Do Humans Trust Advice More if it Comes from AI? An Analysis of Human-AI Interactions

Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou
Stanford University, Stanford, USA

ABSTRACT
In decision support applications of AI, the AI algorithm’s output is framed as a suggestion to a human user. The user may ignore this advice or take it into consideration to modify their decision. With the increasing prevalence of such human-AI interactions, it is important to understand how users react to AI advice. In this paper, we recruited over 1100 crowdworkers to characterize how humans use AI suggestions relative to equivalent suggestions from a group of peer humans across several experimental settings. We find that participants’ beliefs about how human versus AI performance on a given task affects whether they heed the advice. When participants do heed the advice, they use it similarly for human and AI suggestions. Based on these results, we propose a two-stage, “activation-integration” model for human behavior and use it to characterize the factors that affect human-AI interactions.

KEYWORDS
artificial intelligence, AI advice, human-in-the-loop, human interaction with AI, AI trust

1 INTRODUCTION
In safety-critical areas like medicine, there remain significant barriers to widespread adoption of AI. For example, due to the potentially debilitating or even fatal consequences of misdiagnosis, the healthcare system has been wary of implementing AI diagnostic algorithms with unknown failure modes.

One solution, particularly relevant for medicine, is to develop decision support systems with “humans-in-the-loop”; in this mode, humans treat the output of AI algorithms as “advice”, while ultimately making the final decision themselves. The advice becomes another piece of information to aid in the decision-making process, similar to ordering additional lab tests or consulting a colleague. While this mitigates issues related to safety, one of the biggest barriers to adoption remains the black-box nature of current AI systems that limits people’s trust in their advice [6, 19, 23, 30].

In this paper, we seek to better characterize how humans use advice from an AI algorithm. In particular, we ask whether people use advice differently if it comes from a human versus an AI. To answer this question, we employ the judge-advisor paradigm (JAS) from psychology to compare how people use advice from different sources [21, 28]. We developed several experiments for a layperson audience and deployed these experiments on a crowdsourcing platform. Our findings suggest that the advice source (human or AI) affects how likely the individual is to use the advice. Individuals use the advice to the extent that they believe the AI or other humans are good at the given task. However, if an individual chooses to use the advice, there is little difference in how they incorporate human advice versus AI advice. We found similar results when we tested this paradigm on experts—dermatologists—who are given the task of deciding whether or not to biopsy a lesion.

Our contributions. Understanding how users account for AI advice is an important aspect of human-AI interaction and has been under-explored. Here we systematically study how humans use advice in several experiments with both laypeople and an expert group.

- We propose the “activation-integration” model for human behaviour in human-in-the-loop systems, where humans first decide whether to use advice and subsequently decide how to update their judgments.
- We quantify how different demographic and task-related factors contribute to a person’s response to advice. We identify
three key factors related to the task instance – (1) the person’s confidence, (2) the advice confidence, and (3) whether the person’s response agrees with the advice – and one key factor related to the task in general – the person’s prior belief of human vs. AI performance. Other demographic and task-related factors do not have any consistent, significant effect across the variety of settings we test our model under.

- We ran extensive experiments to support our hypotheses. We collected data from a variety of tasks, participants living in a variety geographic regions (US, UK, and Asia), and a range of perceptions of advice quality.
- We release the data we collected for general use by the research community.1

Related work. Advice utilization has been studied in the psychology literature where studies often employ JAS for comparing how humans use advice from various sources when completing a given task [28]. This setup is similar to ours: people are given a task, asked to complete the task, and subsequently receive advice from an outside source. By comparing various types of advice, it is possible to draw conclusions on how different sources of advice are utilized. Previous work has attempted to understand the effects of peer vs. expert advice [17], human vs. algorithmic advice [5, 17, 20, 21], task difficulty [9], and advice confidence [25, 26] on human utilization of advice. New work has also begun to explore AI advice. In particular, how perception of an AI can affect trust and utilization of its advice [16] and advice utilization of an AI relative to peer or expert humans [8, 18, 27].

The findings so far are inconsistent across settings, suggesting that many aspects of the task and the advice presentation affect advice utilization [14]. Some studies show people utilize automated or AI advice more than advice from peers [5, 17, 18, 27], while others have mixed results where AI advice has less or similar utilization [8, 21]. Most studies use student populations in their experiments, with recruited study participants often US-based. All studies have limited sample sizes, generally with n ≤ 300. Additionally, studies typically investigate a single data modality like image-based predictions [5, 8, 9, 17, 25] or forecasting predictions (including future stock price prediction) [20, 21, 27].

We seek to extend this line of work. We ran a comprehensive set of experiments across multiple geographic regions, data modalities including image-based tasks, text-based tasks, and tabular data-based tasks, and a variety of experiment settings varying participants’ beliefs about the quality of the advice given. To run this set of experiments, we recruited over 1100 crowdworkers. This allows us to make more general conclusions about how AI advice is used relative to human advice. In particular, we hypothesize and validate the effectiveness of a two-stage model for human utilization of advice, building on prior work that suggested humans may process advice in two stages [20]. Furthermore, we attribute a human’s prior belief of the advice source’s suitability for a given task as a key factor for determining how receptive the person will be to advice. We also identify factors that are generally important for advice utilization, supporting findings from previous work [9, 25, 26].

2 STUDY DESIGN

We ran multiple experiments to investigate how people use AI advice. All of these experiments follow the same basic design.

![Figure 1: Visualization of our study design.](image)

2.1 Overview

To investigate whether participants use advice from an AI algorithm differently than from humans, we developed a set of experiments on a variety of tasks (Section 2.3). The experiments are setup as the 2-stage procedure shown in Figure 1. Participants were recruited from a crowdsourcing platform and randomly assigned into two groups. The first group received advice from a human while the second group received advice from an AI algorithm. Participants were adults from the US. We also recruited participants from the UK and Asia for a subset of the experiments (Section 3.4.1).

To isolate the advice source’s effect on people’s judgments, the AI-labeled advice and human-labeled advice were identical. The advice was generated using the average response from a previous set of ~50 human labelers who completed the tasks without advice. The resulting advice has an accuracy of 80 ~ 90%, depending on the task. In each condition, participants were informed that the advice source was 80% accurate, and the actual advice participants received was identical in both conditions. Between the conditions, we only varied whether the advice was given by a human versus AI and how the advice was presented by using a small icon of a person or a computer (see Figure 1). When showing the advice to participants, we randomly perturbed the advice by 3 fixed amounts: (1) 20% of the scale width away from the correct response, (2) 20% of the scale width towards the correct response, or (3) no change. Options (1) and (2) each had a 25% probability of being selected, while option (3) had 50% probability of being selected. We added these perturbations as a means for both increasing the variety of advice given and decreasing the accuracy of the advice to <=80%. The expected advice accuracy is given in Table 1. The randomness in these perturbations is specific to each participant. Note that participants did not receive feedback on their or the advice’s performance on individual questions during the experiment.

