Does More Advice Help? The Effects of Second Opinions from Peers in AI-Assisted Decision Making

ZHUORAN LU, Purdue University, USA

DAKUO WANG, Northeastern University, USA

MING YIN, Purdue University, USA

AI assistance in decision-making has become popular, yet people's inappropriate reliance on AI often leads to unsatisfactory human-AI collaboration performance. In this paper, we explore how providing second opinions from human peers may affect decision-makers' behavior and performance in an AI-assisted decision-making setting. Via two pre-registered, randomized experiments, we find that when both AI recommendation and a second opinion from human peers are always presented together, decision-makers reduce their over-reliance on AI, despite their under-reliance on AI also increases. Furthermore, if decision-makers have the control to decide when to solicit a peer's second opinion, their active solicitations of second opinions have the potential to mitigate over-reliance on AI without inducing increased under-reliance in some cases.

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

Additional Key Words and Phrases: Machine learning, second opinions, appropriate reliance, human-AI interaction

ACM Reference Format:

Zhuoran Lu, Dakuo Wang, and Ming Yin. 2018. Does More Advice Help? The Effects of Second Opinions from Peers in AI-Assisted Decision Making. In *CHI TRAIT '23, April 23, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With its rapid development in recent years, Artificial Intelligence (AI) technology has been integrated into many industries, such as business [16, 31], healthcare [19, 20], transportation [25], and more. A common way for AI to augment human workflows in various domains is through *AI-assisted decision-making*, that is, an AI-based decision aid provides decision recommendations to humans while humans make the final decisions. As humans and AI may each possess unique intelligence that is complementary to each other, the human-AI collaborations in this decision-making scenario have the potential to utilize the best of humans and AI and realize a joint performance beyond what can be achieved by each party alone.

In reality, however, the joint decision-making performance of the human-AI team is often not as good as expected. One primary reason underlying such unsatisfactory human-AI collaboration is that humans often rely on the AI recommendations *inappropriately*. Humans may not trust an AI model and hence avoid adopting its recommendations even when the recommendations are highly accurate, resulting in ***under-reliance*** on AI [11]. On the other hand, sometimes humans also show a degree of ***over-reliance*** on AI, as they blindly accept the recommendations of an AI model even when it makes sizable mistakes [6, 8, 29]. To help people establish a more appropriate level of reliance on AI

Manuscript submitted to ACM

and improve the human-AI joint decision-making performance, researchers and practitioners have explored a wide range of methods, such as enhancing humans' understandings of the rationale underlying AI recommendations [3, 32, 33, 40], enforcing people to engage in careful deliberation [6, 27], communicating to people the importance of their decisions [1], multi-human decision-making under AI assistance[36, 37], and building up social transparency in AI systems[12]. However, mixed results have been reported regarding the effectiveness of these methods.

This challenge of humans inappropriately relying on suggestions provided by some "advisors" is not new. Indeed, in the classical paradigm of "Judge-Advisor System" in the advice taking research [5], where a human "judge" receives suggestions from another human "advisor" before making their final judgement on a decision-making problem, it is also observed that the judge may inappropriately discount the advisor's suggestions [42] or over-utilize the advisor's low-quality advice [35]. Interestingly, a common intervention adopted in these scenarios to improve the human judge's decision-making quality is to introduce advice from a second advisor, so that the judge can explore different perspectives and suggestions to make a better final decision [7, 17, 41, 43]. It is thus natural to ask if similar methods will also benefit AI-assisted decision-making—If second opinions from other *human peers* are presented to a decision-maker in addition to the AI recommendation, can they help the decision maker rely on AI more appropriately and achieve a higher level of decision-making performance? As a motivating example, consider an investor who is assisted by an AI model in deciding their stock trading strategies [21, 28, 39]: when an investor is about to buy/sell a stock given an AI model's recommendation, will presenting a second opinion from other investors (e.g., from online discussion forum *wallstreetbets* in reddit [4, 23]) help them make better investment decisions (or the opposite)?

