Understanding the Effect of Accuracy on Trust in Machine Learning Models

Ming Yin
Purdue University
mingyin@purdue.edu

Jennifer Wortman Vaughan
Microsoft Research
jenn@microsoft.com

Hanna Wallach
Microsoft Research
hanna@dirichlet.net

ABSTRACT

We address a relatively under-explored aspect of human–computer interaction: people's abilities to understand the relationship between a machine learning model's stated performance on held-out data and its expected performance post deployment. We conduct large-scale, randomized human-subject experiments to examine whether laypeople's trust in a model, measured in terms of both the frequency with which they revise their predictions to match those of the model and their self-reported levels of trust in the model, varies depending on the model's stated accuracy on held-out data and on its observed accuracy in practice. We find that people's trust in a model is affected by both its stated accuracy and its observed accuracy, and that the effect of stated accuracy can change depending on the observed accuracy. Our work relates to recent research on interpretable machine learning, but moves beyond the typical focus on model internals, exploring a different component of the machine learning pipeline.

CCS CONCEPTS

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

KEYWORDS

Machine learning, trust, human-subject experiments

ACM Reference Format:
Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300509

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300509

1 INTRODUCTION

Machine learning (ML) is becoming increasingly ubiquitous as a tool to aid human decision-making in diverse domains ranging from medicine to public policy and law. For example, researchers have trained deep neural networks to help dermatologists identify skin cancer [8], while political strategists regularly use ML-based forecasts to determine their next move [21]. Police departments have used ML systems to predict the location of human trafficking hotspots [28], while child welfare workers have used predictive modeling to strategically target services to the children most at risk [3].

This widespread applicability of ML has led to a movement to "democratize machine learning" [12] by developing off-the-shelf models and toolkits that make it possible for anyone to incorporate ML into their own system or decision-making pipeline, without the need for any formal training. While this movement opens up endless possibilities for ML to have real-world impact, it also creates new challenges. Decision-makers may not be used to reasoning about the explicit forms of uncertainty that are baked into ML predictions [27], or, because they do not need to understand the inner workings of an ML model in order to use it, they may misunderstand or mistrust its predictions [6, 16, 25]. Prompted by these challenges, as well as growing concerns that ML systems may inadvertently reinforce or amplify societal biases [1, 2], researchers have turned their attention to the ways that humans interact with ML, typically focusing on people's abilities and willingness to use, understand, and trust ML systems. This body of work often falls under the broad umbrella of interpretable machine learning [6, 16, 25].

To date, most work on interpretability has focused explicitly on ML models, asking questions about people's abilities to understand model internals or the ways that particular models map inputs to outputs [20, 24], as well as questions about the relationship between these abilities and people's willingness to trust a model. However, the model is just one component of the ML pipeline, which spans data collection, model selection, training algorithms and procedures, model evaluation, and ultimately, deployment. It is therefore important to study people's interactions with each of these components—not just those that relate to model internals.

One particularly under-explored aspect of the evaluation and deployment components of the pipeline is the interpretability of performance metrics, such as accuracy, precision, or recall. The democratization of ML means that it is increasingly common for a decision-maker to be presented with a "black-box" model along with some measure of its performance—most often accuracy—on held-out data. However, a model's stated performance may not accurately reflect its performance post deployment because the data on which the model was trained and evaluated may look very different from real-world use cases [15]. In deciding how much to trust the model, the decision-maker has little to go on besides this stated performance, her own limited observations of the model's predictions in practice, and her domain knowledge.

This scenario raises a number of questions. To what extent do laypeople—who are increasingly often the end users of systems built using ML models—understand the relationship between a model's stated performance on held-out data and its expected performance post deployment? How does their understanding influence their willingness to trust the model? For example, do people trust a model more if they are told that its accuracy on held-out data is 90% as compared with 70%? If so, will the model's stated accuracy continue to influence their trust in the model even after they are given the opportunity to observe and interact with the model in practice?

In this paper, we describe the results of a sequence of large-scale, randomized, pre-registered human-subject experiments[1] designed to investigate whether an ML model's accuracy affects laypeople's willingness to trust the model. Specifically, we focus on the following three main questions:

- Does a model's stated accuracy on held-out data affect people's trust in the model?
- If so, does it continue to do so after people have observed the model's accuracy in practice?
- How does a model's observed accuracy in practice affect people's trust in the model?

In each of our experiments, subjects recruited on Amazon Mechanical Turk were asked to make predictions about the outcomes of speed dating events with the help of an ML model. Subjects were first shown information about a speed dating participant and his or her date, and then asked to predict whether or not the participant would want to see his or her date again. Finally, they were shown the model's prediction and given the option of revising their own prediction.

In our first experiment, we focus on the first two questions above, investigating whether a model's stated accuracy on held-out data affects laypeople's trust in the model and, if so, whether it continues to do so after they have observed the model's accuracy in practice. Subjects were randomized into one of ten treatments, which differed along two dimensions: stated accuracy on held-out data and amount at stake. Some subjects were given no information about the model's accuracy on held-out data, while others were told that its accuracy was 60%, 70%, 90%, or 95%. Halfway through the experiment, each subject was given feedback on both their own accuracy and the model's accuracy on the first half of the prediction tasks, which was 80% regardless of the treatment. Subjects in all treatments saw exactly the same speed dating events and exactly the same model predictions. This experimental design allows us isolate the effect of stated accuracy on people's trust, both before and after they observe the model's accuracy in practice. As a robustness check, some subjects received a monetary bonus for each correct prediction, while others did not, allowing us to test whether the effect of stated accuracy on trust varies when people have more "skin in the game."

We find that stated accuracy does have a significant effect on people's trust in a model, measured in terms of both the frequency with which subjects adjust their predictions to match those of the model and their self-reported levels of trust in the model. We also find that the effect size is smaller after people observe the model's accuracy in practice. We do not find that the amount at stake has a significant effect.

In our second experiment, we test whether these results are robust to different levels of observed accuracy by running two additional variations of our first experiment: one in which the observed accuracy of the model was low and one in which the observed accuracy of the model was high. We find that a model's stated accuracy still has a significant effect on people's trust even after observing a high accuracy (100%) in practice. However, if a model's observed accuracy is low (55%), then after observing this accuracy, the stated accuracy has at most a very small effect on people's trust in the model.

In our third experiment, we investigate the final question above—i.e., how does a model's observed accuracy in practice affect people's trust in the model? The experimental design used in our first two experiments does not enable us to directly compare people's trust between treatments with different levels of observed accuracy because the prediction tasks (i.e., speed dating events) and the model predictions differed between these treatments. Our third experiment was therefore carefully designed to enable us to make such comparisons. We find that after observing a model's accuracy in practice, people's trust in the model is significantly affected by its observed accuracy regardless of its stated accuracy.