After reading through instructions, each participant completed a series of tasks in randomized order. Within each experiment, the tasks were all of the same type (e.g., image classification on

---

1https://github.com/kailas-v/human-ai-interactions
city images), but the difficulty of each task instance could vary significantly (see Table 1). All tasks are binary classification tasks. Users submitted answers in the form of a confidence on a continuous sliding scale with endpoints labeled “Definitely [Label A]” and “Definitely [Label B]”.

Participants were incentivized to perform well on the task by receiving a bonus based on their averaged judgments across all task instances (both before and after advice). Participants were informed of how the bonus was calculated in the instructions. After completing all tasks in an experiment, participants were asked a series of survey questions to gauge their trust in and familiarity with AI. More details on the instructions and the survey questions are provided in Appendix E.

2.2 Survey Question About Prior

The prior belief value reported in Table 1 was obtained using a survey question asked to every participant after they had finished all task instances. We also ran additional experiments to verify that the location of this survey question had no affect on our analysis and conclusions (see Appendix E.2). The question asked is shown below. The wording was slightly modified for those who received AI advice or human advice to account for the advice they received. The ordering of AI and human in the question was randomized for each participant.

**Human:** “Do you think an artificial intelligence (AI) algorithm or the average person (without help) can do better on this task?”

**AI:** “Do you think the AI or the average person (without help) can do better on this task?”

2.3 Tasks

We provide an overview of the datasets we use for our experiments in Table 1. The socioeconomic score and education level are provided by the crowdsource platform we use for running our experiment (Prolific [22]). For each experiment, every participant completes all tasks from a given dataset twice – before and after receiving advice. More details about each dataset are given below. In Figure 2, we show sample tasks from each dataset. These were the examples shown to participants in the instructions and were selected to be easier than the tasks we use in our experiments. The tasks were designed to cover diverse data modalities (visual, text and tabular) as well as sufficiently challenging so that participants can benefit from advice.

**Art dataset (Image).** This dataset contains images of paintings from 4 art periods: Renaissance, Baroque, Romanticism, and Modern Art. The dataset contains 8 paintings from each time period. Participants were asked to determine the art period a painting is from given a binary choice. The incorrect label was selected to be from adjacent time periods to increase task difficulty (e.g., if Romanticism is the correct label, then the alternate choice would be Baroque or Modern Art; this choice was fixed for all participants). An example task is shown in Figure 2a.

**Cities dataset (Image).** This dataset contains images from Google from 4 major US cities: San Francisco, Los Angeles, Chicago, and New York City. The dataset contains 8 images from each city. The
Table 1: Summary of our 4 laypeople datasets. Baseline accuracy describes task instance-level performance prior to advice (average +/- standard deviation). Prior belief indicates whether participants favored AI or humans; see Section 2.2 for details.

| Dataset             | Data Type | Description            | # of Tasks | Baseline accuracy | Advice accuracy | Prior belief |
|---------------------|-----------|------------------------|------------|------------------|----------------|--------------|
| Art Dataset         | Image     | identify the art period | 32         | 65.7% +/- 17.0%  | 76.6%          | 66.7% (AI)   |
| Cities Dataset      | Image     | identify the city      | 32         | 72.1% +/- 17.1%  | 80.5%          | 62.0% (AI)   |
| Sarcasm Dataset     | Text      | identify sarcasm       | 32         | 72.9% +/- 20.3%  | 78.9%          | 80.0% (human) |
| Census Dataset      | Tabular   | identify income level  | 32         | 70.4% +/- 28.8%  | 73.4%          | 66.0% (AI)   |

3 RESULTS

In Section 3.1, we establish that both AI and human advice are helpful across our tasks. In Section 3.2, we identify the key difference between human and AI advice and propose our activation-integration model to model the observed human behavior. In Section 3.3, we demonstrate the reason for a difference between human and AI advice is largely due to the prior beliefs of an individual. Finally, we validate our activation-integration model and prior belief results across geographic locations and study conditions in Section 3.4 and with expert participants using a dermatology task in Section 3.5.

The results in Sections 3.1, 3.2, and 3.3 are based on the data summarized in Table 2. The data used in Sections 3.4 and 3.5 are described in those sections.

3.1 Effect of advice

First, we consider the effect of receiving any advice on performance. This effect is summarized in the 1st through 3rd rows of Table 3. The 1st row reports the average accuracy across tasks before receiving advice, and the 2nd and 3rd rows report the average change in accuracy after receiving advice. In all four datasets, participants’ responses are more accurate after having received advice in both human and AI conditions. We visualize the average accuracy difference across individual task instances in Figure 3.

3.2 Difference between AI and human advice

Now we ask how this effect differs between human versus AI advice. To help answer this question, we will use the weight of advice (WoA), a common metric in the psychology of advice utilization [11, 31]. The WoA is defined as

\[ WoA = \frac{\text{response}_2 - \text{response}_1}{\text{advice} - \text{response}_1}, \]

where response_1 and response_2 are the participant’s first and second answers respectively, and advice is the advice presented to the participant. WoA is typically clipped to have bounded magnitude (e.g., in the case that the first response happens to be close to the advice given). In the literature, WoA is typically clipped to have a maximum magnitude of 1 (i.e., perfectly following the advice). However, we opt to clip the WoA to have maximum magnitude of 5, allowing us to capture the possibility that people become more confident than the advice after integrating with their own knowledge. A positive value indicates a participant changing their response in accordance with the advice, while a negative value implies the participant changed their response away from the advice. A response of 0 occurs when the participant did not change their response. We define a user to be “activated” for a given task if they change their response by at least a threshold amount after receiving advice. This threshold was chosen to be equal to the width of the slider knob (3.5% of the total slider bar width).

We plot the probability density function of the WoA for each dataset, comparing across conditions in Figure 4. From these plots, we make the following observations:

1. There is a statistically significant difference in how often participants modify their initial response when given AI versus human advice. The magnitude and direction of this difference varies between datasets. For Art, Cities, and Census, the AI condition evokes a larger response; for Sarcasm, the human condition evokes a larger response. We describe this effect as an “activation difference,” where the condition that people are more likely to use advice from is the more balanced across race and gender, biased participants would actually receive negative feedback as they would necessarily receive a low income level. For each image, an alternate city was randomly selected as the negative label presented to participants. The images were selected to be moderately difficult to identify – images of major landmarks were excluded, but differentiating aspects of the cities like architecture were included. An example task is shown in Figure 2b.

