when focusing on T1–T4, we found that overall, claiming the model's stated accuracy to be 90% rather than 60% led to an increase of 2.7% in the subject's belief in model accuracy (Cohen's d=0.26 [0.09, 0.42]), an increase of 2.9% in agreement fraction (Cohen's d=0.17 [0.02, 0.34]), an increase of 5.5% in switch fraction (Cohen's d=0.17 [0.01, 0.34]), and an increase of 7.7% in self-reported trust (Cohen's d=0.37 [0.20, 0.53]). These results highlight the effect of stated accuracy *alone* on people's trust given that the model's prediction between the two treatments (including both binary predicted labels and confidence scores) on each task were kept unchanged. This indicates that despite the fact that fine-grained information of model confidence is provided at the level of individual predictions, an ML model's stated accuracy still casts consistent and significant impact on people's trust in the model—the higher the stated accuracy, the more people trust the model.

To put the effect sizes (as measured by Cohen's d) of stated accuracy on trust in Phase 1 into context, we can compare them against the effect sizes of model confidence on trust. We find that while the model confidence has a larger effect in influencing subjects' belief in model accuracy, the model's stated accuracy tends to have a larger and more reliable impact on subjects' willingness to follow the model's prediction as well as their self-reported trust levels.

5.1.3 RQ3: The Interaction between Confidence and Stated Accuracy in Phase 1. Lastly, we examine the *interaction effect* between model confidence and the model's stated accuracy on influencing people's trust in the model in Phase 1. The only meaningful interaction effect we found is with respect to subject's belief in model accuracy. In particular, when an ML model has a high stated accuracy, the increase in subject's belief in model accuracy brought up by the increase in model confidence is larger than that for an ML model with a low stated accuracy (*T1–T4*: Cohen's d=0.32 [−0.01, 0.64], *T5–T8*: Cohen's d=0.38 [0.07, 0.70]). In other words, the magnitude of the effect of model confidence on how much people believe the predictions to be correct varies depending on the levels of the model's stated accuracy. Alternatively, we can also interpret this as that model confidence moderates the effect of the model's stated accuracy on subject's belief in model accuracy—when the model is more confident, an increase in the model's stated accuracy will lead to a more significant increase in subject's belief of how accurate the model is.



(a) T1–T4 (observed-55%)

(b) T5–T8 (observed-95%)

Figure 5: The difference in trust between subjects with a high stated accuracy treatment and a low stated accuracy treatment in Phase 1. Cohen's *d* values are plotted and listed above each point, and error bars represent 95% bootstrap confidence intervals. An interval to the right (or left) of 0 represents a higher (or lower) mean for the subjects in a high stated accuracy treatment. Measures of trust that see a difference between treatments are also shown with a red and bolded Cohen's *d* value.

5.2 Analyzing Trust in Phase 2

We now turn our attention to the data that we obtained in Phase 2 to examine overall how laypeople's trust in an ML model is affected by the model's stated accuracy, observed accuracy, and model confidence *after* they learn about the model's overall accuracy in Phase 1. Since subjects in all treatments worked on the same prediction tasks and saw the same binary prediction on each of the tasks in Phase 2, we can directly compare the trust measurements across all 8 treatments. In this set of analyses, we are interested in exploring the following research questions after people have observed an ML model's performance in practice:

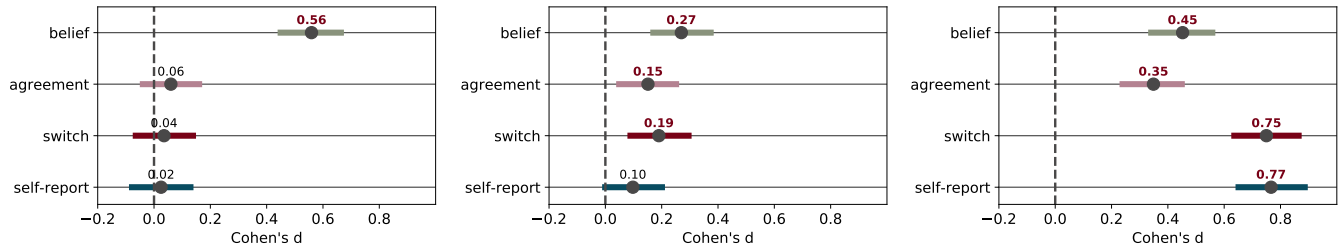
- **RQ4**: Does a model's confidence score still influence people's trust in the model?
- **RQ5**: Do the stated accuracy and observed accuracy of the model still affect people's trust in the model, despite model confidence still being provided for each individual prediction?
- **RQ6**: How do model confidence, the model's stated accuracy, and the model's observed accuracy interact with each other to influence trust?