Finally, via an exploratory analysis, we dig more deeply into the question of how people update their trust after receiving feedback on their own accuracy and the model's accuracy in practice. We analyze differences in individual subjects' trust in the model before and after receiving such feedback. Our experimental data support the conjecture that people compare their own accuracy to the model's observed accuracy, increasing their trust in the model if the model's

[1] All experiments were approved by the Microsoft Research IRB.

observed accuracy is higher than their own accuracy—except in the case where the model's observed accuracy is substantially lower than its stated accuracy on held-out data.

Taken together, our results show that laypeople's trust in an ML model is affected by both the model's stated accuracy on held-out data and its observed accuracy in practice. These results highlight the need for designers of ML systems to clearly and responsibly communicate their expectations about model performance, as this information shapes the extent to which people trust a model, both before and after they are able to observe and interact with it in practice. Our results also reveal the importance of properly communicating the uncertainty that is baked into every ML prediction. Of course, proper caution should be used when generalizing our results to other settings. For example, although we do not find that the amount at stake has a significant effect, it is possible that there would be an effect when stakes are sufficiently high (e.g., doctors making life-or-death decisions).

Related Work

Our research contributes to a growing body of experimental work on trust in algorithmic systems. As a few examples, Dzindolet et al. [7] and Dietvorst et al. [4] found that people stop trusting an algorithm after witnessing it make a mistake, even when the algorithm outperforms human predictions—a phenomenon known as algorithm aversion. Dietvorst et al. [5] found that people are more willing to rely on an algorithm's predictions when they are given the ability to make minor adjustments to the predictions rather than accepting them as is. Yeomans et al. [30] found that people distrust automated recommender systems compared with human recommendations in the context of predicting which jokes people will find funny—a highly subjective domain—even when the recommender system outperforms human predictions. In contrast, Logg et al. [17] found that people trust predictions more when they believe that the predictions come from an algorithm as opposed to a human expert when predicting music popularity, romantic matches, and other outcomes. This effect is diluted when people are given the choice between using an algorithm's prediction and using their own prediction (as opposed to a prediction from another human expert).

The relationship between interpretability and trust has been discussed in several recent papers [16, 22, 25]. Most related to our work, and an inspiration for our experimental design, Poursabzi-Sangdeh et al. [24] ran a sequence of randomized human-subject experiments and found no evidence that either the number of features used in an ML model nor the model's level of transparency (clear or black box) have a significant impact on people's willingness to trust the model's predictions, although these factors do affect people's abilities to detect when the model has made a mistake.

Kennedy et al. [14] touched on the relationship between stated accuracy and trust in the context of criminal recidivism prediction. They ran a conjoint experiment in which they presented subjects with randomly generated pairs of models and asked each subject which model they preferred. The models varied in terms of their stated accuracy, the size of the (fictitious) training data set, the number of features, and several other properties. The authors estimated the effect of each property by fitting a hierarchical linear model and found that people generally focus most on the size of the training data set, the source of the algorithm, and the stated accuracy, while less often taking into account the model's level of transparency or the relevance of the training data.

Finally, a few studies from the human–computer interaction community have examined the relationship between system performance and users' trust in automated systems [31, 32], ubiquitous computing systems [13], recommender systems [23], and robots [26]. For example, in a simulated experimental environment in which users interacted with an automated quality monitoring system to identify faulty items in a fictional factory production line, Yu et al. [31, 32] explored how users' trust in the system varies with its accuracy. Unlike in our work, system accuracy was not explicitly communicated to users. Instead, users "perceived" the accuracy by receiving feedback after interacting with the system. Yu et al. found that users are able to correctly perceive the accuracy and stabilize their trust to a level correlated with the accuracy [31], though system failures have a stronger impact on trust than system successes [32]. In addition, Kay et al. [13] developed a survey tool through which they revealed that, for classifiers used in four hypothetical applications (e.g., electricity monitoring and location tracking), users tend to put more weight on the classifiers' recall rather than their precision when deciding whether the classifiers' performance is acceptable, with the weight varying across applications.

2 EXPERIMENT 1: DOES A MODEL'S STATED ACCURACY AFFECT LAYPEOPLE'S TRUST?

Our first experiment was designed to answer our first two main questions—i.e., does a model's stated accuracy on held-out data affect laypeople's trust in the model, and if so, does it continue to do so after they have observed the model's accuracy in practice? In our experiment, each subject observed the model's accuracy in practice via a feedback screen that was presented halfway through the experiment with information about the subject's own accuracy and the model's accuracy thus far, as described below. Before running the experiment, we posited and pre-registered two hypotheses derived from our questions, which we state informally here:[2]

[2] The pre-registration document is at https://aspredicted.org/uq3hi.pdf.

- **[H1]** The stated accuracy of a model has a significant effect on people's trust in the model *before* seeing the feedback screen.
- **[H2]** The stated accuracy of a model has a significant effect on people's trust in the model *after* seeing the feedback screen.

As a robustness check to guard against the potential criticism that any null results might be due to a lack of performance incentives, we randomly selected some subjects to receive a monetary bonus for each correct prediction. We also posited and pre-registered two additional hypotheses:

- **[H3]** The amount at stake has a significant effect on people's trust in a model *before* seeing the feedback screen.
- **[H4]** The amount at stake has a significant effect on people's trust in a model *after* seeing the feedback screen.

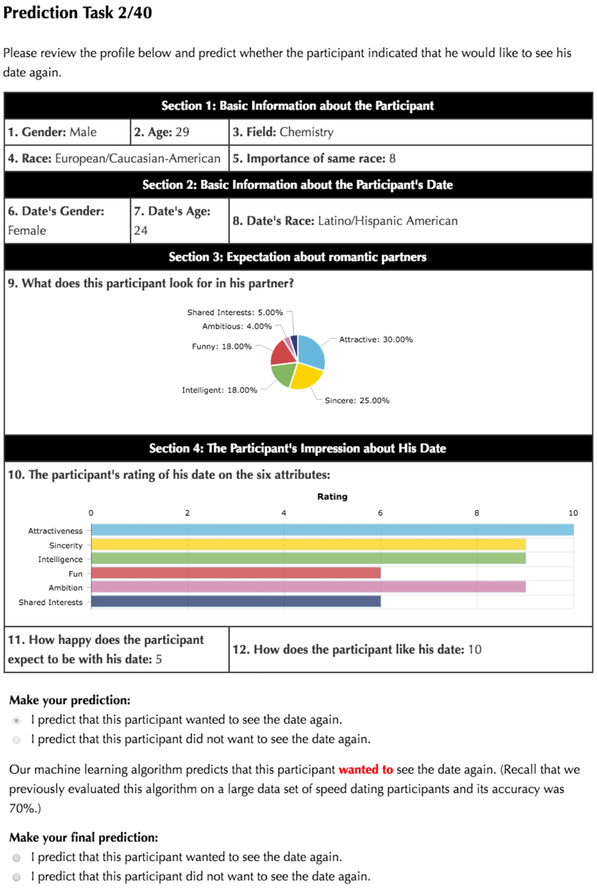
Prediction Tasks

We asked subjects to make predictions about the outcomes of forty speed dating events. The data came from real speed dating participants and their dates via the experimental study of Fisman et al. [9]. Each speed dating participant indicated whether or not he or she wanted to see his or her date again, thereby giving us ground truth from which to compute accuracy. We chose this application for two reasons: First, predicting romantic interest does not require specialized domain expertise. Second, this setting is plausibly one in which ML might be used given that many dating websites already rely on ML models to predict potential romantic partners [18, 29].