Sarcasm dataset (Text). This dataset is a subset of the Reddit sarcasm dataset [15], which includes text snippets from the discussion forum website, Reddit. Specifically, the subset was selected from posts in the “AskReddit” forum and were hand-filtered to exclude posts that either contained potentially offensive content or were too long. Survey participants were asked to identify sarcasm in the text snippet. The dataset was balanced to contain 16 sarcastic posts and 16 non-sarcastic posts. An example task is shown in Figure 2c.

Census dataset (Tabular). This dataset comes from a subset of US census data [29]. The task is to identify an individual’s income level (does the individual make $\geq$ $50,000 annual income or not?) given their demographic information: state of residence, age, education level, marital status, gender, and race. These features were selected based on previously published algorithms for predicting income level [3, 4]. The dataset was balanced to contain 16 individuals who make $\geq$ $50,000 and 16 who make < $50,000 annual income. The dataset was further balanced across race and gender and each binary income category contains 4 Asian, 4 Black, and 8 White individuals. An example task is shown in Figure 2d.

The census dataset may raise concerns of reinforcing gender or race-related biases in participants. However, as income levels were balanced across race and gender, biased participants would actually receive negative feedback as they would necessarily receive a low bonus payment due to their bias.

3.1 Effect of advice

First, we consider the effect of receiving any advice on performance. This effect is summarized in the 1st through 3rd rows of Table 3. The 1st row reports the average accuracy across tasks before receiving advice, and the 2nd and 3rd rows report the average change in accuracy after receiving advice. In all four datasets, participants’ responses are more accurate after having received advice in both human and AI conditions. We visualize the average accuracy difference across individual task instances in Figure 3.

3.2 Difference between AI and human advice

Now we ask how this effect differs between human versus AI advice. To help answer this question, we will use the weight of advice (WoA), a common metric in the psychology of advice utilization [11, 31]. The WoA is defined as

\[ WoA = \frac{\text{response}_2 - \text{response}_1}{\text{advice} - \text{response}_1}, \]

where response_1 and response_2 are the participant’s first and second answers respectively, and advice is the advice presented to the participant. WoA is typically clipped to have bounded magnitude (e.g., in the case that the first response happens to be close to the advice given). In the literature, WoA is typically clipped to have a maximum magnitude of 1 (i.e., perfectly following the advice). However, we opt to clip the WoA to have maximum magnitude of 5, allowing us to capture the possibility that people become more confident than the advice after integrating with their own knowledge. A positive value indicates a participant changing their response in accordance with the advice, while a negative value implies the participant changed their response away from the advice. A response of 0 occurs when the participant did not change their response. We define a user to be “activated” for a given task if they change their response by at least a threshold amount after receiving advice. This threshold was chosen to be equal to the width of the slider knob (3.5% of the total slider bar width).

We plot the probability density function of the WoA for each dataset, comparing across conditions in Figure 4. From these plots, we make the following observations:

1. There is a statistically significant difference in how often participants modify their initial response when given AI versus human advice. The magnitude and direction of this difference varies between datasets. For Art, Cities, and Census, the AI condition evokes a larger response; for Sarcasm, the human condition evokes a larger response. We describe this effect as an “activation difference,” where the condition that people are more likely to use advice from is the more
Table 2: Participant demographics for our initial experiments. Socioeconomic score ([1 – 10]): 5 corresponds to a middle class. Education level ([1 – 8]): 5 corresponds to an associate’s degree. See Appendix E.1 for details.

| Dataset  | # of Participants | Gender (Percent Male / Female) | Age           | Socioeconomic Score | Education Level |
|----------|-------------------|--------------------------------|---------------|---------------------|-----------------|
| Art      | 147               | 47.6% / 52.4%                  | 33.0 +/- 12.3 | 5.2 +/- 1.5         | 5.4 +/- 1.3     |
| Cities   | 94                | 50.0% / 50.0%                  | 26.1 +/- 7.8  | 5.3 +/- 1.5         | 5.4 +/- 1.2     |
| Sarcasm  | 97                | 43.3% / 56.7%                  | 31.2 +/- 11.7 | 5.3 +/- 1.6         | 5.2 +/- 1.4     |
| Census   | 98                | 46.9% / 53.1%                  | 30.5 +/- 10.9 | 5.2 +/- 1.6         | 5.3 +/- 1.2     |

Figure 3: Average accuracy before and after receiving advice. chance is the accuracy expected from random guessing; advice is the average accuracy of the advice; before advice and after advice are the average accuracies across both arms of the experiment.

activated condition. The magnitude and direction of this difference is shown in the “Δ” row of the Activation Rate sub-table in Table 3.

Also note that the sign of the activation difference is reflected in the change in accuracy across the human and AI subsets. For the datasets where AI had a higher activation rate, the participants receiving AI advice had a greater average change in accuracy. On the sarcasm dataset, the change in accuracy is identical between those receiving human versus AI device. While this may seem contradictory (there is a non-zero activation difference), Figure 3 provides an explanation. Note that in the Sarcasm dataset, there are more tasks where the advice is incorrect than in the other three datasets. So, the accuracy improvement we would expect to see is mitigated by bad advice decreasing accuracy. This explanation can also be applied to the Census dataset which has a smaller accuracy difference between AI and human than the Art or Cities dataset.

(2) Conditioned on having changed their label, people behave similarly. As shown in Figure 4, this subset of the population corresponds to between 34% and 54.5% of the responses depending on the experiment. We ran a 2 sample Kolmogorov–Smirnov test for each set of tasks with a p-value threshold of 0.01, including all responses where the label was changed after receiving advice and splitting on the treatment arm to get 2 samples. In all 4 datasets, there was
Figure 4: PDF of the weight of advice (WoA), conditioned on being activated. WoA is clipped to have a maximum magnitude of 5. The % activated refers to the corresponding value in Table 3.

Table 3: Treatment effect (Sections 3.1, 3.2, and 3.3). Accuracy Δ is the accuracy change after receiving advice. "Activation Rate" is the percentage of participants who change their response after receiving advice. Activation Δ is the difference between AI and human activation.

| Dataset  | Art | Cities | Sarcasm | Census |
|----------|-----|--------|---------|--------|
| Accuracy |     |        |         |        |
| Before Advice | 65.7% | 72.5% | 73.1% | 70.5% |
| Δ Human | +7.1% | +3.7% | +3.3% | +2.4% |
| Δ AI | +11.8% | +6.2% | +3.3% | +3.5% |
| Activation Rate |     |        |         |        |
| Human | -46.0% | -48.2% | -39.8% | -41.8% |
| AI | -52.3% | -54.5% | -34.0% | -47.5% |
| Δ | +6.6% | +6.3% | -5.8% | +5.7% |

no statistically significant difference between the samples, suggesting the main effect of the treatment is to modulate the activation rate.