To answer these questions, for each of the four trust measurements, we repeated our Phase 1 analysis method. This time, however, we included all 8 treatments in the analysis, and we included observed accuracy along with confidence and stated accuracy as independent variables to consider. Employing the interval estimate method, we answer **RQ4** by examining the effect of confidence in Phase 2, and we answer **RQ5** by looking into the effects of stated accuracy and observed accuracy in Phase 2. Meaningful interaction effects were identified by conducting a three-way ANOVA (stated accuracy × observed accuracy × confidence) for each of the four trust treatments, and again we followed the interval estimate method by calculating a difference in difference for these interactions to answer **RQ6**.

5.2.1 RQ4: The Main Effect of Confidence in Phase 2. Figure 6a shows the effect of model confidence on each of the four trust measures for Phase 2. Consistent with our Phase 1 results, we found that subjects still tended to believe that models producing higher levels of confidence scores are more likely to be correct in Phase 2 (Cohen's d=0.56 [0.44, 0.67]). This suggests that even after observing the model's accuracy in practice, subjects still believe a prediction with a high confidence score as more likely to be correct. However, it does not appear that the model confidence has any impact on agreement fraction, switch fraction, or self-reported level of trust. That is to say, in Phase 2, subjects did not seem to follow a model prediction more often when a high confidence was associated with it as compared to when a low confidence was associated with it, nor did they feel that they trust a model more when the confidence scores it produced were higher.

5.2.2 RQ5: The Main Effect of Stated and Observed Accuracy in Phase 2. Figures 6b and 6c show that even though model confidence—the model's own accuracy estimate for individual predictions—is provided on each prediction, both the stated accuracy and the observed accuracy of an ML model still have a significant impact on people's trust in the model in Phase 2. After observing the model's performance in practice, subjects not only believed predictions made by a model with higher stated or observed accuracy to be more accurate (*stated*: Cohen's d=0.27 [0.16, 0.38], *observed*: Cohen's d=0.45 [0.33, 0.57]), but also agreed with such predictions more often (*stated*: Cohen's d=0.15 [0.04, 0.26], *observed*: Cohen's d=0.35 [0.23, 0.46]), and switched their answer to match such predictions more often (*stated*: Cohen's d=0.19 [0.08, 0.31], *observed*: Cohen's d=0.75 [0.63, 0.88]). A higher observed accuracy also led subjects to self-report that they trusted the model more (Cohen's d=0.77 [0.64, 0.90]), and it also appears that a higher stated accuracy *may* have a similar effect (Cohen's d=0.10 [−0.01, 0.21]), though this result is less compelling.

When comparing the effect sizes of stated accuracy and observed accuracy on trust in Phase 2 against those of the model confidence, we found the model's stated accuracy and observed accuracy have a larger impact on all trust measures except for subject's belief in



(a) Confidence **(b) Stated Accuracy** **(c) Observed Accuracy**

Figure 6: The difference in trust between subjects with different levels of the given independent variable. Cohen's *d* values are plotted and listed above each point, and error bars represent 95% bootstrap confidence intervals. An interval to the right (or left) of 0 represents a higher (or lower) mean for the subjects in the higher valued treatment. Measures of trust that see a difference between treatments are also shown with a red and bolded Cohen's *d* value.

model accuracy. In particular, a model's observed accuracy dominates the model's stated accuracy and confidence on influencing the subject's willingness to follow the model predictions or their self-reported level of trust in the model in Phase 2. These observations imply that after an ML model's performance has been observed in practice, its stated accuracy and observed accuracy still significantly influence people's trust in the model, and the impact of the observed accuracy is especially substantial.

5.2.3 RQ6: The Interaction between Confidence, Stated Accuracy, and Observed Accuracy in Phase 2. Finally, we investigate whether different factors interact with each other in influencing people's trust in the model in Phase 2. We first note that the three-way interaction between the model's confidence, stated accuracy, and observed accuracy is *not* meaningful for any of the four trust measures, implying that the strength of the interaction between any of the two factors (e.g., stated and observed accuracy) on trust is not dependent on the level of the third factor (e.g., confidence).