For each prediction task (i.e., speed dating event), each subject was first shown a screen of information about the speed dating participant and his or her date, including:

- *The participant's basic information*: the gender, age, field of study, race, etc. of the participant.
- *The date's basic information*: the gender, age, and race of the participant's date.
- *The participant's preferences*: the participant's reported distribution of 100 points among six attributes (attractiveness, sincerity, intelligence, fun, ambition, and shared interests), indicating how much he or she values each attribute in a romantic partner.
- *The participant's impression of the date*: the participant's rating of his or her date on the same six attributes using a scale of one to ten, as well as scores (also using a scale of one to ten) indicating how happy the participant expected to be with his or her date and how much the participant liked his or her date.

The subject was then asked to follow a three step-procedure: First, they were asked to carefully review the information about the participant and his or her date and predict whether or not the participant would want to see his or her date again. Next, they were shown the model's (binary) prediction. Finally, they were given the option of revising their own prediction. A screenshot of the interface is shown in Figure 1.



Prediction Task 2/40

Please review the profile below and predict whether the participant indicated that he would like to see his date again.

Section 1: Basic Information about the Participant		
1. Gender: Male	**2. Age:** 29	**3. Field:** Chemistry
4. Race: European/Caucasian-American		**5. Importance of same race:** 8
Section 2: Basic Information about the Participant's Date		
6. Date's Gender: Female	**7. Date's Age:** 24	**8. Date's Race:** Latino/Hispanic American
Section 3: Expectation about romantic partners		
9. What does this participant look for in his partner?		
Section 4: The Participant's Impression about His Date		
10. The participant's rating of his date on the six attributes:		
11. How happy does the participant expect to be with his date: 5		**12. How does the participant like his date:** 10

Make your prediction:

● I predict that this participant wanted to see the date again.
○ I predict that this participant did not want to see the date again.

Our machine learning algorithm predicts that this participant **wanted to** see the date again. (Recall that we previously evaluated this algorithm on a large data set of speed dating participants and its accuracy was 70%.)

Make your final prediction:

○ I predict that this participant wanted to see the date again.
○ I predict that this participant did not want to see the date again.

Figure 1: Screenshot of the prediction task interface.

Experimental Treatments

We randomized subjects into one of ten treatments arranged in a 5 × 2 design. The treatments differed along two dimensions: stated accuracy on held-out data and amount at stake.

Subjects were randomly assigned to one of five accuracy levels: none (the baseline), 60%, 70%, 90%, or 95%. Subjects assigned to an accuracy level of none were initially given no information about the model's accuracy on held-out data. Subjects assigned to one of the other accuracy levels saw the following sentence in the instructions: "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." Throughout the experiment, we also reminded these subjects of the model's stated accuracy on held-out data each time they were shown one of the model's predictions.

We note that our sentence about accuracy was *not* a deception. We developed four ML models (a rule-based classifier, a support vector machine, a three-hidden-layer neural network, and a random forest) and evaluated them on a held-out data set of 500 speed dating participants, obtaining accuracies of 60%, 70%, 90%, and 95%. To keep the treatments as similar as possible, the models made exactly the same predictions for the forty speed dating events that were shown to subjects.

Subjects were randomly assigned to either low or high stakes. Subjects assigned to low stakes were paid a flat rate of $1.50 for completing the experiment. Subjects assigned to high stakes also received a monetary bonus of $0.10 for each correct (final) prediction[3] in addition to the flat rate of $1.50.

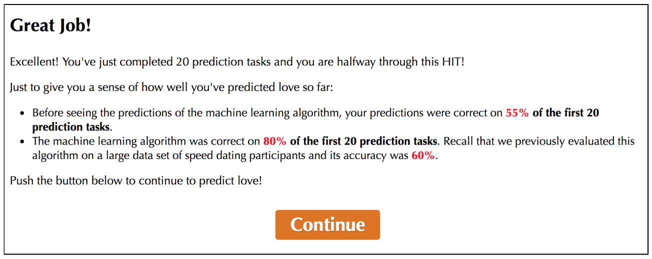
Experimental Design

We posted our experiment as a human intelligence task (HIT) on Amazon Mechanical Turk. The experiment was only open to workers in the U.S., and each worker could participate only once. In total, 1,994 subjects completed the experiment.

Upon accepting the HIT, each subject was randomized into one of the ten treatments described above. Each HIT consisted of exactly the same forty prediction tasks, grouped into two sets A and B of twenty tasks each. As described above, subjects in all ten treatments saw exactly the same model prediction for each task. The experiment was divided into two phases. To minimize differences between the phases, subjects were randomly assigned to see either the tasks in set A during Phase 1 and the tasks in set B during Phase 2, or vice versa; the order of the tasks was randomized within each phase. We chose the tasks in sets A and B so that the observed accuracy on the first twenty tasks would be 80% regardless of the ordering of sets A and B. This experimental design minimizes differences between treatments and allows us to draw causal conclusions about the effect of stated accuracy on people's trust without worrying about confounding factors.

Each subject was asked to make initial and final predictions for each task, following the three-step procedure described above. The subjects were given no feedback on their own prediction or the model's prediction for any individual task; however, after Phase 1, each subject was shown a feedback screen with information about their own accuracy and the model's accuracy (80% by design) on the tasks in Phase 1. A screenshot of the feedback screen is shown in Figure 2.

At the end of the HIT, each subject completed an exit survey in which they were asked to report their level of trust in the model during each phase using a scale of one ("I didn't trust it at all") to ten ("I fully trust it"). Specifically, we asked subjects the following question: "How much did you trust our machine learning algorithm's predictions on the first [last] twenty speed dating participants (that is, before [after] you saw any feedback on your performance and the algorithm's performance)?" We also collected basic demographic information (such as age and gender) about each subject.


Great Job!

Excellent! You've just completed 20 prediction tasks and you are halfway through this HIT!

Just to give you a sense of how well you've predicted love so far:

- Before seeing the predictions of the machine learning algorithm, your predictions were correct on **55% of the first 20 prediction tasks**.
- The machine learning algorithm was correct on **80% of the first 20 prediction tasks**. Recall that we previously evaluated this algorithm on a large data set of speed dating participants and its accuracy was **60%**.

Push the button below to continue to predict love!

Figure 2: Screenshot of the feedback screen shown between Phase 1 and Phase 2 (i.e., after the first twenty tasks).

