Uncalibrated Models Can Improve Human-AI Collaboration

Kailas Vodrahalli* Tobias Gerstenberg† James Zou‡

Abstract

In many practical applications of AI, an AI model is used as a decision aid for human users. The AI provides advice that a human (sometimes) incorporates into their decision-making process. The AI advice is often presented with some measure of "confidence" that the human can use to calibrate how much they depend on or trust the advice. In this paper, we demonstrate that presenting AI models as more confident than they actually are, even when the original AI is well-calibrated, can improve human-AI performance (measured as the accuracy and confidence of the human's final prediction after seeing the AI advice). We first learn a model for how humans incorporate AI advice using data from thousands of human interactions. This enables us to explicitly estimate how to transform the AI's prediction confidence, making the AI *un*calibrated, in order to improve the final human prediction. We empirically validate our results across four different tasks—dealing with images, text and tabular data—involving hundreds of human participants. We further support our findings with simulation analysis. Our findings suggest the importance of and a framework for jointly optimizing the human-AI system as opposed to the standard paradigm of optimizing the AI model alone.

1 Introduction

In safety-critical settings like medicine, AI is often integrated in the form of interactive feedback with a human, allowing the human to decide when and to what extent the AI's "advice" is utilized Vodrahalli et al. [2020]. For example, in medical diagnosis tasks, this essentially places the AI algorithm in a similar category as lab tests or other exams a doctor may order to aid in diagnosis. This form of implementation is important to partially mitigate the often black box nature of AI algorithms that limit a user's trust and usage of AI Feldman et al. [2019], Ribeiro et al. [2016], Xie et al. [2020], Miller [2019].

Typically, AI algorithms are designed and optimized independently of the human users – the AI is designed to be as accurate as possible for its given task using the standard training objectives. This makes sense if the model is used to make isolated decisions and predictions by itself. In this paper, we question this premise and ask whether a joint optimization of the entire human-AI system is possible. In particular, we revisit the conventional wisdom that models with calibrated confidence are desired for collaborative systems, as they better allow accurate transfer of prediction uncertainty between models and/or humans Guo et al. [2017]. We investigate whether explicitly making the AI advice uncalibrated (i.e. overconfident on certain instances) can improve overall human-AI performance, with the intuition that humans rarely are able to accurately integrate calibration information correctly and may in fact benefit from uncalibrated information. We focus on binary classification tasks to simplify our setting. We used the widely-used judge-advisor paradigm (JAS) paradigm from psychology to model the human-AI interaction Van Swol et al. [2018], Prahl and Van Swol [2017].

We then draw on prior work modeling human predictions Vodrahalli et al. [2021] and assume black box access to an AI algorithm which provides advice. From experimental data we collected of thousands of human interactions, we learn a model of how human participants incorporate (or ignore) AI advice in their final predictions. Using this learned human behavior model, we optimize a monotonic transformation that modifies the AI's reported confidence in order to maximize human-AI system performance (measured as the human's final accuracy). We demonstrate first in simulations using our human behavior model and subsequently with real-world data using crowd-sourced human study participants that our modified AI advice results in higher overall performance.

*Stanford University. Email: kailasv@stanford.edu
†Stanford University. Email: gerstenberg@stanford.edu
‡Stanford University. Email: jamesz@stanford.edu

Our contributions We propose a simple framework for optimizing an AI algorithm with respect to a loss function that depends on human utilization of the algorithm's output. We only require black box access to the algorithm. This framework relies on fitting a model to human behavior, which we do using a large collection of empirical data on human interaction with AI advice. We then demonstrate both in simulation and real-world studies that our method, which results in an uncalibrated AI model, improves human-AI performance. These results suggest that optimizing AI algorithms in isolation, as is standard practice, is not optimal. Moreover, to the best of our knowledge, these are some of the first results involving an empirically validated model for human behavior used to optimize AI performance in a collaborative human-AI system, and suggest that this type of framework may have much potential.

Related works The experimental setup we use comes from the advice utilization community. Here, studies often use the JAS framework for measuring and comparing the effect of different forms of advice Van Swol et al. [2018]. Previous work has studied the effect of peer vs. expert advice Madhavan and Wiegmann [2007], automated vs. human advice Prahl and Van Swol [2017], Dzindolet et al. [2002], Madhavan and Wiegmann [2007], Önkal et al. [2009], task difficulty Gino and Moore [2007], and advice confidence Sah et al. [2013], Schultze et al. [2015]. Of particular relevance to our work is Sah et al. [2013], which provides evidence that uncalibrated, confident advice is more utilized by humans in certain settings. We also leverage a proposed model for human behavior from Vodrahalli et al. [2021]. Though we leverage this prior work, our contributions are novel – rather than just attempt to understand factors that affect human-AI collaboration, we develop a novel optimization framework to improve human-AI collaboration in practice.

The conventional approach to human-AI collaboration is to make the AI's predictions explainable Weitz et al. [2019], Shin [2021] and to ensure model and human calibration Guo et al. [2017], Chiou and Lee [2021]. It is standard practice to optimize the AI in isolation (i.e., maximize the AI performance). Recent work suggests that these approaches are not sufficient, and that the objectives used to optimize the AI do not adequately consider the human-AI team. Alternative approaches including learning to defer, where a classifier is trained to determine when expert input is required Madras et al. [2017], Mozannar and Sontag [2020]; learning to complement, where the AI is optimized to perform well on only the tasks that humans struggle with Wilder et al. [2020]; and methods that optimize objectives with costs assigned to requesting AI advice and/or human expert input utilization Bansal et al. [2021]. Our approach is complementary to these prior works and contains several novel aspects. We model human behavior using empirical data, demonstrate that this human model allows us to optimize the AI advice, and demonstrate both in simulation and empirically that our modified advice improves overall system performance. Furthermore, we only require black-box access to the AI advice. There is a large body of work on learning calibrated models. Our focus is different in that we demonstrate how we can improve human performance by making the AI advice uncalibrated in a specific way.