Activation-integration model. Based on these experiments, we propose a two stage model, which we term the activation-integration model, for how participants utilize the advice they receive. In the first (activation) stage, a participant decides whether or not to use the advice. In the second (integration) stage, they decide to what extent they will use the advice. There is evidence in the literature for such a model. Prior work has found evidence that humans process advice in two stages [20]. There is also evidence that people are better at assessing advice than using it [12]. Our activation-integration model captures this effect nicely.

Using this model, we can understand that receiving AI versus human advice affects the activation stage, but not the integration stage. In particular, the rate at which people utilize advice is modulated by the advice source, but the extent to which people utilize advice is not significantly different if they decide to use the advice.

3.3 Effect of participants’ prior beliefs

Now we seek to explain what accounts for the treatment effect we observed. The final column of Table 1 offers an explanation: it shows participants’ answers to a survey question about whether an AI algorithm or the average human would perform better on the given task. Responses were recorded on a sliding scale with human at one end and AI at the other end. We computed the prior by averaging all responses and rescaling to a range of $-100\%$ to $+100\%$. We then set the sign such that a positive prior corresponds to the participant’s belief that the advice source they received would perform better than the alternate source. We visualize the full distribution in Appendix B.

We use 2 mixed effects linear models to test our activation-integration model and show the effect of the prior. The list of
Table 4: Confidence levels of (1) a logistic regression mixed effects model predicting activation and (2) a linear regression mixed effects model predicting integration response. Coefficients printed bold if non-zero (significant with $\alpha = 0.95$), $\beta_{consistent}$ refers to whether the initial response and the advice favor the same label. The $\beta_{response}$ and $\beta_{advice}$ refer to the magnitude of the response and advice respectively.

| Covariates               | Activation Coefficients | Integration Coefficients |
|--------------------------|-------------------------|--------------------------|
| Given AI Advice? ($\beta_{ai}$) | 0.162 +/- 0.134         | -0.023 +/- 0.016         |
| Prior Belief ($\beta_{prior}$) | 0.244 +/- 0.059         | 0.001 +/- 0.008          |
| Response 1 Confidence ($\beta_{response}$) | -0.584 +/- 0.025        | -0.212 +/- 0.004         |
| Advice Confidence ($\beta_{advice}$) | 0.218 +/- 0.024         | 0.025 +/- 0.004          |
| Response 1 and Advice Consistent? ($\beta_{consistent}$) | -1.555 +/- 0.051         | 0.661 +/- 0.008          |
| Age ($\beta_{age}$) | -0.089 +/- 0.072         | -0.026 +/- 0.008         |
| Gender ($\beta_{gender}$) | -0.046 +/- 0.137         | 0.010 +/- 0.016          |
| Education ($\beta_{education}$) | 0.043 +/- 0.076         | 0.021 +/- 0.009          |
| Socioeconomic Status ($\beta_{socioeconomic}$) | 0.139 +/- 0.070         | 0.007 +/- 0.008          |
| Has Programming Experience? ($\beta_{programming}$) | 0.233 +/- 0.147         | 0.014 +/- 0.017          |

3.4 Additional validation studies

3.4.1 Varying geographic location. As perception of AI may be tied to geographic location [10], it is important to question whether our results generalize to across geographic regions. To answer this question, we recruited approximately 200 participants from the UK and 200 participants from Asia and ran experiments with our art and sarcasm datasets. Note that the studies were run identically to those given to US-based participants, with the exception that we now recruit participants exclusively from the corresponding geographic region. More details including participant demographics are included in the Appendix C.

The main finding is that our activation-integration model is largely validated in the UK and in Asia. In the activation model, the prior is associated with activation in the UK. $\beta_{prior, UK} = 0.217$ is similar to the $\beta_{prior, US} = 0.244$, suggesting the two geographical groups are similar. While we do not have sufficient power to say $\beta_{prior, Asia} = 0.101$ is associated with activation, its positive value is in line with the UK and US prior coefficients. $\beta_{response, UK} = 0.584$, $\beta_{response, Asia} = 0.355$, $\beta_{advice, US} = 0.218$, $\beta_{advice, UK} = 0.146$, $\beta_{consistent, US} = 1.555$, $\beta_{consistent, UK} = 1.157$, $\beta_{consistent, Asia} = 1.492$ are associated with activation and are similar in sign and magnitude across all three regions. Gender, education, age, and programming experience are not associated with activation probability across all three regions.

In the integration model, $\beta_{response}$ ($\beta_{response, US} = 0.212$, $\beta_{response, UK} = 0.199$, $\beta_{response, Asia} = 0.226$) and $\beta_{consistent}$ ($\beta_{consistent, US} = 0.661$, $\beta_{consistent, UK} = 0.589$, $\beta_{consistent, Asia} = 0.674$) are associated with integration and have similar sign and magnitude across all three regions. Neither treatment nor prior are associated with integration across all three regions, validating our earlier conclusion that the prior has a strong effect in determining activation, but a weak effect in utilization of advice after activation.

3.4.2 Varying perceived accuracy. We also quantified the effect of the advice’s perceived accuracy on its utilization. In previous experiments, we informed participants that the advice has an accuracy of 80%, for both AI and human advice arms. Here, we varied this value...
While the 4 tasks described in Figure 2 are ideal for crowdsourced integration model are largely consistent across the three groups. We observe the previous findings relating to our activation-integration model using both the start-of-survey and after all tasks are completed for a small subset of our participants. We then check whether there is a statistically significant systematic way by asking the prior belief question both before and after all tasks are completed for a small subset of our participants. We then check whether there is a statistically significant systematic change in prior belief value. We also fit the coefficients of our activation-integration model using both the start-of-survey responses and the end-of-survey responses. We observed no systematic change in prior belief. See Appendix E.2 for more details.

3.5 Real-world medical-AI task

While the 4 tasks described in Figure 2 are ideal for crowdsourced data collection as they are accessible to the general public, it is unclear how well the observed results generalize to real-world tasks with experts (e.g., an AI medical assistant). To validate our activation-integration model on more realistic applications, we designed an additional task involving real medical data and recruited clinicians for our experiment. Additionally, we use a real AI model that we train to provide advice for the clinicians. We use the same experiment setup described in Figure 1.