To quantify a subject's trust in a model, we defined two metrics, calculated separately for each phase, that capture how often the subject "followed" the model's predictions:

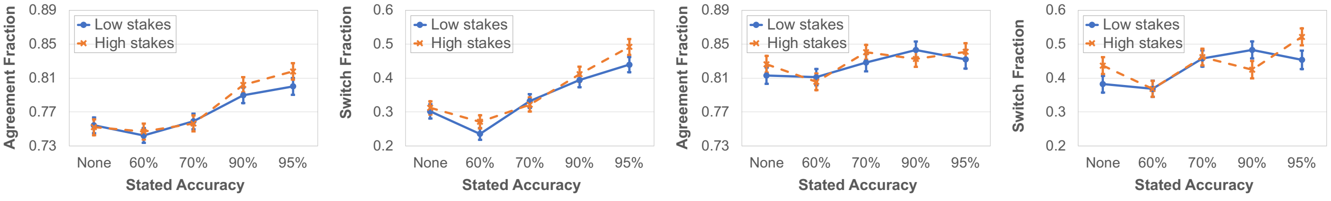
- *Agreement fraction*: the number of tasks for which the subject's final prediction agreed with the model's prediction, divided by the total number of tasks.
- *Switch fraction*: the number of tasks for which the subject's initial prediction disagreed with the model's prediction *and* the subject's final prediction agreed with the model's prediction, divided by the total number of tasks for which the subject's initial prediction disagreed with the model's prediction.

We used these two metrics when formally stating all of our pre-registered hypotheses, while additionally pre-registering our intent to analyze subjects' self-reported trust levels.

Analysis of Trust in Phase 1 (H1 and H3)

We start by analyzing data from Phase 1 to see if subjects' trust in a model is affected by the model's stated accuracy and the amount at stake *before* they see the feedback screen. Figures 3a and 3b show subjects' average agreement fraction and average switch fraction, respectively, in Phase 1, by treatment. Visually, stated accuracy appears to have a substantial effect on how often subjects follow the model's predictions. Subjects' final predictions agree with the model's predictions more often when the model has a high stated accuracy. However, the effect of the amount at stake is less apparent. To formally compare the treatments, we conduct a two-way ANOVA on subjects' agreement fractions and, respectively, switch fractions in Phase 1. The results suggest a statistically significant main effect of stated accuracy on how often subjects follow the model's predictions (effect size $\eta^2 = 0.036$, $p = 4.72 \times 10^{-15}$ for agreement fraction, and $\eta^2 = 0.061$, $p = 5.62 \times 10^{-26}$ for switch fraction) while the main effect of the amount at stake is insignificant ($p = 0.30$ and $p = 0.11$ for agreement fraction and switch fraction, respectively). We do not detect a significant interaction between the two

[3] The highest possible bonus was $40 \times \$0.10 = \$4$—i.e., substantially more than the flat rate of $1.50, thereby making the bonus salient [11].



(a) Phase 1: Agreement frac. **(b) Phase 1: Switch fraction** **(c) Phase 2: Agreement frac.** **(d) Phase 2: Switch fraction**

Figure 3: Comparing how often subjects in different experimental treatments follow an ML model's predictions (average agreement fraction and average switch fraction) during each phase of our first experiment. Error bars represent standard errors.

factors ($p = 0.77$ and $p = 0.62$ for agreement fraction and switch fraction, respectively). In other words, hypothesis H1 is supported by our experimental data, while H3 is not.

An analysis of subjects' self-reported levels of trust reveals a similar pattern. We detect a statistically significant main effect of stated accuracy on subjects' self-reported levels of trust during Phase 1 ($\eta^2 = 0.049$, $p = 1.61 \times 10^{-20}$), while the main effect of the amount at stake is insignificant ($p = 0.92$).

We also conduct a post-hoc Tukey's HSD test to identify pairs of treatments in which subjects exhibit distinct differences in how often they follow the model's predictions. We find that treatments can be clustered into two groups—treatments with an accuracy level of none, 60%, or 70%, and treatments with an accuracy level of 90% or 95%—such that almost all statistically significant results are found for across-group treatment pairs.[4] These results confirm our visual intuition from Figures 3a and 3b: when subjects have not yet observed the model's accuracy in practice, they tend to follow the predictions of models with a high stated accuracy more often than those of models with a low (or no) stated accuracy, even though the models make exactly the same predictions.

Analysis of Trust in Phase 2 (H2 and H4)

We next analyze data from Phase 2 to examine whether subjects' trust in a model is affected by the model's stated accuracy and the amount at stake *after* they see the feedback screen. The feedback screen included information about the subject's own accuracy and the model's accuracy (80% by design) on the tasks in Phase 1, as described above. Figures 3c and 3d show subjects' average agreement fraction and average switch fraction, respectively, in Phase 2, by treatment. Through a two-way ANOVA, we again find a statistically significant main effect of stated accuracy on how often subjects follow the model's predictions during Phase 2 ($p = 0.009$ for agreement fraction and $p = 1.13 \times 10^{-5}$ for switch fraction), but the effect sizes are smaller than those for Phase 1 ($\eta^2 = 0.007$ for agreement fraction and $\eta^2 = 0.014$ for switch fraction). Again, neither the main effect of the amount at stake nor the interaction between the two factors are statistically significant. That is, after receiving feedback on the model's accuracy in practice, stated accuracy still has a substantial effect on how often subjects follow the model's predictions, while the amount at stake does not. In other words, hypothesis H2 is supported by our experimental data, while H4 is not.

As before, an analysis of subjects' self-reported levels of trust tells a similar story. We detect a statistically significant main effect of stated accuracy on subjects' self-reported levels of trust during Phase 2 ($\eta^2 = 0.008$, $p = 0.005$), while the main effect of the amount at stake is insignificant ($p = 0.88$).

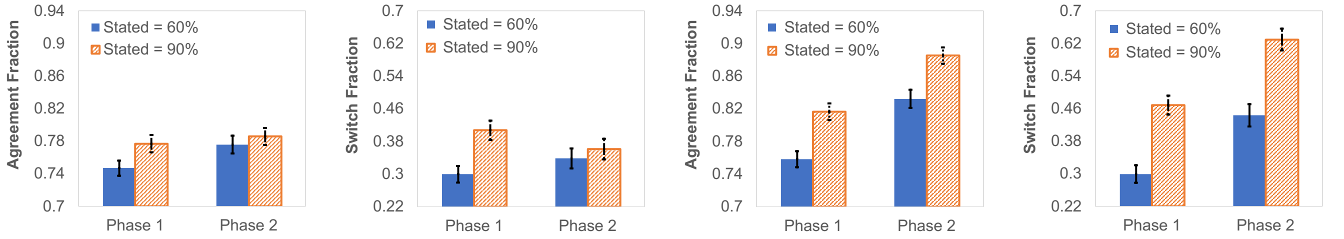
Interestingly, when conducting a post-hoc Tukey's HSD test to identify pairs of treatments in which subjects exhibit distinct differences in how often they follow the model's predictions, we find fewer statistically significant results than we did for Phase 1. For agreement fraction, *none* of the differences across treatment pairs are statistically significant. For switch fraction, we find only five pairs of treatments with statistically significant differences; four of these are between treatments with an accuracy level of 60% and treatments with an accuracy level of 90% or 95%. Unlike in Phase 1, how often subjects follow the predictions of a model with a stated accuracy of 70% is *not* significantly different from how often subjects follow the predictions of a model with a stated accuracy of 90% or 95%. In other words, in Phase 2, a larger difference between two models' stated accuracies is needed to detect a statistically significant effect of stated accuracy on trust, as compared with the difference needed in Phase 1.