2 Experimental setup

Our goal is to augment the performance of a human-AI system by optimizing the AI for use by a human. To make this goal tractable, we assume (1) that we have a model for human interaction with AI advice, and (2) that we have black box access to the AI algorithm producing the advice. We discuss the human behavior model in Section 3 and the AI algorithm in Section 4.2.

2.1 Human-AI system

We assume the human-AI system shown in Figure 1. A human is shown a task and completes it. AI advice is then presented, and the person is allowed to modify their response. As we measure performance before and after receiving advice, this model is conducive to studying the effect of advice and so sees common usage Prahl and Van Swol [2017], Vodrahalli et al. [2021], Mesbah et al. [2021].

2.2 Data and tasks

We focus on binary prediction tasks. As seen in Figure 1, participants respond on a continuous sliding scale. They are instructed to move an icon on the scale to the point that indicates their confidence in their answer.

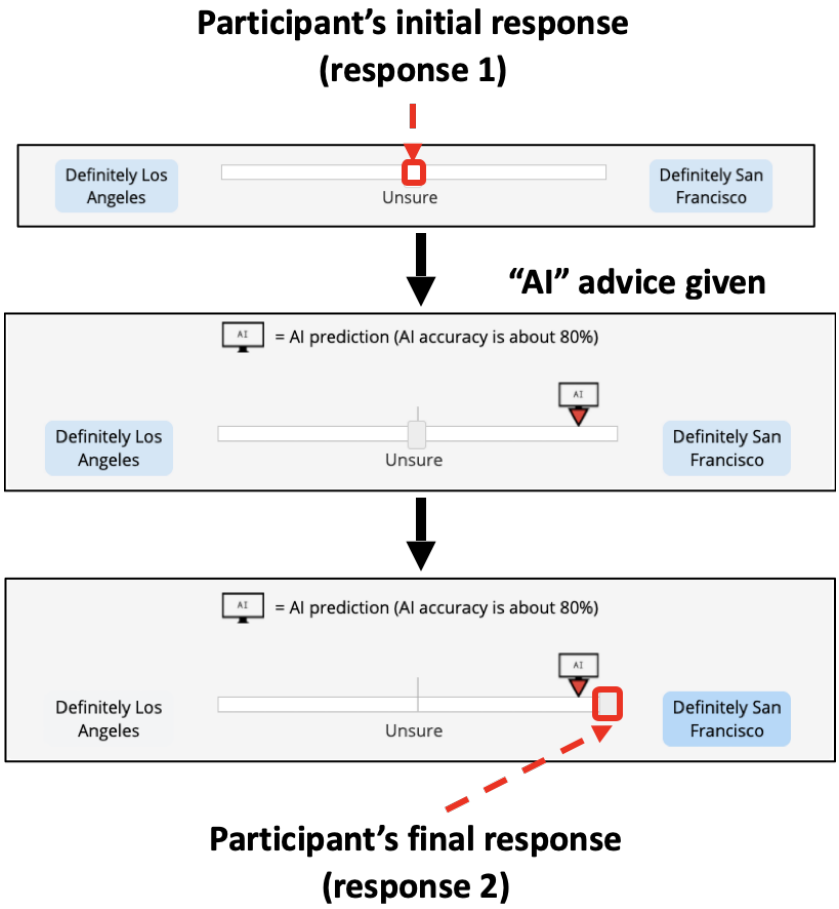


Figure 1: Our experimental setup. A human first completes a task on their own (Response 1). They then are given advice from an AI algorithm and are allowed to change their answer (Response 2).

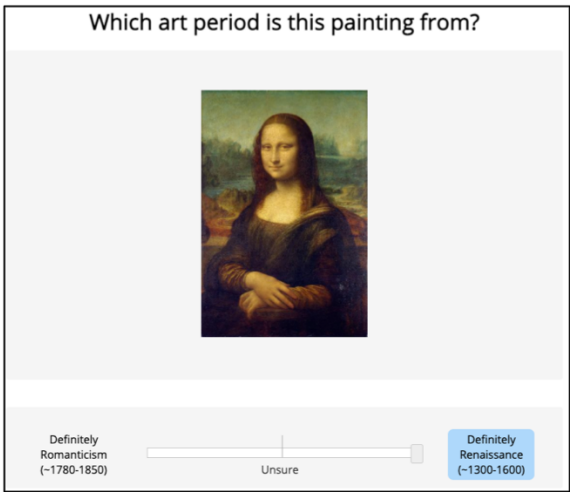
They are incentivized so that higher confidence in correct and incorrect answers increase and decrease reward respectively. The advice also appears on the same sliding scale. The use of the scale allows for a more rich analysis of the user's prediction than just the simple binary prediction label would permit.

We collected data on a set of four tasks from diverse data modalities. These tasks were originally proposed in prior work Vodrahalli et al. [2021]. The tasks were designed to be accessible to the general public to allow use of crowd-sourced study participants for obtaining empirical data. See below for a brief description of the tasks we used; examples of task instances from each of the four tasks are shown in Figure 2. The data collected is summarized in Table 1. Each task consists of 32 binary questions. All participants see the same set of questions, though the order is randomized for each participant. Participants also receive similar advice – a small amount of random noise is added to the advice each participant sees. See Section 4.2 for more details.

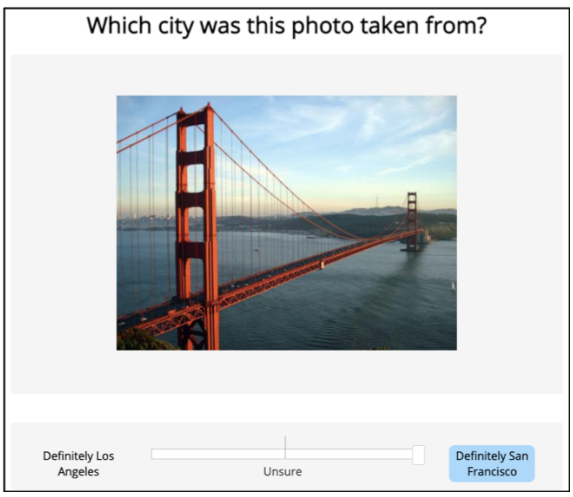
We recruit between 49-79 participants for each task, resulting in several thousand datapoints (each question is one datapoint). Participants are US residents, roughly 50% are female, and the average age is approximately 30 years old.