There are reasons to conjecture the answer to this question either way. On the one hand, it is possible that the decision-maker (e.g., the investor) may perceive the AI model to be more competent than peers (e.g., AI has the "expert power" or authority) [15, 18], which may imply that the presence of second opinions from peers can hardly change how they interact with the AI model. On the other hand, observing potential disagreements between the AI model and the peers in some decision-making cases may nudge the decision-maker into evaluating the AI recommendations more critically and incorporating them into their final decisions more intelligently, which may result in an improvement in their decision-making accuracy—and perhaps second opinions from those who oppose the AI more frequently can lead to a larger accuracy improvement [10]. It is also possible that the decision-maker will leverage the level of agreement between the AI recommendations and the second opinions from peers as a heuristic to gauge the trustworthiness of the AI model and adjust their reliance strategies accordingly, although how changes in the decision-maker's reliance on AI translate to changes in their decision-making accuracy is not straight-forward.

Therefore, to obtain a thorough understanding of how **second opinions from peers** affect the decision maker's behavior (e.g., overall reliance, over-reliance, and under-reliance on AI) and performance (e.g., decision accuracy) in an AI-assisted decision-making scenario, we conducted two pre-registered, randomized human-subject experiments (Experiment 1: $N = 428$, Experiment 2: $N = 336$) on Amazon Mechanical Turk (MTurk). As the second opinions provided by peers may agree with the AI model's decision recommendations at different frequencies (e.g., from agree on almost all decision-making cases to agree very occasionally), we are also interested in examining whether the effects of peer-generated second opinions are moderated by the level of agreement between the peers and the AI model.

In our experiments, subjects were asked to complete a series of sentiment analysis tasks to decide whether a movie review is positive or negative, with the decision recommendations provided by an AI model. In Experiment 1, we created four treatments by varying whether second opinions from peers were presented to subjects on *each* decision-making task, and if so, how frequently they agreed with the AI recommendations. Our results show that when second opinions from peers are *always* available, they result in significant decreases in people's over-reliance on the AI model. However, they also lead to increased levels of under-reliance on AI. These changes are particularly salient as the peers disagree with the AI model more frequently, although, in none of the treatments with second opinions from peers, people's decision accuracy (i.e., appropriate reliance on AI) is significantly different from that in the control treatment where no second opinion is presented.

To mitigate the undesirable side effect of the increased level of under-reliance on AI when second opinions are always present, in our Experiment 2, instead of presenting a second opinion on each task, we provided subjects with the option to actively *request* for a second opinion from peers only if they needed it, hoping that subjects could utilize their perceptions of AI correctness to decrease the presence of second opinions on tasks where the AI model was correct compared to on tasks where the AI model was wrong. Focusing on the comparisons between those subjects who had requested for second opinions *at least on some task* and the comparable subjects in the control treatment who never saw any second opinions (obtained via propensity score matching), we find that people's active solicitations of second opinions from peers may result in a decrease in over-reliance *without* inducing higher levels of under-reliance, but only when the peers have a relatively high level of agreement with the AI model.

Taken together, our results highlight the promise of introducing second opinions generated by human peers as an intervention in the AI-assisted decision-making workflows to help people rely on AI more appropriately and eventually improve their AI-assisted decision-making performance. Meanwhile, the effectiveness of this intervention is shown to be dependent on both the ways that the second opinions are presented and the characteristics of the second opinions.

2 EXPERIMENT 1: PEER'S SECOND OPINIONS ALWAYS PRESENTED

To understand how the presence of second opinions from peers affect people's behavior and performance in AI-assisted decision-making, and how these effects vary with the agreement level between the peers and the AI model, we conduct a randomized experiment on Amazon Mechanical Turk (MTurk).