Together, the results from our first experiment suggest that a model's stated accuracy affects laypeople's trust in the model; however, this effect is diluted once they have observed the model's accuracy in practice. In contrast, the amount at stake does not have an effect on laypeople's trust in a model, at least for the limited range of stakes used in our experiment.

3 EXPERIMENT 2: DOES THIS EFFECT CHANGE IF THE OBSERVED ACCURACY IS LOW/HIGH?

In our first experiment, we found that a model's stated accuracy on held-out data has a significant effect on laypeople's trust in the model. However, in order to make valid comparisons across treatments, we had to design the experiment such that the models made exactly the same predictions for

[4] Among thirty-nine statistically significant results, there are two exceptions: switch fraction Low-60% vs. Low-70% and Low-90% vs High-95%.



(a) Agreement fraction with an observed accuracy of 55% **(b) Switch fraction with an observed accuracy of 55%** **(c) Agreement fraction with an observed accuracy of 100%** **(d) Switch fraction with an observed accuracy of 100%**

Figure 4: Examining the effect of stated accuracy on trust (average agreement fraction and average switch fraction) for different levels of observed accuracy (low vs. high) during each phase of our second experiment. Error bars represent standard errors.

the forty speed dating events that were shown to subjects. We therefore fixed the observed accuracy to be a single arbitrary value—specifically 80%—for all models. It is natural to ask whether we would obtain the same results if a model's observed accuracy were substantially lower or higher than 80%.

Our second experiment consists of two sub-experiments designed to test whether our results are robust to different levels of observed accuracy. In one, the observed accuracy was fixed to a low value (55%), while in the other it was fixed to a high value (100%). Each sub-experiment was a miniature variation of our first experiment with only two treatments: stated accuracy of 60% and stated accuracy of 90%. We developed two ML models, with accuracies of 60% and 90% on held-out data. Before running the experiment, we pre-registered two hypotheses, analogous to H1 and H2:[5]

- **[H5]** The stated accuracy of a model has a significant effect on people's trust in the model *before* seeing the feedback screen, *regardless of its observed accuracy.*
- **[H6]** The stated accuracy of a model has a significant effect on people's trust in the model *after* seeing the feedback screen, *regardless of its observed accuracy.*

Because our first experiment revealed that the amount at stake does not affect people's trust in a model, we did not select any subjects to receive monetary bonuses, nor did we pre-register any hypotheses about the amount at stake.

Experimental Design

We posted our experiment on Amazon Mechanical Turk, with a flat rate of $1.50 for completing the experiment. Workers who participated in our first experiment were prohibited from participating. We obtained data from 757 subjects.

Each subject was first randomized into one of the two sub-experiments and then into one of the two treatments described above. Each sub-experiment was completely analogous to our first experiment. Again, each HIT consisted of forty prediction tasks, grouped into two sets A and B of twenty tasks each. Subjects were randomly assigned to see either the tasks in set A during Phase 1 and the tasks in set B during Phase 2, or vice versa; the order of the tasks was randomized within each phase. For each sub-experiment, we chose the tasks in sets A and B so that subjects in both treatments would see exactly the same model prediction for each task and so that the observed accuracy on the first twenty tasks would be 55% (for the first sub-experiment) or 100% (for the second sub-experiment) regardless of the ordering of sets A and B. This experimental design minimizes differences between treatments *within* each sub-experiment and allows us to draw causal conclusions about the effect of stated accuracy on people's trust when the observed accuracy is low (alternatively, high) by comparing treatments within the first (alternatively, second) sub-experiment. However, because the two sub-experiments consisted of different tasks, we are not able to compare treatments from different sub-experiments.

Experimental Results

We analyze data from Phase 1 of each sub-experiment separately to see if subjects' trust in a model is affected by the model's stated accuracy *before* they see the feedback screen, regardless of its observed accuracy. The left two bars in Figures 4a and 4b show subjects' average agreement fraction and average switch fraction, respectively, in Phase 1 when the observed accuracy was 55%, by treatment. Using two-sample t-tests to compare different treatments, we find a statistically significant effect of stated accuracy on how often subjects follow the model's predictions ($p = 0.033$ for agreement fraction, and $p = 6.33 \times 10^{-4}$ for switch fraction). Similar, and stronger, results can be seen from the left two bars in Figures 4c and 4d, which correspond to the second sub-experiment. Again, we find a statistically significant effect of stated accuracy on how often subjects follow the model's predictions ($p = 4.74 \times 10^{-5}$ for agreement fraction, and $p = 1.71 \times 10^{-7}$ for switch fraction). Together, these results show that hypothesis H5 is supported by our experimental data.

We next analyze data from Phase 2 of each sub-experiment separately to see whether subjects' trust in a model is affected by the model's stated accuracy *after* they see the feedback

[5] The pre-registration document is at https://aspredicted.org/w9t8g.pdf.

screen, regardless of its observed accuracy. The right two bars in Figures 4a–4d show subjects' average agreement fraction and average switch fraction in Phase 2 for both sub-experiments. When the observed accuracy is high (100%), consistent with the results from our first experiment, we find a statistically significant effect of stated accuracy on how often subjects follow the model's predictions after receiving feedback on the model's accuracy in practice ($p = 4.33 \times 10^{-4}$ for agreement fraction, and $p = 1.61 \times 10^{-6}$ for switch fraction). However, this is not the case when the observed accuracy is low (55%). Specifically, the right two bars in Figures 4a and 4b indicate that, after receiving feedback on the model's accuracy in practice, the effect of stated accuracy on how often subjects follow the model's predictions is not significant ($p = 0.506$ for agreement fraction, and $p = 0.515$ for switch fraction) when the observed accuracy is low (55%). Therefore, hypothesis H6 is not supported by our experimental data.

As before, an analysis of subjects' self-reported levels of trust tells a similar story. We detect statistically significant differences in self-reported levels of trust in Phase 1 between between treatments within each sub-experiment, while in Phase 2, we only detect statistically significant differences between treatments when the observed accuracy is high (100%).