We followed standard practice in collecting data – the data collection process was low risk, and we ensured samples contained diverse participants across age, sex, and ethnicity. We also provided compensation above the recommended rate. More details are included in the supplement.

Art dataset (Image data) Contains images of paintings from 4 art periods: Renaissance, Baroque, Romanticism, and Modern Art. Participants were asked to determine the art period a painting is from given a binary choice.

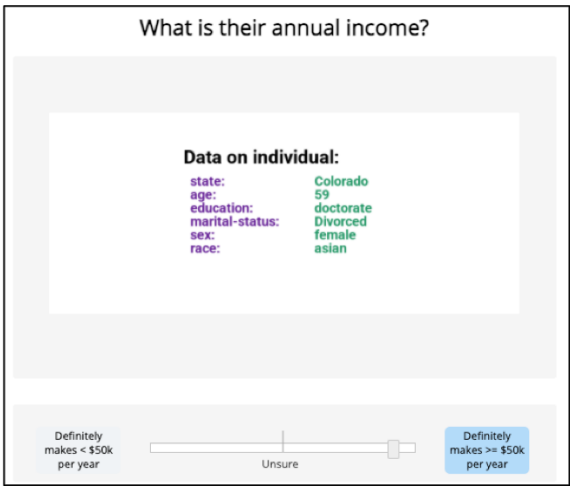


(a) Art dataset



(b) Cities dataset

(c) Sarcasm dataset



(d) Census dataset

Figure 2: Example tasks for each of the 4 datasets we use.

Cities dataset (Image data) This dataset contains images from 4 major US cities: San Francisco, Los Angeles, Chicago, and New York City. The task is to identify which city an image is from given a binary choice.

Sarcasm dataset (Text data) This dataset is a subset of the Reddit sarcasm dataset Khodak et al. [2017], which includes text snippets from the discussion forum website, Reddit. Participants were asked to detect whether sarcasm was present in a given text snippet.

Census dataset (Tabular data) This dataset comes from US census data West and Praturu [2019]. The task is to identify an individual's income level given some of their demographic information: state of residence, age, education level, marital status, sex, and race.

3 Human behavior model

Table 1 provides a summary of the data collected for developing our human behavior model. We find that humans actually changed their response after receiving advice roughly 50% of the time, with the exact

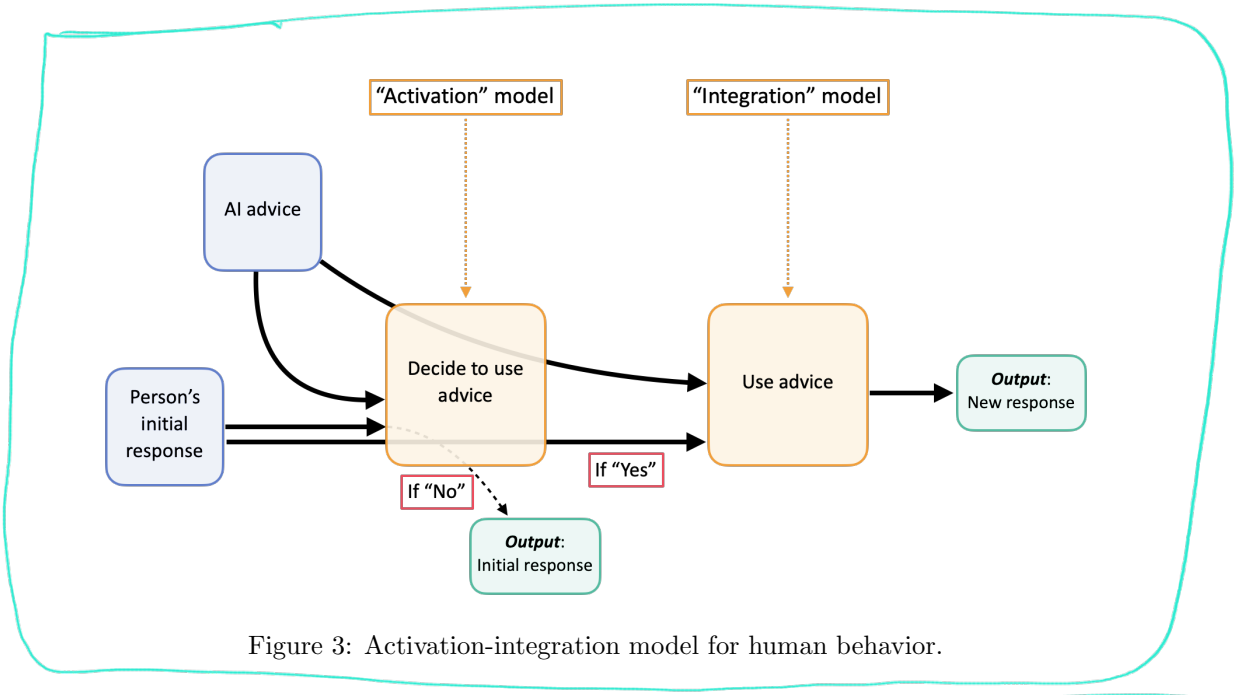


Figure 3: Activation-integration model for human behavior.

Table 1: Summary of data used for fitting our human behavior model.

Task	# Participants	# Observations	Activation Rate	Accuracy (before advice)	Accuracy (after advice)
Art	68	2176	0.523	63.7%	75.5%
Cities	79	2528	0.557	70.7%	77.1%
Sarcasm	49	1567	0.347	72.7%	76.1%
Census	50	1600	0.475	70.1%	73.6%

number shown in the "Activation Rate" column of Table 1. When a person does change their response, we call this person "activated." Specifically, a person is activated if their final response changes by at least a small amount ($\geq 3.5\%$ – the slider bar width).