We then move on to examine the two-way interactions between any pair of influencing factors. Similar to that in Phase 1, we again detected an interaction between model confidence and stated accuracy on subject's belief in model accuracy, suggesting that after a model's performance is observed in practice, increasing the model's confidence on a prediction still leads to a larger increase in subject's belief in the correctness of the prediction when the model's stated accuracy is higher (Cohen's d=0.23 [0.003, 0.46]). Additional meaningful interactions that we found are between a model's stated accuracy and observed accuracy in influencing people's trust in the model with respect to switch fraction (Cohen's d=0.41 [0.18, 0.64]) and self-reported trust (Cohen's d=0.43 [0.19, 0.65]), which are similar to results previously reported in [58]. That is, after subjects have had the opportunity to observe a model's accuracy in practice, higher levels of stated accuracy only push people to follow a model more often and report a higher level of trust in the model if its observed accuracy is high.

In sum, to answer RQ6, we indeed found evidence suggesting that after people observe an ML model's performance in practice, there are some interactions between model confidence, stated accuracy, and observed accuracy that influence people's trust in the model, as quantified by different measures.

6 DISCUSSIONS

In this paper, we conduct an exploratory study to investigate how laypeople's trust in an ML model is affected by both the model's confidence and its accuracy. Our results suggest that model confidence and model accuracy play different roles in influencing people's trust in an ML model. Model confidence mostly influences people's belief in model accuracy, but the model's stated accuracy and observed accuracy consistently impact how much people think they trust a model and how frequently they actually follow the model. In this section, we begin by discussing potential explanations for the limited impact that model confidence had in our study, as well as why we see an inconsistency in our results across different measures of trust. Next, we provide implications for future design and caution readers on the generalizability of our results. Finally, we end with possible directions for future work.

6.1 Understanding the Limited Impact of Model Confidence

Compared to our observations that an ML model's stated accuracy and observed accuracy have significant impact on people's trust in the model, as consistently shown in all four measures of trust that we adopt, the impact of model confidence on trust seems to be more limited. The only consistent effect of model confidence on trust is with respect to people's belief of model accuracy; regardless of whether the model's performance is observed in practice, the more confident a model, the more accurate people believe the model is. In contrast, model confidence doesn't influence the self-reported level of trust in the ML model, and it doesn't reliably influence people's willingness to follow a model.

As to *why* we see this result, it is possible that when both model accuracy and model confidence are presented, people consider accuracy as a *fact*, but deem confidence as an *estimate*, and thus model confidence is treated as a less trustworthy type of performance information. Such perception may become particularly strong after people have a chance to observe a model's accuracy on real-world

trials themselves. This is consistent with what we have observed in our data, which suggests that people substantially rely upon the model's observed accuracy to decide their trust in an ML model in Phase 2, effectively leaving the model confidence almost ignored. This may also explain the difference between our findings and the findings in prior literature indicating that model confidence is directly related to users' levels of trust in the system [3, 36, 54, 61]: when a confidence score is the only performance indicator of a model shown, the score's value may have greater influence on people's trust in its predictions.

6.2 On the (Seemingly) Conflicting Results Across Different Trust Measures

In our experiment, we also find that the effects of model confidence on different trust measures are not always consistent. This raises an important question on how to appropriately operationalize people's "trust" in ML models in experimental studies. To this end, we first note that different trust measures can capture different "types" of trust. For example, one can consider subject's belief in model accuracy and self-reported trust as the "stated trust" or trust perceptions, while agreement and switch fractions represent the "revealed trust," or trusting behaviors.

Our results, for example, show that the effect of model confidence on people's tendency to follow a model (i.e., revealed trust) can be different than that on people's belief in model accuracy or self-reported trust (i.e., stated trust). While this is consistent with various previous literature suggesting that subjective self-reported trust may not be a reliable quantification of trusting behavior [34, 49] (i.e., "saying and doing are two different things"), we conjecture that there may also exist some inherent relationships between different trust measures that could explain the inconsistency. For instance, one possible relationship between the belief in model accuracy and the decision to follow a prediction of an ML model could be that the subject would only be willing to adjust her own prediction to match with that of the model's if her belief in the accuracy of the prediction is *above a threshold*. If this was true, then it's possible to see model confidence has a meaningful impact on subject's belief in model accuracy but not on agreement or switch fractions, just as what we have seen in Phase 2 of our experiment, and to some extent, Phase 1 as well. Exploring how various trust measures relate to one another, perhaps by understanding the reasoning process behind people's trust decisions, can be another important direction which may significantly advance our understanding of trust in ML.