One possible explanation for the lack of support for hypothesis H6 is that the difference between the models' stated accuracies (60% and 90%) is simply not large enough to detect a statistically significant effect of stated accuracy on trust in Phase 2 when the models' observed accuracies are low, at least with our experiment's sample size. Indeed, as we show in our third experiment in the next section, if we keep the higher accuracy level at 90% but decrease the lower accuracy level from 60% to 50%, then we again find a statistically significant effect, even when the observed accuracy is low (55%). Combining these results with the results from our first experiment—i.e., when the observed accuracy is 80%, increasing the stated accuracy from 60% to 90% has a statistically significant effect on subjects' trust during both phases—we conjecture that the effect of stated accuracy on people's trust *before* they observe the model's accuracy in practice is not too dependent on observed accuracy, but the effect of stated accuracy on people's trust *after* they observe the model's accuracy in practice *is* dependent on observed accuracy.

4 EXPERIMENT 3: DOES A MODEL'S OBSERVED ACCURACY AFFECT LAYPEOPLE'S TRUST?

In our first and second experiments, we found that a model's stated accuracy on held-out data has a significant effect on laypeople's trust in the model, though this effect is diluted once they have observed the model's accuracy in practice and can even disappear if the observed accuracy is sufficiently low. Our third and final experiment was designed to answer our third main question—i.e., how does a model's observed accuracy in practice affect laypeople's trust in the model?

Answering this question required us to modify our experimental design because the design used in our first two experiments does not enable us to directly compare people's trust between treatments with different levels of observed accuracy. In our third experiment, we therefore randomized subjects into one of six treatments arranged in a 2×3 design. The treatments varied along two dimensions: stated accuracy on held-out data (50% or 90%) and observed accuracy *on the tasks in Phase 1 only* (55%, 80%, or 100%). Before running the experiment, we pre-registered the following hypothesis:[6]

- **[H7]** After seeing the feedback screen, the observed accuracy of a model has a significant effect on people's trust in the model, *regardless of its stated accuracy.*

Experimental Design

As with our first two experiments, we posted the experiment on Amazon Mechanical Turk with a flat rate of $1.50 for completing the experiment. The experiment was only open to U.S. workers who had not participated in our previous experiments. In total, we obtained data from 1,042 subjects.

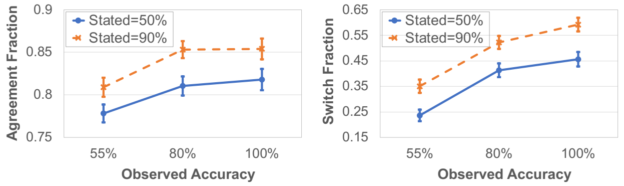
Each subject was randomized into one of the six treatments described above. Each HIT consisted of exactly the same forty prediction tasks. However, models with different observed accuracies made different predictions for the tasks in Phase 1. To allow us to compare treatments with different observed accuracies, subjects in all six treatments saw exactly the same model prediction for each task in Phase 2, corresponding to an observed accuracy of 80%. Because the models' observed accuracies could differ between the phases, we did not randomize the set of tasks used in each phase, but did randomize the order of the tasks within each phase.

Experimental Results

Figures 5a and 5b show subjects' average agreement fraction and average switch fraction, respectively, in Phase 2, by treatment. Visually, observed accuracy appears to have a substantial effect on how often subjects follow a model's predictions, regardless of the stated accuracy of the model.

To formally compare the treatments, we conduct one-way ANOVAs on subjects' agreement fractions and switch fractions in Phase 2 across treatments with the same stated accuracy but different observed accuracies. When the stated accuracy is 50%, we find that there is a statistically significant effect of observed accuracy on how often subjects follow the model's predictions ($p = 0.032$ for agreement fraction, and $p = 3.4 \times 10^{-9}$ for switch fraction). We similarly find a statistically significant effect of observed accuracy on trust ($p = 0.005$ for agreement fraction, and $p = 3.57 \times 10^{-10}$ for

[6] The pre-registration document is at https://aspredicted.org/7yf66.pdf.



(a) Agreement fraction **(b) Switch fraction**

Figure 5: Examining the effect of observed accuracy on trust (average agreement fraction and average switch fraction) for different levels of stated accuracy during Phase 2 of our third experiment. Error bars represent standard errors.

switch fraction) when the stated accuracy is 90%. In other words, hypothesis H7 is supported by our experimental data.

As before, analyzing subjects' self-reported levels of trust tells a similar story. We find a statistically significant effect of observed accuracy on subjects' self-reported levels of trust in Phase 2 when the stated accuracy is 50% ($p = 1.58 \times 10^{-6}$) and when the stated accuracy is 90% ($p = 5.24 \times 10^{-14}$).

Given the results from our first two experiments—and, in particular, the lack of support for hypothesis H6—it is natural to ask about the extent to which a model's stated accuracy affects subjects' trust in the model during Phase 2 of this experiment. If we compare treatments with the same observed accuracy on the tasks in Phase 1 but different stated accuracies, we find that stated accuracy has a substantial effect on trust, regardless of the observed accuracy. As discussed in the previous section, we suspect that this apparent discrepancy with the results from our second experiment when the observed accuracy is low (55%) is due to the fact that the difference between the models' stated accuracies (50% and 90%) is larger in this experiment than in our second experiment. Indeed, it is possible that that for our range of sample sizes, there could exist a statistically significant difference between treatments with stated accuracies of 50% and 90% without there being any statistically significant difference between treatments with stated accuracies of 50% and 60% or between treatments with stated accuracies of 60% and 90%. Regardless, our results imply that after receiving feedback on a model's accuracy in practice, the model's stated accuracy *and* its observed accuracy affect people's trust in the model.

5 EXPLORATORY ANALYSIS: HOW DOES TRUST CHANGE AFTER RECEIVING FEEDBACK?

In our three experiments, we found that both a model's stated accuracy on held-out data and its observed accuracy in practice affect laypeople's trust in the model. In this section, we report the results of an exploratory analysis of the data from our first two experiments, aimed at digging more deeply into the mechanisms behind this finding. We analyze differences in individual subjects' trust between Phase 1 and Phase 2, and consider three possible mechanisms to explain them.

We posit three possible mechanisms by which subjects might update their trust after seeing the feedback screen:

- *Stated vs. observed:* Subjects increase their trust in the model if its observed accuracy is higher than its stated accuracy, and decrease their trust otherwise.
- *Self vs. stated:* Subjects increase their trust in the model if the model's stated accuracy is higher than their own accuracy in Phase 1, and decrease their trust otherwise.
- *Self vs. observed:* Subjects increase their trust in the model if the model's observed accuracy is higher than their own accuracy in Phase 1, and decrease their trust otherwise.