2.1 Experimental Task

In our experiment, we asked subjects to determine the sentiment of movie reviews with the help of an AI model. Specifically, in each task, subjects were presented with a movie review taken from the IMDB movie review dataset [24] (with lengths between 280 and 300 words). Along with the movie review, we also showed subjects an AI model's binary prediction of the review's sentiment (i.e., positive vs. negative), while subjects in some experimental treatments also had access to the judgement of the review's sentiment made by a peer (i.e., a randomly selected crowd worker; see Section 2.2 for details). After reviewing all this information, subjects were asked to make a final decision on whether the sentiment expressed in the movie review was positive or negative. In total, each subject needed to review the same set of 20 movie reviews in our experiment.

On each sentiment analysis task, all subjects in our experiment were presented with the prediction given by the *same* AI model. In particular, to obtain this AI model, we fine-tuned a pre-trained RoBERTa model [22] from the Huggingface's transformers library—First, from the IMDB movie review dataset, we sampled a subset of movie reviews (5000 as training set, another 500 as the test set). We used the representation embeddings of the last layer of the pre-trained RoBERTa model as the input of a multi-layer perceptron and tuned parameters of both the pre-trained model and the perceptron based on the training data. Our final model achieved an accuracy of 77.6% on the held-out test set. On the set of 20 movie reviews we used in our experiment, the model's accuracy was 75% (i.e., correct on 15 tasks and incorrect on 5 tasks), which closely reflected the model's overall performance on the test set. We also intentionally did not train a model with very high performance so that we would have sufficient data to understand how subjects behave in AI-assisted decision-making both when the AI model is correct and wrong (e.g., analyze subjects' under-reliance and over-reliance separately).

We used sentiment analysis tasks in our experiment both because it reflects real-world AI-assisted decision-making scenarios well, and because it requires no specific domain knowledge from our human subjects. Similar tasks have also been used in previous studies to investigate human behavior in AI-assisted decision-making [9, 14, 34], and to explore ways to promote humans' appropriate reliance on AI in AI-assisted decision-making [3]. We note that for the IMDB dataset from which we draw our decision-making tasks, the ground-truth label for a movie review's sentiment is decided by the review poster's *own* star rating of the movie (on a scale of 1 to 10) associated with their review text.

2.2 Experimental Treatments

In total, we created four treatments for our experiment by varying the presence of second opinions from peers and the level of agreement between peers' judgements and the AI model's predictions.

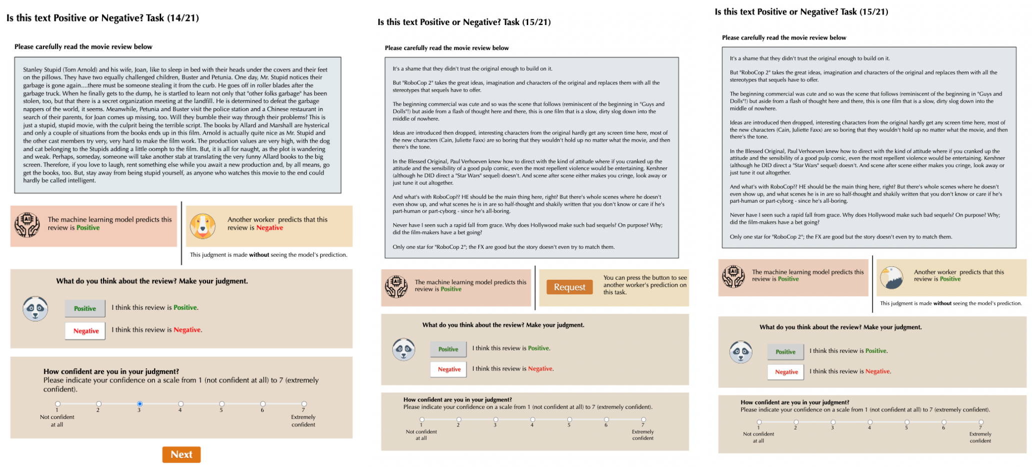
Specifically, to enable the presence of second opinions from real human decision-makers to ensure ecological validity, we first ran a pilot study in which 34 MTurk workers were recruited to review the same set of 20 movie reviews that we selected for our experiment. These workers were asked to determine the sentiment of each review *independently*, i.e., without seeing our AI model's prediction. For each worker, we then computed the fraction of tasks in which their independent judgement was the same as our AI model's prediction across all 20 tasks; we denoted this fraction as the worker's "level of agreement" with the AI model. After each worker's level of agreement with the AI model was computed, we identified three subsets of workers from the entire pool of 34 workers, with each subset containing three workers—The first subset contained the three workers who agreed with the AI model most frequently, and we referred to them as the "*high agreement peers*"; the second subset contained three workers who agreed with the AI model on about 50% of the tasks, and we referred to them as the "*medium agreement peers*"; finally, the last subset contained the three workers who agreed with the AI model least frequently, whom we referred to as the "*low agreement peers*"[1].