The task is to determine whether a skin lesion should be biopsied or not based on a single image of the lesion (i.e., is the lesion likely malignant or benign); a sample task instance is shown in Figure 5. The images used in the study were selected by a board-certified dermatologist from the International Skin Imaging Collaboration (ISIC) database. For the advice source, we trained a ResNet-18 model [13] to predict whether a lesion is malignant or benign on the ISIC dataset [24]. We then took the ResNet model output (after softmax) as our advice source. Across the data sample we show in our task, the model is well-calibrated with an expected calibration error of 0.068.

Participants were dermatologists recruited by the authors through public outreach via email lists and social media. Most recruited dermatologists were practicing in the US (23); we also recruited dermatologists practicing in the UK (5), Korea (3), Japan (2), Canada (2), Brazil (2), and Poland (1). Unlike our other experiments, the recruited dermatologists were not given monetary compensation. Additionally, we did not measure socioeconomic scores or computer programming experience of the participants. Due to the more stringent recruitment parameters (the participants must be a practicing dermatologist), we were limited in the number of participants we could recruit. Demographic information on the recruited participants is provided in Table 5.

Performance was similar across the two conditions, with advice improving accuracy by a similar amount (Table 6), and only resulting in moderately higher average confidence in the correct answer. This concurs with prior work that found radiologists benefit similarly from human or AI advice when examining chest x-rays [8]. The activation rate was also similar across human and AI advice. The initial and final accuracies are most comparable with the art task (for laypeople), where we observed a large difference in advice utilization across conditions (4.7% accuracy increase and 6.4% activation increase for AI advice). These findings suggest that the dermatologists surveyed did not differentiate between human and AI advice. However, this may be a limitation of our study – in
practice, human advice generally comes from a known and trusted colleague rather than the unknown human source in our study.

We show our activation-integration model coefficients fit to the biopsy data in Table 7. As we could not collect the same demographic information as in previous experiments, we are limited in the number of factors we include in our model. However, we observe similar effects as previously described (Table 4. The signs of $\beta_{\text{prior}}, \beta_{\text{response}}, \beta_{\text{advice}},$ and $\beta_{\text{consistent}}$ are consistent with our previous results. While $\beta_{\text{advice}}$ in the integration stage and $\beta_{\text{prior}}$ in both stages do not have a statistically significant effect, we can partially attribute this to the limited dataset size – < 40% of the data collected for our other experiments. We also note that $\beta_{\text{experience}} = -0.845$ is significant for activation – suggesting that as a dermatologist becomes more experienced, they are less likely to heed advice. However, $\beta_{\text{experience}}$ does not have a significant effect in the integration setting, suggesting that experience serves to gate when advice is used but not how it is used, similar to the effect of prior belief discussed in the laypeople experiments.

These results are important as they indicate our activation-integration model largely holds up in a real-world setting.

Table 5: Participant demographics for Biopsy task.

| Dataset | # of Participants | # of Years Practicing | Education Level |
|---------|-------------------|------------------------|-----------------|
| Biopsy  | 37                | 8.6 +/- 8.6            | 8.0             |

Table 6: Advice effect on Biopsy. “Conf” refers to the average confidence in the correct answer.

| Advice source | Baseline Accuracy (Conf) | Final Accuracy (Conf) | Activation Rate | Advice Accuracy |
|---------------|--------------------------|-----------------------|-----------------|----------------|
| AI            | 67.6% (0.295)            | 74.0% (0.376)         | 29.7%           | 83.3%          |
| Human         | 67.9% (0.295)            | 74.2% (0.387)         | 29.8%           | 83.3%          |

4 DISCUSSION AND FUTURE WORK

This work investigates how individuals incorporate external advice that comes from an AI algorithm. This is an important topic for the broader impact of AI and has been under-explored [8, 14]. We propose the activation-integration model for advice utilization to explain our empirical findings and validate its utility across our diverse set of experimental settings. In the activation stage, a person decides whether to use the advice. In the integration stage, they decide how to incorporate the advice. We also find that the advice source affects whether the advice is acted on (activation), but now how it is integrated. This finding holds over a variety of settings with three different data modalities, across participants from the US, UK and Asia, and across perceived advice quality.

These results provide a better understanding of how humans incorporate advice from AI. Several exciting research directions lie ahead. For example, our work primarily focuses on common tasks—e.g. paintings, locations—that require no special expertise. This is a good starting point given that there has not been similar experiments before. An interesting question is to what extent our activation-integration model applies to settings where human experts incorporate external advice (e.g. a doctor is suggested a medical diagnoses by an AI or another doctor). Our initial results suggest the model holds; however, our data is limited and validating across other, similar tasks is important. Another important question we can ask is how these results generalize across different types of advice. Here we focus on the most direct type of external feedback (i.e. a prediction with confidence score), because this is the most common type of output in machine learning. It would be interesting to explore more complex advice: e.g. where the AI provides an explanation together with its prediction.

Ethical Considerations. We collect data from human subjects in our work through survey questions, where we asked crowd-sourced workers to complete simple tasks (image classification, sarcasm identification in text, and income level predictions from tabular data). We do not collect or release any personally identifiable data associated with the workers. Given the low risk nature of this data, IRB approval was not needed. For the medical task, we did obtain IRB approval. More information on the specifics of the data we collected including compensation and survey design are in Appendix E.3.

This work and its potential extensions also raise some ethical questions. If it is possible to manipulate human utilization of AI advice either through the presentation of the algorithm or through the form of the advice given, there is risk that miscalibrated beliefs about the AI could harm performance. Note that this may not be done intentionally, but can still be harmful. So while understanding what factors influence human perception and utilization of AI advice is critical to designing systems where humans and AI work jointly, it must be done with care to ensure an AI agent is not over-trusted.

Table 7: Activation-integration model coefficients for Biopsy. Coefficients bolded if non-zero with confidence > 0.95.

| Covariates                                      | Activation Coefficients | Response Change Coefficients |
|-------------------------------------------------|-------------------------|------------------------------|
| Given AI Advice? ($\beta_{\text{ai}}$)          | 0.143 +/- 0.602         | -0.002 +/- 0.603             |
| Prior Belief ($\beta_{\text{prior}}$)           | -0.283 +/- 0.427        | 0.059 +/- 0.050              |
| Response 1 Confidence ($\beta_{\text{response}}$)| -0.531 +/- 0.121        | -0.201 +/- 0.026             |
| Advice Confidence ($\beta_{\text{advice}}$)    | 0.260 +/- 0.111         | 0.024 +/- 0.024              |
| Response 1 and Advice Consistent? ($\beta_{\text{consistent}}$) | -0.615 +/- 0.111 | 0.271 +/- 0.023 |
| Exposure to AI (in life) ($\beta_{\text{ai}\text{-exposure}}$) | 1.185 +/- 1.269 | 0.147 +/- 0.140 |
| Years of Clinical Experience ($\beta_{\text{experience}}$) | -0.845 +/- 0.411 | -0.056 +/- 0.051 |
REFERENCES