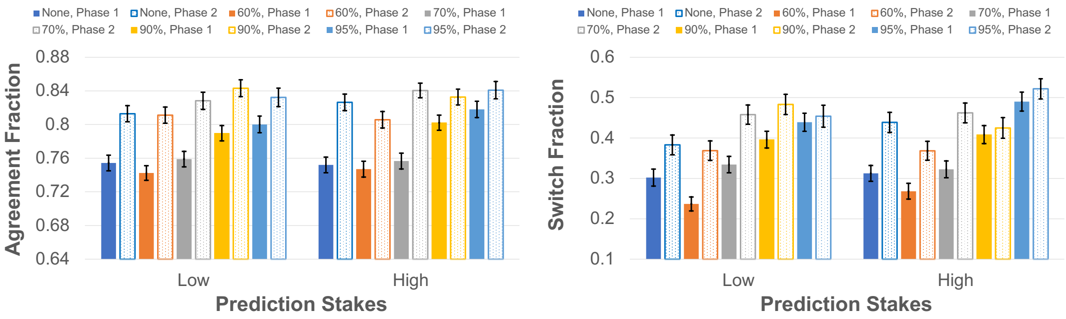
We consider the evidence for and against each mechanism. Of course, this list is not exhaustive, and subjects' behavior may be some combination of the three. Our goal is only to determine which mechanism is most consistent with our data.

Analysis of Differences in Trust Between Phases

We first argue that the experimental design used in our first two experiments allows us to compare subjects' trust between Phase 1 and Phase 2. This is because for each subject, we randomized the set of tasks (*A* or *B*) used in each phase. Approximately half of the subjects saw the tasks in set *A* during Phase 1 and the tasks in set *B* during Phase 2, while the other half saw the reverse. Therefore any inherent differences between the sets of tasks or the models' predictions for them should be mitigated by averaging trust over subjects. This is not the case for our third experiment. We therefore restrict our analysis to the data from our first two experiments.

We start by analyzing data from our first experiment. Figures 6a and 6b show subjects' average agreement fraction and average switch fraction, respectively, in Phase 1 (solid bars) and Phase 2 (dotted bars), by treatment. For all ten treatments, subjects appear to increase their trust between Phase 1 and Phase 2. To confirm this, we conduct paired t-tests to determine whether the increase for each treatment is statistically significant. We find that subjects in *all* treatments significantly increase their agreement fractions between Phase 1 and Phase 2 (the largest p-value is 0.019 for the High-95% treatment). Subjects in all treatments also increase their switch fractions, and this increase is statistically significant ($p < 0.001$) in all but three treatments (the exceptions are High-90%, Low-95%, and High-95%). These results provide evidence against the stated vs. observed mechanism, since we do not see a decrease in subjects' trust when the stated accuracy is higher than the observed accuracy (80%).

Next, we investigate whether and how a subject's own accuracy affects their trust in a model. Among the 1,994 subjects in the first experiment, only 3.4% (68 subjects) were at least as accurate as the model during Phase 1. To differentiate between the self vs. stated and self vs. observed mechanisms, we therefore restrict our analysis to those subjects whose



(a) Phase 1 vs. Phase 2: Agreement fraction

(b) Phase 1 vs. Phase 2: Switch fraction

Figure 6: Comparing agreement fraction and switch fraction between Phase 1 and 2 across treatments in our first experiment.

Treatment	N	Phase 2 - Phase 1 (δ)	p-value
(Low stakes, Stated-60%)	75	0.035	0.014^{*}
(High stakes, Stated-60%)	71	0.005	0.721
(Low stakes, Stated-70%)	14	0.032	0.375
(High stakes, Stated-70%)	16	0.05	0.037^{*}

(a) Agreement fraction

Treatment	N	Phase 2 - Phase 1 (δ)	p-value
(Low stakes, Stated-60%)	75	0.15	$1.7\times10^{-4***}$
(High stakes, Stated-60%)	71	0.11	0.006^{**}
(Low stakes, Stated-70%)	14	0.227	0.302
(High stakes, Stated-70%)	16	0.222	0.066

(b) Switch fraction

Table 1: Differences in trust between Phase 1 and Phase 2, by treatment, for subjects in our first experiment whose own accuracy on the tasks in Phase 1 was higher than the model's stated accuracy, but lower than the model's observed accuracy. N denotes the number of such subjects. p-values are reported for paired t-tests; the symbols *, **, and *** represent significance levels of 0.05, 0.01, and 0.001, respectively.

own accuracy on the tasks in Phase 1 was higher than the model's stated accuracy, but lower than its observed accuracy. Tables 1a and 1b show differences in these subjects' average agreement fraction and average switch fraction, respectively, between Phase 1 and Phase 2, by treatment. These subjects appear to increase their trust between Phase 1 and Phase 2.[7] Paired t-tests indicate that these increases are often statistically significant. These results provide evidence against the self vs. stated mechanism; only the self vs. observed mechanism remains consistent with our experimental data.

Finally, we analyze the data from our second experiment to consider evidence for and against the self vs. observed mechanism. We find that our data are mostly consistent with this mechanism, although we do find one exception. Specifically, we would expect that for the treatment with a stated accuracy of 90% and an observed accuracy of 55%, subjects whose own accuracy was less than 55% would increase their trust between Phase 1 and Phase 2; however, this was *not* the case. Although we see an insignificant increase in the average agreement fraction ($\delta = 0.017, p = 0.137$), we also see an insignificant decrease in the average switch fraction ($\delta = -0.012, p = 0.741$) and in subjects' self-reported levels of trust in the model ($\delta = -0.3609, p = 0.137$). These results suggest that, to the extent that subjects do compare their own accuracy with the model's observed accuracy, the effect of this comparison can be diluted by other factors, such as the difference between the model's stated and observed accuracies.

6 DISCUSSION

In this paper, we investigate whether the accuracy of an ML model affects laypeople's willingness to trust the model via a sequence of large-scale, randomized, pre-registered human-subject experiments. We find that a model's stated accuracy on held-out data affects people's trust in the model, but that the effect size is smaller after people observe the model's accuracy in practice. Furthermore, if a model's observed accuracy is low, then its stated accuracy has at most a very small effect on people's trust in the model. We also find that after observing a model's accuracy in practice, people's trust in the model is significantly affected by its observed accuracy regardless of its stated accuracy. Via an exploratory analysis, we find that after observing a model's accuracy in practice, people are more likely to increase their trust in the model if the model's observed accuracy is higher than their own accuracy. We do find one exception though: if a model's observed

[7] We note that restricting our analysis to subjects whose own accuracy is in a particular range may inadvertently introduce a bias toward subjects who saw a particular set of tasks (i.e., A or B) first. We therefore calculated, for each treatment in Table 1, the fraction of subjects whose own accuracy was between the model's stated and observed accuracy and who saw tasks in set A (or B) first. The only severe imbalance was for the High-60% treatment, where 36.6% of these subjects saw set A first and 63.4% saw set B first. A closer look at the data suggests, though, that both groups of subjects revised their predictions to match those of the model more often during Phase 2 than during Phase 1, which is consistent with the overall results in Table 1.

accuracy is substantially lower than its stated accuracy, people do not revise their predictions to match those of the model after observing its accuracy in practice, even when their own accuracy is lower than the model's observed accuracy.

At the end of each experiment, we asked each subject for some basic demographic information—specifically their age, gender, and education level—via an exit survey. This information enabled us to investigate whether people's trust in a model varies with their demographics. In general, we did not find any differences between demographic groups. Of course, it is possible that demographic information other than age, gender, and education level might affect trust; we therefore highlight this as a potential avenue for future exploration.