Utilizing these three sets of "peers" that we identified from our pilot study, we designed the following 4 treatments:

- **Treatment 1 (Control)**: Subjects had access to the predictions of the AI model when completing each task. However, they did not see any judgement made by other peer workers.
- **Treatment 2 (With high agreement peers)**: Subjects had access to predictions made by the AI model when completing each task. In addition, on each task, we randomly selected a worker from the set of three *high agreement peers*, and presented the selected worker's judgement on that task to the subject as a second opinion[2].
- **Treatment 3 (With medium agreement peers)**: Subjects had access to predictions made by the AI model when completing each task. In addition, on each task, we randomly selected a worker from the set of three *medium agreement peers*, and presented the selected worker's judgement on that task to the subject as a second opinion.
- **Treatment 4 (With low agreement peers)**: Subjects had access to predictions made by the AI model when completing each task. In addition, on each task, we randomly selected a worker from the set of three *low agreement peers* and presented the selected worker's judgement on that task to the subject as a second opinion.

[1] For the subset of high, medium, and low agreement peers, the average level of agreement between the crowd workers and the AI model was 76.67%, 50%, and 30%, respectively.

[2] While in this experiment, the second opinions presented to subjects have already been collected from crowd workers in the pilot study, in reality, they can be solicited from peers when the decision maker is about to make their decision on a task. We decided to collect second opinions ahead of time in this study to simplify the experimental procedure and to enable the quantification of the level of agreement between second opinions and the AI model.



Is this text Positive or Negative? Task (14/21)

Please carefully read the movie review below

Stanley Stupid (Tom Arnold) and his wife, Joan, like to sleep in bed with their heads under the covers and their feet on the pillows. They have two equally challenged children, Buster and Petunia. One day, Mr. Stupid notices their garbage is gone again....there must be someone stealing it from the curb. He goes off in roller blades after the garbage truck. When he finally gets to the dump, he is startled to learn not only that "other folks garbage" has been stolen, too, but that there is a secret organization meeting at the landfill. He is determined to defeat the garbage nappers of the world, it seems. Meanwhile, Petunia and Buster visit the police station and a Chinese restaurant in search of their parents, for Joan comes up missing, too. Will they bumble their way through their problems? This is just a stupid, stupid movie, with the culprit being the terrible script. The books by Allard and Marshall are hysterical and only a couple of situations from the books ends up in this film. Arnold is actually quite nice as Mr. Stupid and the other cast members try very, very hard to make the film work. The production values are very high, with the dog and cat belonging to the Stupids adding a little oomph to the film. But, it is all for naught, as the plot is wandering and weak. Perhaps, someday, someone will take another stab at translating the very funny Allard books to the big screen. Therefore, if you love to laugh, rent something else while you await a new production and, by all means, go get the books, too. But, stay away from being stupid yourself, as anyone who watches this movie to the end could hardly be called intelligent.

The machine learning model predicts this review is Positive

Another worker predicts that this review is Negative

This judgment is made **without** seeing the model's prediction.

What do you think about the review? Make your judgment.

I think this review is Positive.

I think this review is Negative.

