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

Supplemental Material

ZHUORAN LU, Purdue University, USA
MING YIN, Purdue University, USA

1 DATA AND ANALYSIS CODE

All of the data and analysis code are published on the GitHub repository at https://github.com/ZhuoranLu/Trustworthy-ML.

2 CATEGORIZING TASK INSTANCES THROUGH A PILOT STUDY

In order to implement our experiment design, we need to first identify a set of prediction tasks where on each task, *most* people make the same prediction so that in our experimental studies we can easily vary the level of agreement between people and an ML model by changing the model's prediction to agree/disagree with the prediction of the majority. In addition, we also need both prediction tasks that people are confident about their own predictions and the ones that people are not.

Therefore, we conducted a pilot study on a subset of 214 speed dating events sampled from the dataset of Fisman et al. [1]. We recruited workers from Amazon Mechanical Turk (MTurk) to make predictions on these speed dating events through a human intelligence task (HIT). Specifically, each worker was asked to complete a random sample of 20 prediction tasks in a HIT. In each task, the worker made a binary prediction on the outcome of the speed dating event (i.e., the participant would/would not be willing to see the date again), and indicated her confidence in the prediction by selecting from one of the three options—"not confident", "somewhat confident", and "very confident". An attention check question was also included in each HIT, where the worker was instructed to select some pre-specified options. Our HITs were open to US workers only.

A total of 315 workers participated in our pilot study. We then filtered out all the predictions made by workers who failed to answer the attention check question correctly. Furthermore, we focused on considering only those task instances where at least 10 workers made a prediction. Our final dataset, therefore, included 4,537 predictions from MTurk workers on 201 task instances.

Based on the data we collected, we defined the following properties for each task instance:

- *Average accuracy*: the fraction of workers who made a correct prediction on the task among all workers who made a prediction on it.
- *Average confidence*: the average level of confidence workers reported for their predictions on the task, where the confidence values were standardized within each worker[1].

Figure 1 shows the distribution of the 201 task instances on average accuracy and confidence.

[1]The three confidence levels—not confident, somewhat confident, and very confident—were mapped to numeric values of 0, 1, 2, respectively. We standardized the confidence values within each worker to remove individual biases in reporting confidence.

Authors' addresses: Zhuoran Lu, lu800@purdue.edu, Purdue University, 305 N. University Street, West Lafayette, IN, USA, 47907; Ming Yin, mingyin@purdue.edu, Purdue University, 305 N. University Street, West Lafayette, IN, USA, 47907.

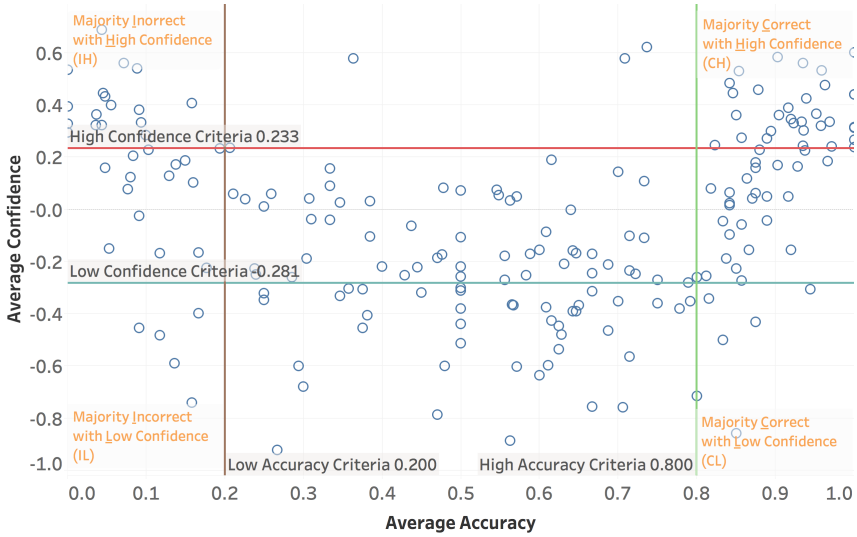
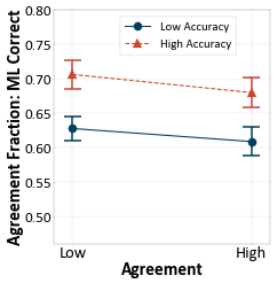
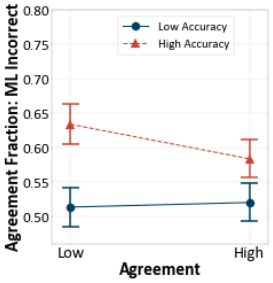