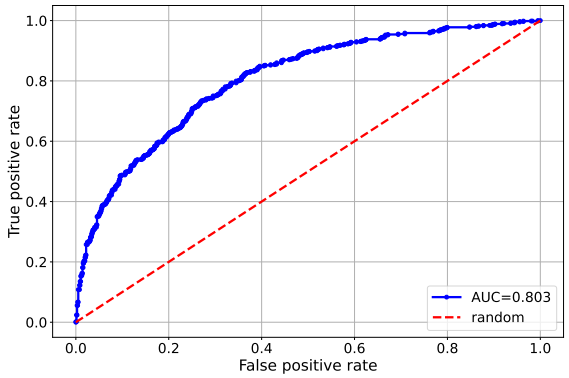
This observation motivates a simple, two-stage model for human behavior: a human first decides whether to modify their response ("activation stage"), and subsequently, if they were activated, they modify their response ("integration stage"). A diagram of this model is shown in Figure 3. This two-stage model is consistent with previously proposed models from the psychology and HCI literature Harvey and Fischer [1997], Vodrahalli et al. [2021].

We fit functions for each of the two stages: an activation model that predicts how likely it is that a person is activated, and an integration model that predicts how the person modifies their advice, assuming they are activated. Taken together, we can model human behavior in the two-stage human-AI system described in Figure 1. We define this model's behavior more rigorously in Section 4.

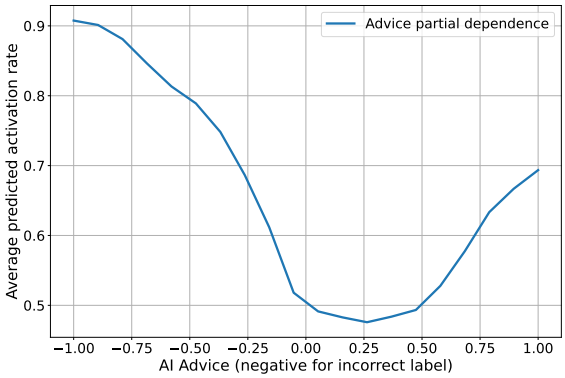
Our fitted models will only be used to optimize our advice-modifying function, and will not be used a test time. As such, we can utilize demographic information about the participants to fit the model, though such information may not be available during a real-world deployment. We opt to fit a small neural network for each of these stages, as we require a differentiable model (for easy optimization) with moderately more complexity than a linear model.

3.1 Activation model

Our activation model takes 12 input features extracted from a single person-task instance interaction. It outputs a probability for whether the person will be activated on the given task instance. The input features



(a) ROC-AUC curve



(b) Partial dependence plot for AI advice

Figure 4: Performance of our activation model for human behavior: (a) ROC-AUC curve for model performance on our validation set. (b) Partial dependence plot measuring the effect of the advice on activation.

are based on the person's initial response, the AI advice, and several demographic features including age and sex. We do not encode the initial response or AI advice directly, as the label (the left or right slider location of the correct answer) is randomized for each participant. Instead, we use the response and advice confidence, as well as a binary feature that encodes whether the response and advice agreed on the label. This implies a symmetry in the model's predictions. Please see the Appendix for a description of all the features used.

We use a 3 layer neural network with ReLU non-linearities for the activation model. We also attempted to use other simple algorithms for modeling the activation behavior of humans, like linear models, but found the neural networks to perform better. We trained the network using binary cross-entropy loss with an Adam optimizer Kingma and Ba [2014] and stopped training when validation loss increased, as is standard. A visualization of the activation model performance is shown in Figure 4a. Here we show the ROC-AUC curve of our model; it achieves an AUC > 0.8 in predicting activation on held-out test participants, suggesting reasonable performance despite the high variability in our dataset.

In Figure 4b, we show a partial dependence plot for the learned activation model. This plot shows the average behavior (across our entire test dataset) of the activation model when we fix the AI advice to be a certain value (the x-value) for all data points. The AI advice can be positive (correct advice) or negative (incorrect advice), while the magnitude corresponds to the advice confidence. We note that (1) our activation model predicts people will be most activated when the advice is confidently incorrect. This occurs because (a) most people in our dataset get answers correct, on average, and seeing confident advice of the opposite label tends to make people change their initial response (e.g. by decreasing confidence even if the predicted label does not change). And because (b) confident advice that agrees with a person's own initial response tends to increase their confidence. We also observe that (2) the activation rate decreases when advice is less confident, but increases for confident and correct advice due to the same reasons noted previously for confident and incorrect advice. Activation rate is lower for correct advice than for incorrect advice as most people in our dataset are, on average correct, which tends to invoke a smaller response.

3.2 Integration model

Our integration model is similar to our activation model, but optimized for a different function. In particular, we use the same input features, model architecture, and training procedure. The integration model outputs a prediction for $sign(r_1)(r_2 - r_1)$, where $r_1, r_2 \in [-1, 1]$ are the initial and final responses respectively. Note that r_1, r_2 are probabilities normalized to the $[-1, 1]$ scale. The training minimizes the RMSE of the predicted change and the actual change. The optimization is performed on only the subset of the dataset that was actually activated (roughly 50% of the data). Our integration model has a reasonable performance, with an RMSE of 0.25 (the maximum absolute error possible is 2) and R^2 of 0.73 on test data.