Interestingly, in our experiments, we also found the effects of model confidence on the two stated trust measures (i.e., belief in model accuracy and self-reported trust) can be different, which seems puzzling at first glance. One explanation for this is that subjects experienced the anchoring effect when providing belief in model accuracy. In particular, model confidence was described to subjects as the chance that the model believes for its prediction to be correct, and we solicited subject's belief in model accuracy by asking how likely she thought the model's individual prediction was correct. In addition, both confidence and belief in model accuracy were represented as a value between 0 and 1, perhaps making them analogous to one another. This may have made it easy for subjects to use model confidence as a reference point for their belief in the model's accuracy, causing them to subconsciously weigh model confidence higher than other factors such as stated and observed accuracy in their final decision.

On the other hand, the self-reported trust measure was asked upon completing the prediction tasks, and it was asked in the form of a Likert scale. As such, the belief in model accuracy can affect, but does not necessarily determine, one's self-reported trust. Self-reported trust is a more holistic judgement of ML model's performance, integrity and intention, reflecting the subject's feeling about the model as a whole, and can also be influenced by one's emotional state like surprise or confusion.

Finally, we acknowledge that trust is a complex concept and a multidimensional construct. Though we attempted to measure trust using a diverse set of metrics including both behavioral and self-reported methods, further studies are needed to understand how to, for example, design reliable scales to probe into various aspects of trust in ML models. In particular, there has been extensive literature in various fields that discusses how to define, model, and quantify trust in computational environments [7, 12], which can provide useful guidance in the future on how to appropriately define trust in the context of interactions between humans and ML models.

6.3 Design Implications

In our findings we saw that after the model's performance is observed in practice, people's trust in the model is affected dominantly by model's observed accuracy, yet it is hardly affected by model's confidence. This indicates the potential needs of helping laypeople to better understand the uncertainty inherent in performance calculation and calibration estimation based on a small set of predictions. As we have discussed earlier, if people indeed simply ignore model confidence in deciding their trust in a model, especially after observing its accuracy in practice, then it is crucial to help people recognize that doing so can be sub-optimal, since any accuracy measurements computed based on a small number of trials may not reflect the model's overall performance accurately. Even if the accuracy measurement is reliable, there is still great value in utilizing a *calibrated* confidence score to calibrate trust in an ML model. On the other hand, people might have actually adjusted their interpretation of confidence scores after they have seen the model's accuracy in practice and thus obtained an estimation of how calibrated the model's confidence is. If this is the case, the key message that needs to be conveyed to people becomes that the degree of confidence calibration estimated based on a limited number of predictions can also be inaccurate, especially when confidence scores for these predictions all lie in a small range (e.g., 0.95–1).

In other words, it is critical for people to see the value of continued use of a calibrated confidence score to adjust their trust in an ML model, even after observing a very high or a very low accuracy of the model in practice (instead of simply always trusting or not trusting the model, regardless of the model confidence). Meanwhile, the estimated degree of calibration for a model's confidence based on a small set of predictions may not be very accurate. To this end, a tool that assists people in updating their estimation of how calibrated a model's confidence is as they interact with the model, and perhaps even quantifying and visualizing the uncertainty of such estimation [21, 29], can potentially be very helpful.

6.4 On the Generalizability of Our Results

Despite our findings, proper caution should be used when generalizing our results to different experimental settings, task domains, and subject populations. In terms of settings, to keep the size of the experiment manageable, we included only two levels for each of our influencing factors. The levels we chose for each influencing factor are relatively extreme, with the hope of identifying clear effects by dichotomizing the levels. It is thus unclear what our experimental results would be if an influencing factor takes a value that is substantially different from our chosen levels (e.g., after people observe an intermediate level of model accuracy, say 80%, whether and how the model confidence will affect people's trust in the ML model?). Our task interface may have further constrained our results. Had we communicated multiple performance-related information of the ML model to people in a different way (e.g., communicate model confidence to people verbally, using terms like "virtually certain to be correct", "likely correct", etc., or using visualizations), we may have obtained different results. Limiting or expanding the amount of information that we showed to subjects for each speed dating task profile itself may have also had an impact on how much attention was given to the model's performance information.