How confident are you in your judgment?
Please indicate your confidence on a scale from 1 (not confident at all) to 7 (extremely confident).

○ 1 Not confident at all ○ 2 ● 3 ○ 4 ○ 5 ○ 6 ○ 7 Extremely confident

Is this text Positive or Negative? Task (15/21)

Please carefully read the movie review below

It's a shame that they didn't trust the original enough to build on it.

But "RoboCop 2" takes the great ideas, imagination and characters of the original and replaces them with all the stereotypes that sequels have to offer.

The beginning commercial was cute and so was the scene that follows (reminiscent of the beginning in "Guys and Dolls"!) but aside from a flash of thought here and there, this is one film that is a slow, dirty slog down into the middle of nowhere.

Ideas are introduced then dropped, interesting characters from the original hardly get any screen time here, most of the new characters (Cain, Juliette Faxx) are so boring that they wouldn't hold up no matter what the movie, and then there's the tone.

In the Blessed Original, Paul Verhoeven knew how to direct with the kind of attitude where if you cranked up the attitude and the sensibility of a good pulp comic, even the most repellent violence would be entertaining. Kershner (although he DID direct a "Star Wars" sequel) doesn't. And scene after scene either makes you cringe, look away or just tune it out altogether.

And what's with RoboCop?? HE should be the main thing here, right? But there's whole scenes where he doesn't even show up, and what scenes he is in are so half-thought and shakily written that you don't know or care if he's part-human or part-cyborg - since he's all-boring.

Never have I seen such a rapid fall from grace. Why does Hollywood make such bad sequels? On purpose? Why; did the film-makers have a bet going?

Only one star for "RoboCop 2"; the FX are good but the story doesn't even try to match them.

The machine learning model predicts this review is Positive

You can press the button to see another worker's prediction on this task.

What do you think about the review? Make your judgment.

I think this review is Positive.

I think this review is Negative.

How confident are you in your judgment?
Please indicate your confidence on a scale from 1 (not confident at all) to 7 (extremely confident).

○ 1 Not confident at all ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 Extremely confident

Is this text Positive or Negative? Task (15/21)

Please carefully read the movie review below

It's a shame that they didn't trust the original enough to build on it.

But "RoboCop 2" takes the great ideas, imagination and characters of the original and replaces them with all the stereotypes that sequels have to offer.

The beginning commercial was cute and so was the scene that follows (reminiscent of the beginning in "Guys and Dolls"!) but aside from a flash of thought here and there, this is one film that is a slow, dirty slog down into the middle of nowhere.

Ideas are introduced then dropped, interesting characters from the original hardly get any screen time here, most of the new characters (Cain, Juliette Faxx) are so boring that they wouldn't hold up no matter what the movie, and then there's the tone.

In the Blessed Original, Paul Verhoeven knew how to direct with the kind of attitude where if you cranked up the attitude and the sensibility of a good pulp comic, even the most repellent violence would be entertaining. Kershner (although he DID direct a "Star Wars" sequel) doesn't. And scene after scene either makes you cringe, look away or just tune it out altogether.

And what's with RoboCop?? HE should be the main thing here, right? But there's whole scenes where he doesn't even show up, and what scenes he is in are so half-thought and shakily written that you don't know or care if he's part-human or part-cyborg - since he's all-boring.

Never have I seen such a rapid fall from grace. Why does Hollywood make such bad sequels? On purpose? Why; did the film-makers have a bet going?

Only one star for "RoboCop 2"; the FX are good but the story doesn't even try to match them.

The machine learning model predicts this review is Positive

Another worker predicts that this review is Positive

This judgment is made **without** seeing the model's prediction.

What do you think about the review? Make your judgment.

I think this review is Positive.

I think this review is Negative.

How confident are you in your judgment?
Please indicate your confidence on a scale from 1 (not confident at all) to 7 (extremely confident).