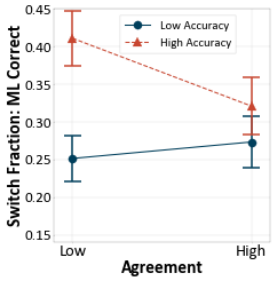
Fig. 1. Average accuracy and average confidence for task instances with at least 10 predictions in the pilot study. Each circle represents one task instance.



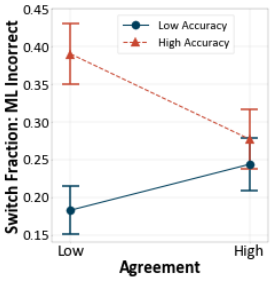
(a) Agreement Fraction: ML Correct



(b) Agreement Fraction: ML Wrong



(c) Switch Fraction: ML Correct



(d) Switch Fraction: ML Wrong

Fig. 2. The average values of agreement fraction and switch fraction in Phase 2 across four treatments in Experiment 2, on the subset of Phase 2 tasks where the ML model was correct (Fig. 2a, 2c), and on the subset of Phase 2 tasks where the ML model was incorrect (Fig. 2b, 2d). Error bars represent the standard errors of the mean.

With the knowledge of the distribution of the tasks, we then used the first and third quantile of the average confidence across all task instances as our thresholds to differenciate task instances where people are confident or unconfident

about their own predictions. We also required that *at least* 80% of workers need to make the same prediction on a task instance for this task to be considered as satisfying the criteria of "most people provide the same prediction" (which means the average accuracy of these tasks is either not lower than 0.8 or not higher than 0.2).

3 EXPERIMENT 1: ADDITIONAL RESULTS

As a robustness check, in addition to the one-way ANOVA, we also conducted additional statistical tests on the subject's reliance data to understand whether the level of agreement between subjects and an ML model on tasks that subjects are highly confident affects their reliance on the model. Results of these analyses are largely consistent with our one-way ANOVA results. For example, using the Kruskal-Wallis tests, we found that the differences in subject's reliance on the ML model across the three treatments were significant or marginally significant (agreement fraction: $H(2) = 5.27$, $p = 0.072$; switch fraction: $H(2) = 11.21$, $p = 0.004$). Furthermore, when using beta regressions to understand the relationship between the level of high confidence human-model agreement and people's reliance on the model, we also found that higher levels of agreement result in significantly higher levels of reliance on the model (agreement fraction: estimated coefficient $\beta = 0.009, p < 0.001$; switch fraction: $\beta = 0.008, p = 0.005$)[2].

4 EXPERIMENT 2: ADDITIONAL RESULTS

4.1 Additional analysis on subject's reliance on the model in Phase 2

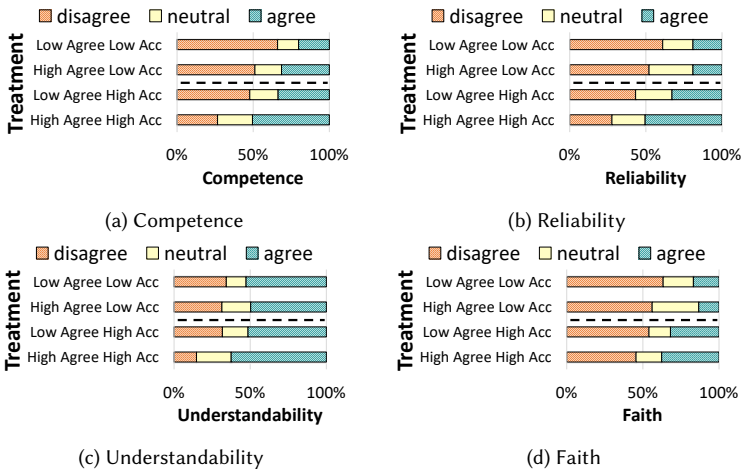
First, beyond the two-way ANOVA tests, we conducted beta regressions to analyze how the level of high confidence human-model agreement and the model's accuracy, together, affect people's reliance on an ML model. Consistent with our ANOVA results, we found that higher levels of model accuracy significantly increase people's reliance on the ML model in Phase 2 (agreement fraction: $\beta = 0.38$, $p = 0.006$, switch fraction: $\beta = 0.56$, $p = 0.001$), but higher levels of human-model agreement in Phase 1 do not ($p > 0.05$). Besides, we also found a significant interaction between the level of high confidence human-model agreement and the level of model accuracy on influencing the switch fraction ($\beta = -0.52$, $p = 0.031$).