Additionally, we recognize the potential for "confirmation bias" from our subjects in determining whether to trust the ML model, given that they were always asked to make an independent prediction first before seeing the ML model's prediction. In this way, subjects whose independent predictions often agreed with the model may have speculated that the model's accuracy was higher than it was in practice, leading to over-trust in its predictions, and vice-versa. Indeed, prior research has shown that trust in peer assessment is lowered when prior expectations are not met [30], and within the context of trust in ML and its predictions, in the absence of information about an ML model's performance, people's reliance on the model is dependent on how often the model's predictions agree with their own [39]. We note that the possible existence of this confirmation bias mainly influences the absolute magnitude of the four trust measures we used in this study, but is unlikely to be a confounding variable for our main results regarding whether and how model accuracy and model confidence together affect trust. This is because all our analyses were conducted only within the set of treatments where the corresponding ML model shared the *same* binary predictions on tasks (i.e., within $T1$–$T4$ and within $T5$–$T8$ separately for Phase 1 analysis, and within $T1$–$T8$ for Phase 2 analysis), so the average degree that subjects suffered from their biases should be similar across treatments in such a set due to the random assignment. However, we acknowledge that had our experimental tasks been designed in a different way (e.g., showing the model's prediction before eliciting the subject's), our findings might change.

Finally, while the type of task (i.e., predicting speed dating outcomes) and the kind of human subjects (i.e., MTurk workers) we chose suit the purpose of our study—that is, to understand how laypeople's trust in an ML model is affected by multiple performance indicators of the model in their low-stakes decision making—we caution readers to generalize our results to other populations or other tasks. For example, it is unclear whether our results will still hold when users of the ML model have better knowledge on uncertainty quantification (e.g., data scientists) or on the domain itself (e.g., experts of human behavior). Our goal of understanding how laypeople trust ML models in low-stakes decision making implies that the decision of how much to trust the ML model does not have particularly impactful or long-term consequences. Whether similar results on the effects of various performance indicators can be obtained for tasks with higher stakes, such as making predictions of prognosis or recidivism, is unknown. We also note that though the task of predicting speed dating outcomes is easy enough for laypeople to understand and make meaningful predictions, it turns out that our subjects are not good at making correct predictions on this task by themselves. Had the prediction task been significantly easier or more difficult for people, different conclusions may have been drawn.

6.5 Future Work

With the exploratory evidence obtained in this study, we would like to conduct more confirmatory studies in the future to validate these results and examine their generalizability under different settings, for different populations, and on different tasks. There are also many other types of performance indicators of an ML model beyond accuracy and confidence, such as precision, recall, and F-1 score. Exploring how different combinations of performance indicators of a model affect people's trust in the model differently is another future direction. Ultimately, we seek to gain understanding as to why various indicators of model performance affect people's trust in the way that they do, so as to build a theoretical framework on humans' perception, processing, and comprehension of performance information of machine learning.

7 CONCLUSIONS

As machine learning becomes more ubiquitous in everyday life, understanding how laypeople trust the predictions of ML models becomes increasingly important. In this work, we explore the impact of multiple performance indicators of an ML model on laypeople's trust in the model. Our results suggest that, in general, the confidence an ML model ties to its predictions significantly influences how much laypeople claim to believe in the model's individual predictions, but performance measurements like the model's accuracy on a held-out set of data and observed accuracy in practice have greater influence on how often laypeople will actually follow the model as well as their self-reported overall trust in the model. We aim for this work to be a step towards further study into the process that people's perception and confidence in the predictions of machine learning models are shaped by diverse information about the model that they receive and how this process eventually impacts trust. We hope exploratory evidence we report in this work will inspire more discussions in this direction.

ACKNOWLEDGMENTS

We thank Siddharth R. Shah for his participation in the early design of this study. 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*. ACM, 3.

[2] Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. Towards improving trust in context-aware systems by displaying system confidence. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*. ACM, 9–14.

[3] Stavros Antifakos, Adrian Schwaninger, and Bernt Schiele. 2004. Evaluating the effects of displaying uncertainty in context-aware applications. In *International Conference on Ubiquitous Computing*. Springer, 54–69.

[4] 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 Seventh AAAI Conference on Human Computation and Crowdsourcing*.

[5] 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.

[6] James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. Citeseer, 35.

[7] Diego De Siqueira Braga, Marco Niemann, Bernd Hellingrath, and Fernando Buarque De Lima Neto. 2018. Survey on Computational Trust and Reputation Models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 101.

[8] David V Budescu and Adrian K Rantilla. 2000. Confidence in aggregation of expert opinions. *Acta psychologica* 104, 3 (2000), 371–398.

[9] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[10] 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*. ACM, 559.

[11] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. 120–129.