[1] Dale J Barr. 2013. Random effects structure for testing interactions in linear mixed-effects models. Frontiers in psychology 4 (2013), 328.
[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823 (2014).
[3] Navoneel Chakrabarty and Sanket Bhiwase. 2018. A statistical approach to adult census income level prediction. In 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCCN). IEEE, 207–212.
[4] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/census+income
[5] Mary T Drinidalet, 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.
[6] Robin C Feldman, Ehrik Aldana, and Kara Stein. 2019. Artificial intelligence in the health care space: how we can trust what we cannot know. Stan. L. & Pol’y Rev 30 (2019), 399.
[7] Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 6 (1981), 381–395.
[8] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lerner, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ digital medicine 4, 1 (2021), 1–8.
[9] Francesca Gino and Don A Moore. 2007. Effects of task difficulty on use of advice. Journal of Behavioral Decision Making 20, 1 (2007), 21–35.
[10] Alexa Hagerty and Igor Rubinov. 2019. Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv preprint arXiv:1907.07892 (2019).
[11] Nigel Harvey and Ilan Fischer. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. Organizational behavior and human decision processes 70, 2 (1997), 117–133.
[12] Nigel Harvey, Clare Harries, and Ilan Fischer. 2000. Using advice and assessing its quality. Organizational behavior and human decision processes 81, 2 (2000), 252–273.
[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 770–778.
[14] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2020. Why are we averse to AI? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 1135–1144.
[15] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Cambia, Stephen Dunza, Pascale Guitera, David Gutman, et al. 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific data 8, 1 (2021), 1–8.
[16] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society 5, 1 (2018), 2053951718756684.
[17] Poornima Madhavan and Douglas A Wiegmann. 2007. Effects of information source, pedigree, and reliability on operator interaction with decision support systems. Human Factors 49, 5 (2007), 773–785.
[18] Neda Meshab, Christoph Tauchert, and Peter Buxmann. 2021. 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. ScholarSpace, 4083.
[19] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence 267 (2019), 1–38.
[20] Dilek Önkul, Paul Goodwin, Mary Thomson, Simon Gönül, and Andrew Pollock. 2009. The relative influence of advice from human experts and statistical methods on forecast adjustments. Journal of Behavioral Decision Making 22, 4 (2009), 390–409.
[21] Andrew Prahl and Lyn Van Swol. 2017. Understanding algorithm aversion: When is advice from automation discounted? Journal of Forecasting 36, 6 (2017), 691–702.
[22] Prolific 2022. Prolific Academic. https://www.prolific.co.
[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why should i trust you?” Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 1135–1144.
[24] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Cambia, Stephen Dunza, Pascale Guitera, David Gutman, et al. 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific data 8, 1 (2021), 1–8.
[25] Sunita Sah, Don A Moore, and Robert J MacCoun. 2013. Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. Organizational Behavior and Human Decision Processes 121, 2 (2013), 246–255.
[26] Thomas Schultz, Anne-Fernandine Rakotoarisoa, and Stefan Schulz-Hardt. 2015. Effects of distance between initial estimates and advice on advice utilization. Judgment & Decision Making 10, 2 (2015).
[27] Christoph Tauchert, Neda Meshab, et al. 2019. Following the Robot? Investigating Users’ Utilization of Advice from Robo-Advisors. In ICS. AIS.
[28] Lyn M Van Swol, Jihyun Esther Paik, and Andrew Prahl. 2018. Advice recipients: The psychology of advice utilization. (2018).
[29] Alex West and Anusha Praturu. 2019. Enhancing the Census Income Prediction Dataset. https://people.ischool.berkeley.edu/~alexwest/w210_census_income.html/. Accessed: 2021-05-15.
[30] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang’Anthony’ Chen. 2020. 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. ACM, 1–13.
[31] Iian Yaniv and Dean P Foster. 1997. Precision and accuracy of judgmental estimates. Journal of Behavioral Decision Making 10, 1 (1997), 21–32.
A APPENDIX

We provide more data about participants’ prior beliefs about the effectiveness of AI in Section B. Section C provides more details about our UK and Asia experiments which further validated our findings. Section E provides more information about our survey design. Our study followed standard practices for studying human evaluations of external advice [28]. All participants consented to join the study and we protect their privacy by only using de-identified data in our analysis.

B DISTRIBUTION OF PRIOR

We show the distributions of prior belief across our four datasets in Figure 6. For the art, cities, and census datasets, the prior belief favors the AI advice; for the sarcasm dataset, the prior belief favors the human advice. Note that for all datasets there is a high variance, and there are always participants who strongly believe either the AI or human advice is better regardless of the average belief.

C UK AND ASIA VALIDATION EXPERIMENTS

C.1 Dataset Overview

We show an overview of the datasets used for experiments in the UK and Asia in Table 8. Comparing with Table 1, note that the baseline accuracies in the UK are similar to the US, while the baseline accuracies in Asia are lower in both datasets. The prior belief across both datasets in Asia and the UK is in the same direction as in the US (AI or human favored). In the UK, the prior belief is stronger (e.g., higher percentage of people favor AI for the art dataset and human for the sarcasm dataset); in Asia, it is only stronger for the art dataset.

C.2 Demographics

Demographic information for the UK and Asia experiments is shown in Table 9. All values are similar to the US datasets with two slight differences: (1) the average age of the Sarcasm dataset in the UK is close to 10 years higher, and (2) the socioeconomic score and education level in Asia are both about 0.5 higher than the US data.

Participants from Asia came from a diverse set of countries. At least 10 participants across both art and sarcasm datasets are Indian, Chinese, Vietnamese, Filipinos, Pakistani, Indonesian, or Korean.

C.3 Activation-integration model

Here we include the coefficient values referenced in Section 3.4.1. In Table 10, we show the coefficients for the activation model trained using US data, UK data, and Asia data. As was noted earlier, the magnitude and sign of the $\beta_{\text{prior}}, \beta_{\text{response}}, \beta_{\text{advice}}$, and $\beta_{\text{consistent}}$ are largely the same across all three regions. Demographic values are observed to be similar across the three datasets, as expected.

D PERCEIVED ACCURACY VALIDATION EXPERIMENTS

D.1 Dataset Overview

We show an overview of the datasets used for experiments on the perceived accuracy variation in Table 12. We use the art dataset (Figure 2 a); the 80% accurate advice data is the same data as used in previous sections. Note the only difference between the three datasets is the information given to the participants about the accuracy of the advice; the actual advice used is identical across all three datasets. Demographic values are observed to be similar across the three datasets, as expected.