Moreover, we separated Phase 2 tasks where the ML model made correct predictions from those tasks where the ML model made wrong predictions, and we looked into subjects' reliance on the model on these two subsets of Phase 2 tasks respectively. Figure 2 shows the average values of agreement fraction and switch fraction across four treatments in Experiment 2, on Phase 2 tasks where the ML model was correct (Figures 2a, 2c), and on Phase 2 tasks where the ML model was incorrect (Figures 2b, 2d). The trends are consistent with the subject's reliance on the ML model on all Phase 2 tasks.

4.2 Impact of high confidence human-model agreement on perceptions of the model after some performance feedback is obtained

To see that in Experiment 2, how subject's reliance on the model might have been influenced by their perceptions of the model, we plotted the proportions of subjects across all 4 treatments who agreed, disagreed, or stood neutral on the four statements about the model's competence, reliability, understandability, and their faith in the model after they received the information about the model's overall accuracy in Phase 1 in Figures 3a–3d. Proportion test results suggest that significant differences exist across the 4 treatments on almost all four aspects of model perceptions ($p < 0.01$,

[2]Both our dependent variables (i.e., agreement fraction and switch fraction) take values in the interval $[0, 1]$. To map the dependent variable value x into the range of $(0, 1)$, for all beta regressions that we conducted, we applied the transformation $x' = \frac{x+\epsilon}{1+2\epsilon}$, where ϵ is a small value close to 0.



(a) Competence (b) Reliability

(c) Understandability (d) Faith

Fig. 3. Proportions of subjects who agreed, disagreed, or stood neutral on the statements on the model's competence, reliability, understandability, and faith in the model, across the four treatments in Experiment 2.

except for the fraction of subjects who disagreed the statement on their faith in the model, $p = 0.059$, and the fraction of subjects who agreed with the statement on the model's understandability, $p = 0.24$).

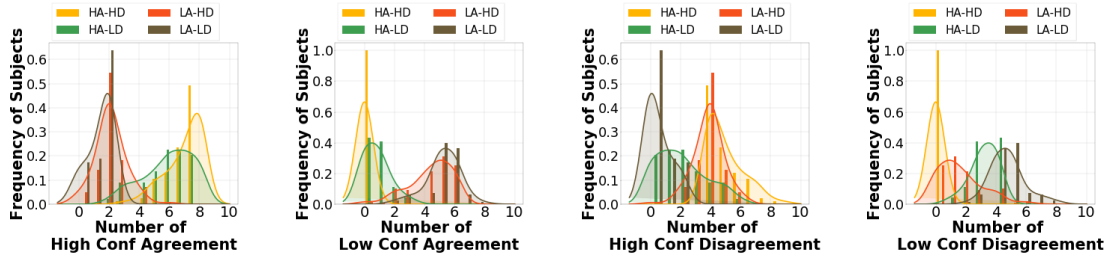
However, using post-hoc pairwise proportion tests, we found that after subjects observed the model's accuracy in Phase 1 was 50%, the fraction of subjects who agreed/disagreed that the model was competent, reliable, understandable, and they have faith in the model was *not* affected by the level of agreement with the model subjects had experienced in Phase 1 ($p > 0.1$). On the other hand, after subjects observed the model's accuracy in Phase 1 was 80%, compared to subjects who experienced a low level of agreement with the model in Phase 1, those who had experienced a higher level of agreement tended to feel the model was slightly more competent (agree: $p = 0.073$, disagree: $p = 0.010$), reliable (agree: $p = 0.050$, disagree: $p = 0.076$), and understandable (agree: $p = 0.68$, disagree: $p = 0.029$), though they did not have significantly higher levels of faith in the model (agree & disagree: $p > 0.1$).