In Figure 5, we plot a heatmap of the fitted integration model's behavior. We plot the average output of our integration model across our test data split when we fix the person's initial response (x-axis) and

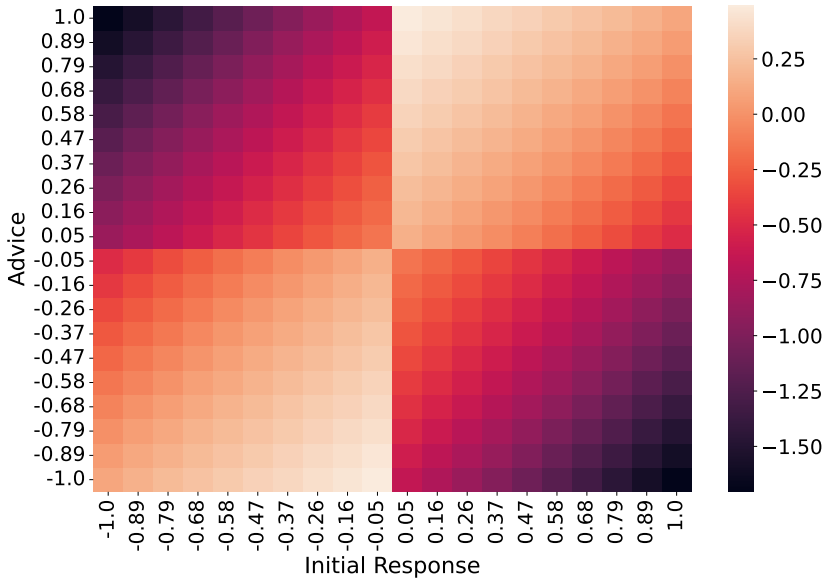


Figure 5: Heatmap of the average behaviour of our integration model for human behavior when we vary AI advice and the human's initial response.

AI advice (y-axis) for all datapoints to a certain value. The initial response and AI advice can be positive (correct advice) or negative (incorrect advice), while the magnitude corresponds to the advice confidence. The integration model prediction can also be positive (increasing confidence in initial predicted label) or negative (decreasing confidence in initial predicted label / change in predicted label).

We observe three features in the heatmap: (1) The heat map is symmetric across quadrants 1, 3 and 2,4. This is by design. As the model does not use the initial response and AI advice values directly – it uses their magnitudes (confidence) as well as a third, binary term indicating whether they agree on label – the integration model can only predict a delta change relative to the sign of the initial response. (2) The output has largest magnitude when the advice is confident and opposite the person's initial response. (3) if the advice is the same label as the person's initial response, the integration model output is largest when the person's initial response has low confidence and the model has high confidence. These observations confirm our integration model has reasonable operation behavior.

4 Modifying the AI advice

Here we discuss how we modify the raw AI advice output. First let's establish some notation. The notation described below is in reference to a single datapoint; later, we will add the subscript i to index across datapoints.

Let $A \in \mathbb{R}$ denote the inverse sigmoid of the advice presented to participants (i.e., we present the AI advice as $\sigma(A)$). We will work with A rather than $\sigma(A)$ here, as we will aim to replace the σ transform with an optimized function that improves human-AI performance.

Let $g : \mathbb{R} \to [0, 1]$ be a function (which we will optimize) that maps the AI advice to a probability that the user is shown. Note that for the baseline, "unmodified advice," we set g to be the sigmoid function,

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

Let r_1, r_2 be the initial and final responses from a participant (r_2 is a function of r_1, A, and other features), and let $y \in \{0, 1\}$ denote the true label. Let $\mathbf{u}$ represent demographic features of a person, and let $\mathbf{x} = (r_1, A, \mathbf{u})$ be the input feature vector to our $f_{\text{activation}}$ and $f_{\text{integration}}$, our activation and integration models for human behavior models described in Section 3. To simplify notation, we let $f_{\text{integration}}$ denote

the function that outputs the predicted r_2 for an activated person (rather than the delta from r_1 that we optimize for). Finally, let f_{HB} represent the entire model for human behavior. We also assume all feature preprocessing is hidden inside $f_{\text{activation}}$ and $f_{\text{integration}}$ for convenience of notation.

4.1 Optimizing for human-AI performance

Now that we have established the notation, we discuss how we optimize a function to modify the raw AI advice. First, we consider g of the form

$$g(A) = g_{\alpha,\beta}(A) = \frac{1}{1 + e^{-\text{sign}(A)(\alpha|A|+\beta)}}, \tag{1}$$

where $\alpha, \beta \in \mathbb{R}_{\geq 0}$. Note that setting $(\alpha, \beta) = (1, 0)$ results in $g_{1,0} = \sigma$, the sigmoid function used to produce the unmodified advice. Note that we do not exactly perform a linear transformation on A: the $|A|$ and $\text{sign}(A)$ terms are included to ensure that g does not change the label recommended as it makes the function symmetric around $A = 0$. α modulates the rate at which the presented advice increases with AI confidence, while β adjusts the minimum confidence level of the presented advice. For example, if $\beta = 1$, the presented confidence is > 0.73 (relative to the chosen label). In general, we can use other differentiable function for g; we chose the form in Equation 1 because it is easy to optimize, simple to understand, and is still flexible enough to fit the data.

As a shorthand for when we modify the features input to f_{activate} and $f_{\text{integrate}}$, we denote

$$g(\mathbf{x}) = (g(A), r_1, \mathbf{u}).$$

We now optimize for α, β to maximize the expected final accuracy using our human behavior model. The human behavior model discussed in Section 3 has the following mathematical representation:

$$f_{HB}(g(\mathbf{x})) = f_{HB}(g_{\alpha,\beta}(\mathbf{x})) = \begin{cases} r_1 & \text{w.p. } 1 - f_{\text{activate}}(g_{\alpha,\beta}(\mathbf{x})) \\ f_{\text{integrate}}(g_{\alpha,\beta}(\mathbf{x})) & \text{w.p. } f_{\text{activate}}(g_{\alpha,\beta}(\mathbf{x})) \end{cases}$$

We then solve the following optimization problem to obtain α, β:

$$(\alpha^*, \beta^*) = \arg\min_{\alpha,\beta} \mathbb{E}\left[L(f_{HB}(g(\mathbf{x})),\ y)|\mathbf{x}\right], \tag{2}$$

where L is the binary cross entropy loss function. Note that the expectation is taken with respect to the randomness in f_{HB} and the data distribution. So given training dataset D, we optimize the following expression:

$$\arg\min_{\alpha,\beta} \sum_{(\mathbf{x}_i, y_i) \in D} \mathbb{E}[L(f_{HB}(g(\mathbf{x}_i)),\ y_i)] =$$
$$\arg\min_{\alpha,\beta} \sum_{(\mathbf{x}_i, y_i) \in D} L(r_{1_i},\ y_i) \cdot (1 - f_{\text{activate}}(g(\mathbf{x}_i))) + L(f_{\text{integrate}}(g(\mathbf{x}_i)),\ y_i) \cdot f_{\text{activate}}(g(\mathbf{x}_i))$$

In practice, we carry out this optimization using stochastic gradient descent, which is possible as we chose f_{activate} and $f_{\text{integrate}}$ to be differentiable functions. Once we have optimized for α^*, β^*, we can apply it to new data without needing the user demographics $\mathbf{u}$ (i.e., we do not personalize the confidence transformation to individual users).