[12] Jin-Hee Cho, Kevin Chan, and Sibel Adali. 2015. A survey on trust modeling. *ACM Computing Surveys (CSUR)* 48, 2 (2015), 28.

[13] Geoff Cumming. 2013. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

[14] Geoff Cumming. 2014. The new statistics: Why and how. *Psychological science* 25, 1 (2014), 7–29.

[15] David Dearman, Alex Varshavsky, Eyal De Lara, and Khai N Truong. 2007. An exploration of location error estimation. In *International Conference on Ubiquitous Computing*. Springer, 181–198.

[16] Abhijit Dharia and Hojjat Adeli. 2003. Neural network model for rapid forecasting of freeway link travel time. *Engineering Applications of Artificial Intelligence* 16, 7-8 (2003), 607–613.

[17] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[18] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2016. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2016), 1155–1170.

[19] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[20] Pierre Dragicevic. 2016. Fair statistical communication in HCI. In *Modern statistical methods for HCI*. Springer, 291–330.

[21] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 144.

[22] 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.

[23] Yuan Gao, Sanmay Das, and Patrick Fowler. 2017. Homelessness service provision: a data science perspective. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

[24] Inga Großmann, André Hottung, and Artus Krohn-Grimberghe. 2019. Machine learning meets partner matching: Predicting the future relationship quality based on personality traits. *PloS one* 14, 3 (2019).

[25] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1321–1330.

[26] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[27] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.

[28] Samantha Joel, Paul W Eastwick, and Eli J Finkel. 2017. Is romantic desire predictable? Machine learning applied to initial romantic attraction. *Psychological science* 28, 10 (2017), 1478–1489.

[29] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5092–5103.

[30] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2390–2395.

[31] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 411.

[32] Amanda Kube, Sanmay Das, and Patrick J Fowler. 2019. Allocating interventions based on predicted outcomes: A case study on homelessness services. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[33] Volodymyr Kuleshov and Percy S Liang. 2015. Calibrated structured prediction. In *Advances in Neural Information Processing Systems*. 3474–3482.

[34] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 487.

[35] 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*. ACM, 29–38.

[36] Brian Y Lim and Anind K Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 415–424.

[37] Zachary C Lipton. 2018. The mythos of model interpretability. *Commun. ACM* 61, 10 (2018), 36–43.

[38] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[39] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[40] Xiaolei Ma, Zhuang Dai, Zhengbing He, Jihui Ma, Yong Wang, and Yunpeng Wang. 2017. Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. *Sensors* 17, 4 (2017), 818.

[41] Polina Marinova. 2017. How Dating Site eHarmony Uses Machine Learning to Help You Find Love. https://fortune.com/2017/02/14/eharmony-dating-machine-learning/

[42] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 427–436.

[43] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 625–632.

[44] Ellen Peters, Daniel Västfjäll, Paul Slovic, CK Mertz, Ketti Mazzocco, and Stephan Dickert. 2006. Numeracy and decision making. *Psychological science* 17, 5 (2006), 407–413.

[45] Robert Pinsker. 2011. Primacy or recency? A study of order effects when nonprofessional investors are provided a long series of disclosures. *Behavioral Research in Accounting* 23, 1 (2011), 161–183.

[46] John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.

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

[48] Valerie F Reyna and Charles J Brainerd. 2008. Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and individual differences* 18, 1 (2008), 89–107.

[49] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations.. In *IUI*.

240–251.

[50] Aaron Springer and Steve Whittaker. 2018. Progressive Disclosure: Designing for Effective Transparency. *arXiv preprint arXiv:1811.02164* (2018).

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

[52] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 272–283.

[53] Brandon M Turner, Mark Steyvers, Edgar C Merkle, David V Budescu, and Thomas S Wallsten. 2014. Forecast aggregation via recalibration. *Machine learning* 95, 3 (2014), 261–289.

[54] Jhim Kiel M Verame, Enrico Costanza, and Sarvapali D Ramchurn. 2016. The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study. In *Proceedings of the 2016 chi conference on human factors in computing systems*. ACM, 4908–4920.

[55] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[56] Robert L Winkler, Yael Grushka-Cockayne, Kenneth C Lichtendahl, and Victor Richmond R Jose. 2018. *Averaging Probability Forecasts: Back to the Future*. Harvard Business School.

[57] 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.

[58] 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*. ACM, 279.

[59] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, 307–317.

[60] Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 694–699.

[61] 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).

[62] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2979–2989.