Similarly, we also examined subject's self-reported overall trust in the model across all treatments. The two-way ANOVA test suggests significant main effects for both the model accuracy ($p < 0.001$) and the human-model agreement ($p = 0.011$). However, post-hoc Tukey's HSD test showed that the effect of human-model agreement was not significant when the model's observed accuracy was 50% ($p = 0.65$), and the effect is marginal when the model's observed accuracy was 80% ($p = 0.057$).

5 EXPERIMENT 3: ADDITIONAL RESULTS

5.1 Confirming the validity of experimental design

In Experiment 3, we restrict our analyses to the subset of subjects who agreed with the ML model on at most 8 tasks in Phase 1. To see that within this subset, whether subjects' confidence in the tasks that they agreed or disagreed with the model across different treatments aligned well with our expectations, we showed the histograms of the number of high/low confidence tasks that subjects in each treatment agreed or disagreed with the ML model in Phase 1 in Figure 4. Overall, the subject's actual agreement/disagreement with the model in different treatments was consistent with our design. For example, subjects in HA–HD and HA–LD treatments agreed with the model on high confidence tasks significantly more often than subjects in LA–HD and LA–LD treatments (Figure 4a), while subjects in HA–HD and LA–HD treatments also tended to disagree with the model on high confidence tasks more compared to subjects



(a) number of high confidence agreement (b) number of low confidence agreement (c) number of high confidence disagreement (d) number of low confidence disagreement

Fig. 4. Comparing the number of agreement/disagreement between subjects and the ML model on high confidence tasks and low confidence tasks in Phase 1 of Experiment 3 across the 4 experimental treatments (only subjects who agreed with the ML model on 8 or fewer tasks in Phase 1 are included). Gaussian kernels are used for kernel density estimation.

in HA–LD and LA–LD treatments (Figure 4c). In other words, our experimental design successfully varied subjects' confidence in the tasks that they agreed or disagreed with the model across different treatments.

5.2 Additional analysis on subject's reliance on the model in Phase 2

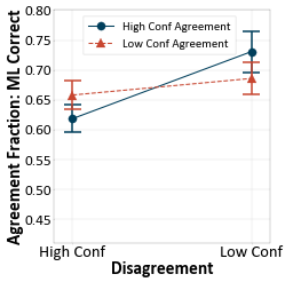
First, beyond the two-way ANOVA tests, we conducted beta regressions to analyze how subject's confidence in their agreement and disagreement with an ML model affects their reliance on the model. Again, we found the interactions between subject's confidence in agreement and subject's confidence in disagreement are significant on affecting both measures of reliance (agreement faction: $\beta = -0.94$, $p < 0.001$, switch fraction: $\beta = -0.97$, $p = 0.003$). These results align well with the ones that we obtained from the two-way ANOVA tests.

In addition, Figures 5a and 5c showed that in Experiment 3, subject's reliance on the ML model on Phase 2 tasks where the model made correct predictions. Similarly, Figures 5b and 5d showed subject's reliance on the ML model on Phase 2 tasks where the model made wrong predictions. Here, we observed the same interaction patterns as those seen when examining subject's reliance on the ML model on all Phase 2 tasks. Two-way ANOVA tests further confirm that in Figures 5a–5d, the interaction effects between the two factors (subject's confidence in agreement/disagreement with the model) on influencing subject's reliance on the model are almost always statistically significant at the level of $p < 0.05$ (the only exception is Figure 5a, $p = 0.133$).