D.2 Activation-integration model

Here we include the coefficient values referenced in Section 3.4.2. In Table 13, we show the coefficients for the activation model trained using the three perceived accuracy bins. As was noted earlier, the magnitude and sign of the $\beta_{\text{prior}}, \beta_{\text{response}}, \beta_{\text{advice}}$, and $\beta_{\text{consistent}}$ are largely the same across all three groups.

In Table 14, we show the coefficients for the integration model across all three groups. $\beta_{\text{response}}, \beta_{\text{advice}}$, and $\beta_{\text{consistent}}$ have identical sign and similar magnitude across all three groups. The prior effect is not significant in any of the regions, supporting our two-stage model.

E SURVEY DETAILS

E.1 Demographic Information Collection

We conducted our experiments using Prolific [22]. Prolific provides access to various demographic information that we used in our study. These include age, gender, education level, socioeconomic status, and whether the participant has programming experience. Education level is defined on a scale from 1 to 8. The interpretation of the education level is given below:

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. The question asked to participants to determine the socioeconomic status is included in Figure 7.

E.2 Survey Question About Prior

Here we add some additional details on the experiments we ran to check whether completing the task has a systematic effect on prior belief response. As described in the main paper, we asked the survey question to measure prior belief after all task instances have been completed in the majority of our experiments. This was done to avoid any potential biasing in the participants’ prior to completing the tasks, the participants are not aware that there are two types of advice they could have received. However, a concern is that completing the task instances will bias a participant’s response.

To check whether the location of the survey affected people’s responses, we ran additional experiments on the Art and Sarcasm
datasets where we asked the prior belief survey question both prior to showing participants any tasks and after all tasks are completed. In Figure 8, we plot the prior belief survey response before and after the participant has completed all task instances. We plot a robust linear fit of the data computed using RANSAC [7] (in red) as well as the expected fit if prior belief response is not affected by completing the task instances. The robust fit in both the Art and Sarcasm datasets indicate that the location of prior belief survey does not have a systematic affect on prior belief response.

We further confirmed this conclusion by running our mixed effects analysis using the prior belief responses from the start of the survey. We found that our main findings hold—prior belief has a significant effect for the activation stage, but not for the integration stage.

### E.3 Participant Compensation

Participants were compensated at a rate of $10.00 per hour as per the Prolific recommended rates. The survey was estimated to take 10 minutes based on several trial runs by the authors. Participants were compensated with this assumption.

### Table 8: Summary of the datasets we use in our experiments in the UK and Asia. Follows same format as Table 2

| Dataset       | Data Type | Description      | # of Tasks | Baseline accuracy | Prior belief |
|---------------|-----------|------------------|------------|-------------------|--------------|
| Art (UK)      | Image     | identify the art period | 32         | 65.5% +/- 16.6%   | 74.4% (AI)   |
| Sarcasm (UK)  | Text      | identify sarcasm  | 32         | 70.6% +/- 20.8%   | 83.0% (human) |
| Art (Asia)    | Image     | identify the art period | 32         | 62.2% +/- 17.0%   | 76.3% (AI)   |
| Sarcasm (Asia)| Text      | identify sarcasm  | 32         | 62.1% +/- 21.9%   | 78.2% (human) |
Table 9: Summary of the participant demographics from the 4 datasets we use in our experiments. See Section E.1 for information on the socioeconomic score and education level.

| Dataset            | # of Participants | Gender (Percent Male / Female) | Age          | Socioeconomic Score | Education Level |
|--------------------|-------------------|--------------------------------|--------------|---------------------|-----------------|
| Art Dataset (UK)   | 86                | 45.3% / 54.7%                  | 32.2 +/- 12.3| 5.2 +/- 1.6         | 5.1 +/- 1.4     |
| Sarcasm Dataset (UK)| 88              | 55.7% / 44.3%                  | 41.4 +/- 15.4| 4.9 +/- 1.8         | 5.1 +/- 1.6     |
| Art Dataset (Asia) | 97                | 56.7% / 43.3%                  | 27.4 +/- 6.8 | 5.8 +/- 1.6         | 5.9 +/- 1.1     |
| Sarcasm Dataset (Asia)| 101        | 48.5% / 51.5%                  | 29.0 +/- 7.3 | 5.8 +/- 1.4         | 6.1 +/- 1.3     |

Table 10: Coefficients for activation stage mixed effects logistic regression model. Comparison across different geographical regions.

| Covariates                                  | Coefficients (US) | Coefficients (UK) | Coefficients (Asia) |
|---------------------------------------------|-------------------|-------------------|---------------------|
| Given AI Advice? ($\beta_{ai}$)             | 0.162 +/- 0.134   | -0.253 +/- 0.218  | 0.389 +/- 0.207     |
| Prior Belief ($\beta_{prior}$)              | 0.244 +/- 0.059   | 0.217 +/- 0.107   | 0.101 +/- 0.100     |
| Response 1 Confidence ($\beta_{response}$)  | -0.584 +/- 0.025  | -0.491 +/- 0.037  | -0.355 +/- 0.038    |
| Advice Confidence ($\beta_{advice}$)        | 0.218 +/- 0.024   | 0.146 +/- 0.036   | 0.143 +/- 0.033     |
| Response 1 and Advice Consistent? ($\beta_{consistent}$) | -1.555 +/- 0.051 | -1.157 +/- 0.076  | -1.492 +/- 0.073    |
| Age ($\beta_{age}$)                         | -0.089 +/- 0.072  | -0.074 +/- 0.084  | 0.110 +/- 0.192     |
| Gender ($\beta_{gender}$)                   | -0.046 +/- 0.137  | 0.201 +/- 0.229   | 0.157 +/- 0.224     |
| Education ($\beta_{education}$)             | 0.043 +/- 0.076   | 0.027 +/- 0.106   | -0.118 +/- 0.133    |
| Socioeconomic Status ($\beta_{socioeconomic}$) | 0.139 +/- 0.070   | 0.013 +/- 0.109   | -0.010 +/- 0.109    |
| Has Programming Experience? ($\beta_{programming}$) | 0.233 +/- 0.147  | -0.209 +/- 0.256  | 0.031 +/- 0.217     |

Table 11: Coefficients for integration stage mixed effects linear regression model. Comparison across different geographical regions.