Additionally, we acknowledge that trust is notoriously difficult to isolate and measure. In our experiments, we measured people's trust in terms of both behavioral measures (i.e., agreement fraction and switch fraction) and self-reports; these approaches gave consistent results. However, due to our experimental design (in particular, the difference between a model's stated and observed accuracies), it is possible that trust is affected by other factors such as surprise, confusion, and cognitive dissonance, all of which may mediate the effect of accuracy. Further research is needed to carefully determine whether this is the case. It would also be interesting to see whether people trust a model more when there is no difference between its stated and observed accuracies.

Our results offer a number of actionable implications for the human–computer interaction and ML communities to improve the trustworthiness of ML. First, our results highlight the need for designers of ML systems to clearly and responsibly communicate their expectations about model performance, as this information shapes the extent to which people trust a model, both before and after they are able to observe and interact with it in practice. Moreover, our experiments also show that people put substantial weight on their own interactions with a model when deciding how much to trust it, even if their interactions are limited to a few prediction tasks. This behavior is likely desirable in settings where there is a substantial mismatch between held-out data and real-world use cases, meaning that a model's stated performance on held-out data does not accurately reflect its performance post deployment. However, a model's observed performance on a small set of prediction tasks is not necessarily a good indicator of its average performance in practice. It is therefore important for designers of ML systems to convey the uncertainty inherent in performance calculations based on a small set of prediction tasks so that people do not mistakenly distrust an accurate model if its performance on the first few prediction tasks they observe in practice is low.

More broadly, it is crucial for designers of ML systems to properly communicate the uncertainty that is baked into *every* ML prediction. How best to achieve this goal is an area for future research, but could mean, for example, accompanying a model with a summary of characteristics of the data on which it was trained and evaluated in order to help end users reason about scenarios in which the model is most likely to achieve its stated performance on held-out data [10, 19]. This may help end users develop a better sense of how systematic discrepancies between evaluation and deployment environments might affect performance. In particular, we highlight the importance of further human–computer interaction research aimed at helping laypeople understand uncertainty. For example, visualization techniques could help people infer a model's expected performance post deployment from its stated and observed accuracies. Additional studies should be conducted to understand how the ways that accuracy is expressed (e.g., different ways to put accuracy in context/perspective) affect people's trust in a model.

Our finding that people are more likely to increase their trust in a model if the model's observed accuracy is higher than their own accuracy suggests a simple real-world intervention that can be used to encourage appropriate levels of trust in a model: ask people to make predictions themselves before using the model so that they can assess how good they are the tasks in question. By doing so, they will be better able to determine how much they should trust the model.

We note that proper caution should be used when generalizing our results to other settings. Because we focused on asking laypeople, recruited on Amazon Mechanical Turk, to make predictions about speed dating events, our results may not hold for settings where, for example, significant domain expertise is required or where the stakes are especially high (e.g., doctors making life-or-death decisions). In addition, subjects observe the model's accuracy in practice during a relatively short time window. Investigating how accuracy affects trust over repeated interactions, spanning a much longer time period, is therefore an important direction to explore. For example, how long does it take people to detect that a model's stated accuracy on held-out data is substantially different from its accuracy post deployment?

Ultimately, our work highlights a pressing need for more experimental studies aimed at understanding people's interactions with different components of the ML pipeline, thereby expanding the umbrella of interpretable machine learning beyond its current focus on model internals. We hope this paper will inspire more work in this direction.

7 ACKNOWLEDGEMENTS

We are grateful to the CHI 2019 anonymous reviewers who provided many helpful comments, and to Miro Dudík, Dan Goldstein, Jake Hofman, Ryan Kennedy, Forough Poursabzi-Sangdeh, and Marco Tulio Ribeiro for valuable discussions. We thank all the Amazon Mechanical Turk workers who participated in our experiments and made this work possible.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016).

[2] 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.

[3] Alexandra Chouldechova, Diana Benavides Prado, Oleksandr Fialko, Emily Putnam-Hornstein, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the First Conference on Fairness, Accountability, and Transparency.*

[4] 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.

[5] 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).

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

[7] Mary T. Dzindolet, Linda G. Pierce, Hall P. Beck, and Lloyd A. Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human Factors* 44, 1 (2002), 79–94.

[8] 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.

[9] 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.

[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. (2018). CoRR arXiv:1803.09010.

[11] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing High Quality Crowdwork. In *Proceedings of the Twenty-Fourth International World Wide Web Conference.*

[12] Kartik Hosanagar and Apoorv Saxena. 2017. The Democratization of Machine Learning: What It Means for Tech Innovation. Knowledge@Wharton, retrieved from http://knowledge.wharton.upenn.edu/article/democratization-ai-means-tech-innovation/.

[13] Matthew Kay, Shwetak N Patel, and Julie A Kientz. 2015. How Good is 85%?: A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* 347–356.

[14] Ryan Kennedy, Philip D. Waggoner, and Matthew Ward. 2018. Trust in Public Policy Algorithms. Working paper.

[15] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence.*

[16] Zachary C. Lipton. 2016. The mythos of model interpretability. (2016). CoRR arXiv:1606.03490.

[17] Jennifer M. Logg, Julia Minson, and Don A. Moore. 2018. Algorithm Appreciation: People prefer algorithmic to human judgment. (2018). Harvard Business School NOM Unit Working Paper No. 17-086.

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

[19] 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 Second Conference on Fairness, Accountability, and Transparency.*

[20] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. (2018). CoRR arXiv:1802.00682.

[21] David W Nickerson and Todd Rogers. 2014. Political campaigns and big data. *Journal of Economic Perspectives* 28, 2 (2014), 51–74.

[22] Dilek Önkal, Paul Goodwin, Mary Thomson, and Sinan Gönül. 2009. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* 22 (2009), 390–409.

[23] Umberto Panniello, Michele Gorgoglione, and Alexander Tuzhilin. 2016. Research note–In CARSs we trust: How context-aware recommendations affect customers? Trust and other business performance measures of recommender systems. *Information Systems Research* 27, 1 (2016), 182–196.

[24] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. (2018). CoRR arXiv:1802.07810.

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining.*

[26] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction.* 141–148.

[27] Jennifer Wortman Vaughan and Hanna Wallach. 2017. The Inescapability of Uncertainty. In *CHI Workshop on Designing for Uncertainty in HCI: When Does Uncertainty Help?*

[28] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.*

[29] Peng Xia, Hua Jiang, Xiaodong Wang, Cindy X Chen, and Benyuan Liu. 2014. Predicting User Replying Behavior on a Large Online Dating Site.. In *Proceedings of the International Conference on Web and Social Media.*

[30] Michael Yeomans, Anuj K. Shah, Sendhil Mullainathan, and Jon Kleinberg. 2018. Making sense of recommendations. Working paper.

[31] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2016. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization.* 223–227.

[32] 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.* 307–317.