4.2 Original AI advice

Here we describe the AI advice used in our empirical and simulation results. Note that the previously described framework can work with any black-box AI model.

Since the datasets we are using are relatively small and curated primarily for human use rather than AI training, we opt to use aggregated human responses as a proxy for an AI model. In particular, we use the average the initial response of a previous group of participants as our black-box AI model. This allows us to ensure the advice given is "reasonable" in the sense that the advice is generally less confident on difficult tasks and more confident on easy tasks. We also check whether this advice is calibrated. In Figure 6, we show the calibration plot of the AI advice. We also compute the expected calibration error (ECE) Naeini et al. [2015]. The ECE for AI advice is 0.074, indicating the advice is roughly calibrated.

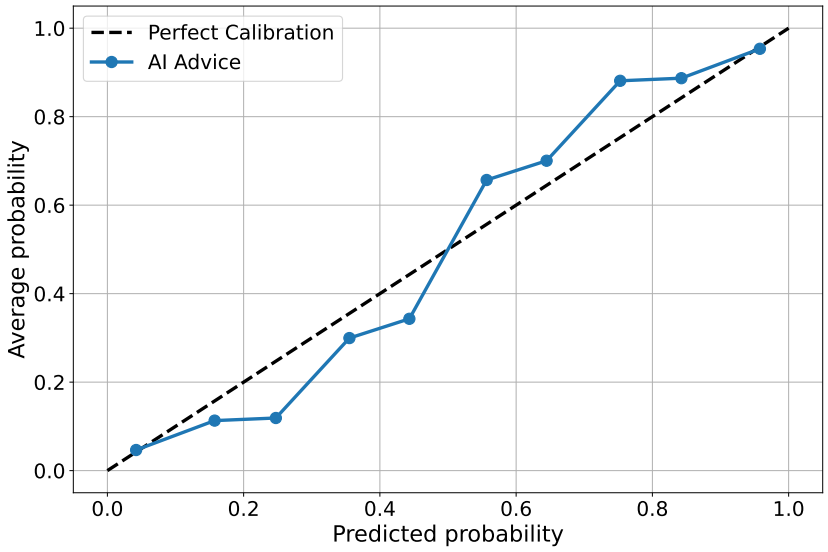


Figure 6: Calibration plot of the AI advice, aggregated across all 4 tasks. We note that the advice is roughly calibrated with an expected calibration error of 0.074.

Table 2: Simulated performance difference of modified advice ("Advice given" curve in Figure 7) to unmodified advice. Positive difference indicates higher value for modified advice. Results shown across 3 average advice accuracy levels.

Advice Accuracy	Final Accuracy	Correct Confidence	Activation Rate
75%	+0.7%	+0.097	+0.083
79%	**+1.9%**	**+0.114**	**+0.091**
85%	+3.1%	+0.131	+0.111

5 Results

5.1 Simulation results

Optimizing Equation 2 results in an optimal α^*, β^*. Using f_{HB}, we can then analyze the expected benefit of g_{α^*,β^*} over $g_{1,0}$. In Figure 7, we show g_{α^*,β^*}. We plot three separate curves. The curve labeled "Advice given" is optimized using the previously described dataset, and this is the function we use in our empirical results. We also modify the dataset by biasing the advice towards the correct label and adding noise to the advice to respectively increase and decrease the average advice accuracy in the dataset to 75% and 85%. We then fit a new (α, β) to the modified dataset and plot these curves. Note that f_{HB} is fit to the unmodified dataset and is fixed for optimizing all three of these curves.

For all three curves, we notice there is a large discontinuity at $A = 0.5$. This is a result of a large β^*, and suggests that advice is not useful to humans unless it is at a certain level of confidence (e.g., to make the human activated).

That all three curves follow relatively similar paths suggests that the performance of g_{α^*,β^*} may be robust to small changes in advice accuracy. This is confirmed (in simulation) in Table 2, where we show the simulated performance difference between modified and unmodified advice. We see here that the final accuracy, confidence in correct result, and activation rate all increase with the modified advice.