| Covariates                                  | Coefficients (US) | Coefficients (UK) | Coefficients (Asia) |
|---------------------------------------------|-------------------|-------------------|---------------------|
| Given AI Advice? ($\beta_{ai}$)             | -0.023 +/- 0.016  | 0.008 +/- 0.024   | -0.000 +/- 0.027    |
| Prior Belief ($\beta_{prior}$)              | 0.001 +/- 0.008   | -0.023 +/- 0.012  | 0.005 +/- 0.013     |
| Response 1 Confidence ($\beta_{response}$)  | -0.212 +/- 0.004  | -0.199 +/- 0.007  | -0.226 +/- 0.007    |
| Advice Confidence ($\beta_{advice}$)        | 0.025 +/- 0.004   | 0.033 +/- 0.007   | 0.016 +/- 0.006     |
| Response 1 and Advice Consistent? ($\beta_{consistent}$) | 0.661 +/- 0.008  | 0.589 +/- 0.014   | 0.674 +/- 0.011     |
| Age ($\beta_{age}$)                         | -0.026 +/- 0.008  | -0.013 +/- 0.009  | -0.014 +/- 0.025    |
| Gender ($\beta_{gender}$)                   | 0.010 +/- 0.016   | 0.027 +/- 0.025   | -0.031 +/- 0.029    |
| Education ($\beta_{education}$)             | 0.021 +/- 0.009   | -0.016 +/- 0.012  | 0.015 +/- 0.017     |
| Socioeconomic Status ($\beta_{socioeconomic}$) | 0.007 +/- 0.008   | -0.020 +/- 0.012  | -0.028 +/- 0.014    |
| Has Programming Experience? ($\beta_{programming}$) | 0.014 +/- 0.017  | 0.044 +/- 0.028   | -0.007 +/- 0.028    |

Participants were also informed that they could receive up to a 30% bonus. This bonus was calculated as

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

where $S$ is the average performance of the participant across all tasks, both before and after receiving advice. Performance for a single task is computed as

$$\text{sign}(\text{correct response}) \times \text{response}_1$$

775
Table 12: Summary of the datasets we use in our experiments varying the perceived accuracy. Follows same format as Table 2.

| Dataset                  | # of Participants | Gender (Percent Male / Female) | Age       | Socioeconomic Score | Education Level |
|--------------------------|-------------------|--------------------------------|-----------|---------------------|-----------------|
| Art (65% accurate advice)| 103               | 42.7% / 57.3%                  | 34.1 +/- 12.8 | 5.0 +/- 1.5        | 5.7 +/- 1.2     |
| Art (80% accurate advice)| 147               | 47.6% / 52.4%                  | 33.0 +/- 12.3 | 5.2 +/- 1.5        | 5.4 +/- 1.3     |
| Art (95% accurate advice)| 104               | 39.4% / 60.6%                  | 29.0 +/- 9.8  | 4.9 +/- 1.7        | 5.3 +/- 1.2     |

Table 13: Coefficients for activation stage mixed effects logistic regression model. Comparison across different perceived advice accuracies.

| Covariates                          | Coefficients (65%) | Coefficients (80%) | Coefficients (95%) |
|-------------------------------------|--------------------|--------------------|--------------------|
| Given AI Advice? (βai)              | -0.006 +/- 0.297   | 0.167 +/- 0.278    | 0.582 +/- 0.339    |
| Prior Belief (βprior)               | -0.003 +/- 0.176   | 0.223 +/- 0.151    | -0.222 +/- 0.156   |
| Response 1 Confidence (βresponse)   | -0.729 +/- 0.058   | -0.679 +/- 0.050   | -0.665 +/- 0.059   |
| Advice Confidence (βadvice)         | 0.196 +/- 0.052    | 0.156 +/- 0.042    | 0.088 +/- 0.052    |
| Response 1 and Advice Consistent? (βconsistent) | -1.616 +/- 0.103  | -2.066 +/- 0.093   | -2.004 +/- 0.114   |
| Age (βage)                          | -0.155 +/- 0.134   | 0.123 +/- 0.118    | 0.006 +/- 0.173    |
| Gender (βgender)                    | 0.173 +/- 0.306    | -0.001 +/- 0.264   | -0.348 +/- 0.311   |
| Education (βeducation)              | -0.141 +/- 0.174   | -0.132 +/- 0.136   | -0.180 +/- 0.170   |
| Socioeconomic Status (βsocioeconomic) | 0.323 +/- 0.167   | 0.220 +/- 0.137    | 0.387 +/- 0.151    |
| Has Programming Experience? (βprogramming) | -0.198 +/- 0.342  | 0.251 +/- 0.283    | -0.495 +/- 0.346   |

Table 14: Coefficients for integration stage mixed effects linear regression model. Comparison across different perceived advice accuracies.

| Covariates                          | Coefficients (65%) | Coefficients (80%) | Coefficients (95%) |
|-------------------------------------|--------------------|--------------------|--------------------|
| Given AI Advice? (βai)              | 0.046 +/- 0.045    | -0.082 +/- 0.039   | -0.019 +/- 0.035   |
| Prior Belief (βprior)               | -0.038 +/- 0.028   | 0.017 +/- 0.021    | -0.026 +/- 0.016   |
| Response 1 Confidence (βresponse)   | -0.186 +/- 0.012   | -0.237 +/- 0.009   | -0.259 +/- 0.010   |
| Advice Confidence (βadvice)         | -0.024 +/- 0.010   | 0.009 +/- 0.008    | 0.021 +/- 0.009    |
| Response 1 and Advice Consistent? (βconsistent) | 0.599 +/- 0.019    | 0.681 +/- 0.014    | 0.677 +/- 0.017    |
| Age (βage)                          | -0.011 +/- 0.021   | -0.028 +/- 0.016   | -0.011 +/- 0.018   |
| Gender (βgender)                    | 0.011 +/- 0.047    | -0.010 +/- 0.037   | 0.029 +/- 0.032    |
| Education (βeducation)              | -0.018 +/- 0.026   | 0.004 +/- 0.019    | 0.027 +/- 0.018    |
| Socioeconomic Status (βsocioeconomic) | 0.035 +/- 0.026   | 0.030 +/- 0.019    | -0.012 +/- 0.016   |
| Has Programming Experience? (βprogramming) | -0.001 +/- 0.052   | 0.031 +/- 0.039    | -0.002 +/- 0.036   |

where $-1 \leq \text{response}_1 \leq 1$. Note that this performance metric penalizes incorrect responses.

The total cost of running all of our experiments (including the participants we used to calibrate the advice) was around $2,500.
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 7: Question participants are asked by Prolific [22] to obtain participant’s socioeconomic status.

Figure 8: Scatter plot of prior belief survey response before and after tasks for (a) Art and (b) Sarcasm datasets. The dotted black line shows the expected linear fit if location of the survey does not affect prior belief response; the red line shows a robust linear fit of the data.