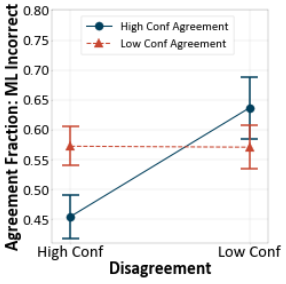
5.3 Impact of human confidence in agreement/disagreement on perceptions of the model

Subject's evaluations on the model's competence, reliability, understandability, as well as their faith and overall trust in the model across the 4 treatments in Experiment 3 are shown in Figures 6a–6e. Interestingly, we again detected the pattern that people's confidence in their agreement and disagreement with the model *interact* with each other in influencing people's perceptions of the model's competence and understandability, as well as people's overall trust in the model. Two-way ANOVA tests show that such interaction is statistically significant for overall trust ($p = 0.034$) and marginally significant for model competence ($p = 0.051$).

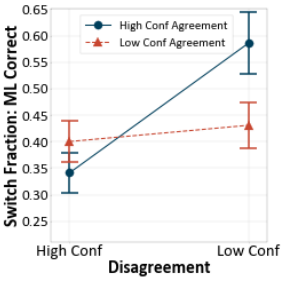
On the other hand, we did not observe similar cross-over interaction pattern for subject's perception in the reliability of a model (Figure 6b) or their faith in a model (Figure 6d). To the contrary, we found that compared to low confidence disagreement, high confidence disagreement consistently results in a significantly lower perceived level of model



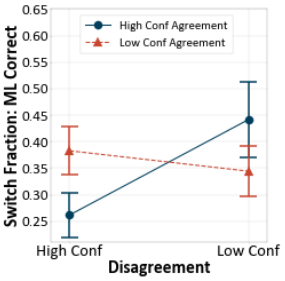
(a) Agreement Fraction: ML Correct



(b) Agreement Fraction: ML Wrong



(c) Switch Fraction: ML Correct



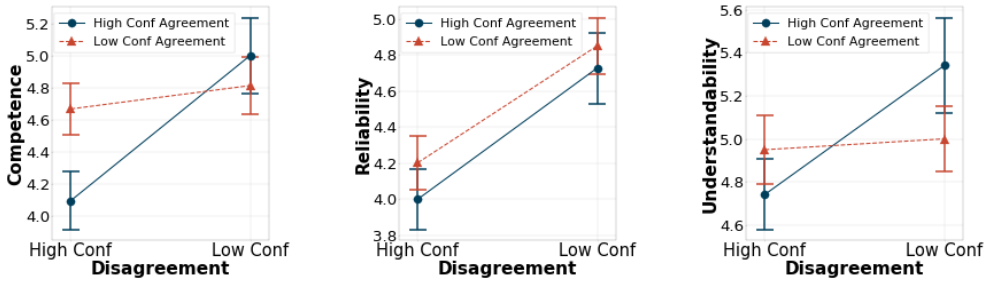
(d) Switch Fraction: ML Wrong

Fig. 5. The average values of agreement fraction and switch fraction in Phase 2 across four treatments in Experiment 3, on Phase 2 tasks where the ML model was correct (Fig. 5a, 5c), and on Phase 2 tasks where the ML model was incorrect (Fig. 5b, 5d). Error bars represent the standard errors of the mean.

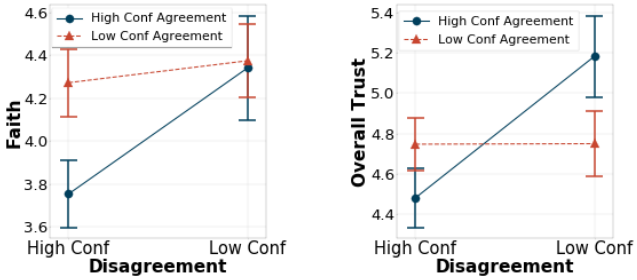
reliability ($p < 0.001$). Additionally, low confidence human-model agreement leads to higher faith in the model compared to high confidence human-model agreement ($p = 0.042$).

REFERENCES

[1] 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.



(a) Competence (b) Reliability (c) Understandability



(d) Faith (e) Overall trust

Fig. 6. Average values of subject's assessment on model competence, reliability, understandability, as well as their faith and overall trust in the model in Experiment 3 (constrained to subjects who agreed with the ML model's predictions on at most 8 tasks in Phase 1). Error bars represent the standard errors of the mean.