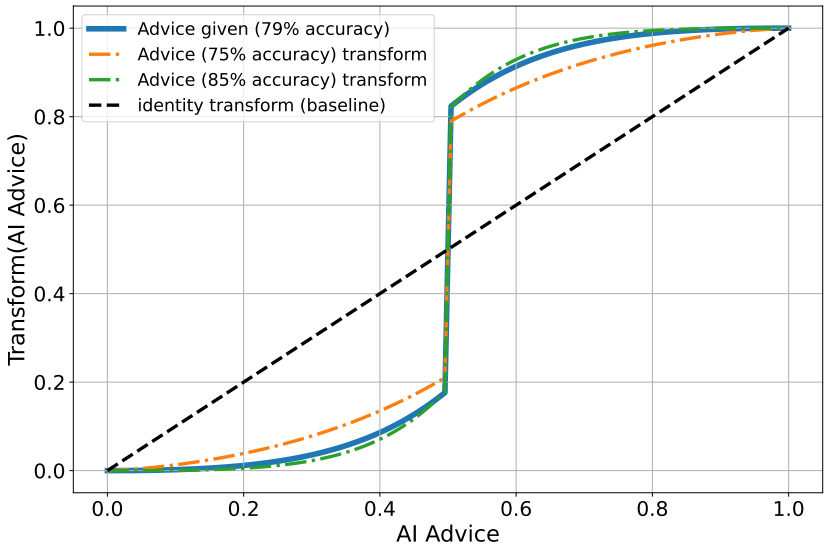
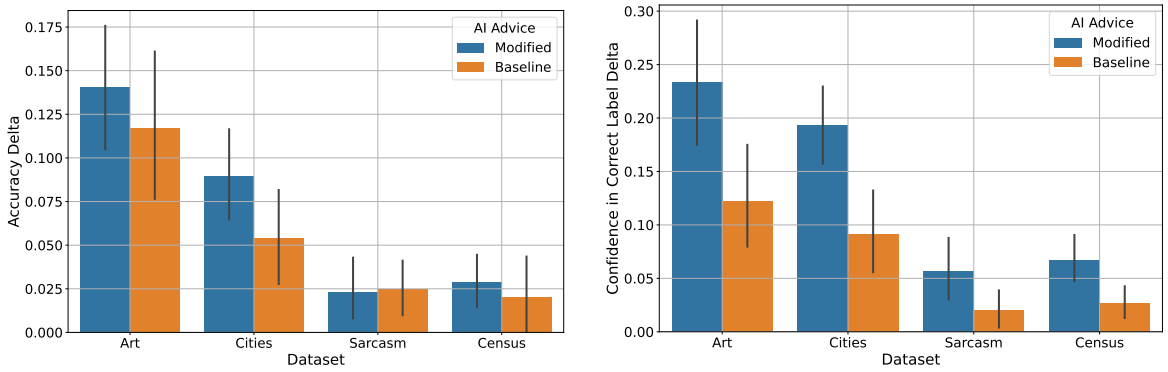


Figure 7: Our fitted modified AI advice vs. the original unmodified advice. We plot multiple fitted functions, where we vary the average advice accuracy in the training data. The curve labeled "Advice given" is what we actually show people in our experiments.

5.2 Human experiments



(a) Accuracy (b) Confidence

Figure 8: Our empirical results: we plot the average change between the first (before advice) and second (after advice) responses in (a) accuracy and (b) confidence of the correct label, averaged by participant. These results are consistent with our simulation findings of Table 2. The error bar represents ±1 standard deviation.

We verify our simulation results across the same four, previously described tasks. Adult, US-based participants were recruited through the Prolific crowdworker platform Prolific. We recruited 50 participants for each task, and randomly assigned either the unmodified (baseline) or modified advice to each person. In Figure 8, we show the average change in (a) accuracy and (b) confidence of the correct label (the average confidence, with correct and incorrect labels receiving positive and negative weight respectively) across all 4 tasks, partitioning by the advice received by participants. Accuracy and confidence increase across all tasks when using any advice. This increase is significant for confidence (Two-Sample Student's t-test, $p = 3.84 \times 10^{-6}$). Moreover, we can largely confirm our simulation results. Accuracy and confidence increased due to our modified AI advice across nearly all tasks. Additionally, activation rate (not plotted) increases with the modified advice, with an average increase of 3.3% across tasks due to the modified advice.

Table 3: Empirical performance on UK-based participants. Showing the difference in metric between modified advice and baseline (value is > 0 when modified advice resulted in a larger value). We recruited 50 participants for each task.

Task	Final Accuracy	Correct Confidence	Activation Rate
Art (90%)	+8.5%	+0.150	+0.101
Cities (87%)	+1.7%	+0.140	+0.169
All (89%)	**+5.2%**	**+0.149**	**+0.143**

Another interesting finding is that tasks with higher advice accuracy (Art and Cities) exhibit larger increases in accuracy and confidence. The AI advice is more accurate in Art and Cities (accuracy: 90% and 87%) compared to Sarcasm and Census (accuracy: 78% for both). This intuitively makes sense – we are making the advice more confident in its predictions, and so while this benefits all tasks, it disproportionately benefits tasks where the advice is more accurate. This effect was predicted by our simulation results.

Our primary human experiments involved US-based participants. To show our findings generalize, we took the advice modification learned on the US data and applied it to new participants based in the UK for the Art and Cities tasks. The findings are highly consistent: modified advice increases activation and improves final human accuracy and confidence (Table 3). Combining US and UK data, performance improvement is significant for both accuracy ($p = 8.86 \times 10^{-3}$) and confidence ($p = 4.43 \times 10^{-11}$).

6 Discussion and Future Work

In this paper, we proposed optimizing a simple function to modify the reported confidence of AI advice used in a collaborative human-AI system. To optimize the reported confidence, we learned a human behavior model from experimental data that allowed us to simulate the full human-AI system. The modified AI advice are explicitly uncalibrated, but as we showed in both simulated and empirical results, it resulted in substantially higher final human performance. We further validated our model using new participants from a different country.

These results highlight the importance of recognizing that AI, when deployed in collaborative systems, does not operate in isolation and should not be optimized under that assumption. While we proposed one method of optimizing the AI algorithm with consideration for the human part of the system, there are many avenues to explore in future work. One limitation of this work is that the experiments we run have the average AI advice accuracy higher than the average human accuracy. Though this is a reasonable assumption in many cases, it may not always hold true. It is important to understand whether making a model uncalibrated is always beneficial, or if there are modes of operation where calibrated advice is indeed preferred.

An interesting modification to our setup is to give full access to the AI advice rather than treating it like a black box as we do here. In this setting, we can imagine directly optimizing the algorithm with respect to the loss after applying the human behavior model to the algorithm's advice. A thorough analysis of the resulting model and how it differs from baseline (an algorithm optimized independently of the human component) would likely provide interesting insights into human-AI systems. Some prior work attempts to take this step, but does not factor in an empirically validated model for human behavior Wilder et al. [2020], Bansal et al. [2021]. We believe our method of explicitly modeling human behavior using empirical data may be important for achieving gains in real-world performance.

Finally, we recognize that the human behavior model we used in this work does not capture all real-world settings. It may be important to consider different models for different applications, and our results will need to be validated in these new settings.

References

Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414, 2021.

Erin K Chiou and John D Lee. Trusting automation: Designing for responsivity and resilience. *Human Factors*, page 00187208211009995, 2021.

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

Robin C Feldman, Ehrik Aldana, and Kara Stein. Artificial intelligence in the health care space: how we can trust what we cannot know. *Stan. L. & Pol'y Rev.*, 30:399, 2019.

Francesca Gino and Don A Moore. Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, 20(1):21–35, 2007.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Nigel Harvey and Ilan Fischer. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes*, 70(2):117–133, 1997.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Poornima Madhavan and Douglas A Wiegmann. Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors*, 49(5):773–785, 2007.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *arXiv preprint arXiv:1711.06664*, 2017.

Neda Mesbah, Christoph Tauchert, and Peter Buxmann. Whose advice counts more–man or machine? an experimental investigation of ai-based advice utilization. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 4083. ScholarSpace, 2021.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38, 2019.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Dilek Önkal, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, 22(4):390–409, 2009.

Andrew Prahl and Lyn Van Swol. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6):691–702, 2017.

Prolific. Prolific academic. https://www.prolific.co, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

Sunita Sah, Don A Moore, and Robert J MacCoun. Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes*, 121(2):246–255, 2013.

Thomas Schultze, Anne-Fernandine Rakotoarisoa, and Stefan Schulz-Hardt. Effects of distance between initial estimates and advice on advice utilization. *Judgment & Decision Making*, 10(2), 2015.

Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551, 2021.

Lyn M Van Swol, Jihyun Esther Paik, and Andrew Prahl. Advice recipients: The psychology of advice utilization. In *The Oxford handbook of advice*, page 21–41. Oxford University Press, 2018.

Kailas Vodrahalli, Roxana Daneshjou, Roberto A Novoa, Albert Chiou, Justin M Ko, and James Zou. Trueimage: A machine learning algorithm to improve the quality of telehealth photos. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 220–231. World Scientific, 2020.

Kailas Vodrahalli, Tobias Gerstenberg, and James Zou. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. *arXiv preprint arXiv:2107.07015*, 2021.

Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. " do you trust me?" increasing user-trust by integrating virtual agents in explainable ai interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 7–9, 2019.

Alex West and Anusha Praturu. Enhancing the census income prediction dataset. https://people.ischool.berkeley.edu/~alexwest/w210_census_income_html/, 2019. Accessed: 2021-05-15.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020.

Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM, 2020.

Appendix

A Features for Human behavior Model

Here we detail the features used for our activation and integration models ($f_{\text{activation}}$ and $f_{\text{integration}}$. Features and a description are given in Table 4.

Table 4: Description of input features for $f_{\text{activation}}$ and $f_{\text{integration}}$ models.

Feature #	Value Range	Description
1	$[0, 1]$	Response 1 Confidence
2	$[0, 1]$	Advice Confidence
3	$\{-1, +1\}$	$+1$ if Advice and Response 1 agree on label, otherwise -1
4	$[-1, 1]$	Feature 1 $\cdot$ Feature 3
5	$[-1, 1]$	Feature 2 $\cdot$ Feature 3
6	$[-1, 1]$	Survey question measuring human's perception of AI performance on the given task
7	$\mathbb{R}_{\geq 18}$	Age
8	$\{0, 1\}$	Sex
9	$\{0, 1\}$	Does the person have prior experience with computer programming?
10	$[1, 10]$	Self-reported socioeconomic status
11	$[0, 1]$	Survey question assessing perceived presence of AI in person's life
12	$[1, 8]$	Self-reported education level

B More Details on Crowd-Source Data Collection

Here we include additional details on the crowd-sourced data collection procedure we run.

B.1 Demographic Information Collection

Our experiments were run on Prolific Prolific. The demographic information used by our activation-integration model was provided by Prolific. Education level is defined on a scale from 1 to 8, and should be interpreted as:

1 – Don't know / not applicable

2 – No formal qualifications

3 – Secondary education (e.g. GED/GCSE)

4 – High school diploma/A-levels

5 – Technical/community college

6 – Undergraduate degree (BA/BSc/other)

7 – Graduate degree (MA/MSc/MPhil/other)

8 – Doctorate degree (PhD/other)

Socioeconomic status is defined on a scale from 1 to 10 and was assessed by asking participants to answer the question shown in Figure 9.

Socioeconomic Status

Participants were asked the following question: Think of a ladder (see image) as representing where people stand in society. At the top of the ladder are the people who are best off—those who have the most money, most education and the best jobs. At the bottom are the people who are worst off—who have the least money, least education and the worst jobs or no job. The higher up you are on this ladder, the closer you are to people at the very top and the lower you are, the closer you are to the bottom. Where would you put yourself on the ladder? Choose the number whose position best represents where you would be on this ladder.

Figure 9: Socioeconomic status question.

B.2 Survey Questions

Participants were asked two survey questions. We used their responses in our activation and integration models. The questions were:

1. Do you think the AI or the average person (without help) can do better on this task?
2. How often do you use AI systems to aid you in your everyday life and/or at work?

Participants responded to each question on a slider scale. The first question was intended to measure participant's perceived confidence in the AI prior to doing any tasks. The second question was intended to measure the participant's familiarity and comfort with AI.

B.3 Participant Compensation

We compensated participants at $\geq$ \$10.00 per hour (depending on bonus pay) as per the recommended rates by Prolific. We informed participants of the possibility for bonus pay. Bonus was calculated as:

$$\text{bonus} = \begin{cases} 0 & \text{if } S < 0.3 \\ S * 0.3 & \text{otherwise} \end{cases}.$$

Here, S is the average performance, computed as (the average of)

$$sign(\text{correct response}) \cdot \text{response}_1,$$

where $-1 \leq \text{response}_1 \leq 1$.

The total cost of collecting the data for training our human behavior model (Section 3) was around \$600. The total cost of collecting data for our human experiments (Section 5.2) was around \$750.