○ 1 Not confident at all ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 Extremely confident

(a) Experiment 1

(b) Experiment 2: Before request

(c) Experiment 2: After request

Fig. 1. An illustration of our task interface in our experiments. Fig. 1a shows an example interface in Experiment 1 for treatments where second opinions were presented. Fig. 1b and 1c show an example of our task interface in Experiment 2 (for treatments where subjects can solicit the second opinions from peer workers), before (1b) and after (1c) subjects clicked the "Request" button.

Figure 1a shows an example of the task interface for treatments where the second opinions from peers are presented (i.e., Treatment 2, 3, or 4). With this design, we expect that subjects in Treatment 2 will find the second opinions generated by peer workers agree with the AI model more frequently than subjects in Treatment 3, who in turn will observe a higher level of agreement between the peers and the AI model than subjects in Treatment 4.

2.3 Experimental Procedure

Our experiment was posted on Amazon Mechanical Turk (MTurk) as a human intelligence task (HIT). The HIT was open to workers in the U.S. only, and each worker could only take the HIT once. Each HIT contained the same 20 movie review tasks, which were arranged in random order. In addition, we included an attention check question in our HIT, in which the subject was instructed to select a pre-specified option. We only considered the data generated by subjects who passed the attention check as valid data in our analysis.

Upon arrival at the HIT, subjects were randomly assigned to one of the four treatments. Subjects first received instruction on the movie review task. In order to show that they understood how to complete the movie review tasks, subjects needed to complete a qualification task, in which they were asked to review a simple movie review and determine its sentiment. Subjects could proceed to the actual experiment only if they answered the qualification question correctly. In the actual experiment, subjects were first asked to pick an avatar to represent themselves throughout the experiment. Then, as we have discussed in Section 2.1–2.2, subjects completed the 20 movie review tasks, and depending on the treatment they were assigned, on each task, they saw the decision recommendation generated by our AI model and possibly by other peer workers. In each task, beyond making a decision on the movie review's sentiment, subjects were also asked to indicate how confident they were in their decision using a 7-point Likert scale from 1 ("not confident at all") to 7 ("extremely confident"). After completing all 20 tasks, subjects needed to fill out an exit-survey to report their demographics.

The base payment of our experiment is $1.2. To encourage subjects to carefully deliberate about the decision recommendations made by the AI model and the peers, we also provided a performance-based bonus to subjects—If the subject's accuracy in our HIT was higher than 65%, we paid them an additional 5-cent bonus for each correct prediction they made; thus, subjects could earn up to $1 bonus in our experiment in addition to the base payment.

2.4 Analysis Methods

To understand how second opinions from peers affect people's behavior and performance in AI-assisted decision-making, we pre-registered a set of dependent variables for this experiment[3]. Specifically, to examine how peers' judgements change people's reliance on AI models in AI-assisted decision-making, and whether these changes are desirable or not, we consider the following dependent variables:

- **Overall reliance**: The chance for a subject's decision to be the same as the AI model's prediction.
- **Over-reliance**: The chance for a subject's decision to be the *same* as the AI model's prediction, when the AI model's prediction was *incorrect*.
- **Under-reliance**: The chance for a subject's decision to be *different* from the AI model's prediction, when the AI model's prediction was *correct*.
- **Appropriate reliance**: The chance for a subject's decision to be the same as the AI model's correct prediction or different from the AI model's incorrect prediction; this effectively represents the subject's *decision accuracy*.

A subject's overall reliance quantifies the subject's reliance behavior in AI-assisted decision-making without differentiating whether such reliance is desirable. We then used over-reliance, under-reliance, and appropriate reliance to understand whether the reliance behavior that the subject exhibited was desirable or not. Intuitively, a desirable reliance behavior requires lower levels of over-reliance and under-reliance, and higher levels of appropriate reliance.

Based on our pre-registration, for all dependent variables, we conducted the one-way analysis of variance (ANOVA) to examine whether there are any significant differences in them across the 4 experimental treatments. When a significant difference was found, we used the Tukey HSD tests to conduct post-hoc pairwise comparisons.

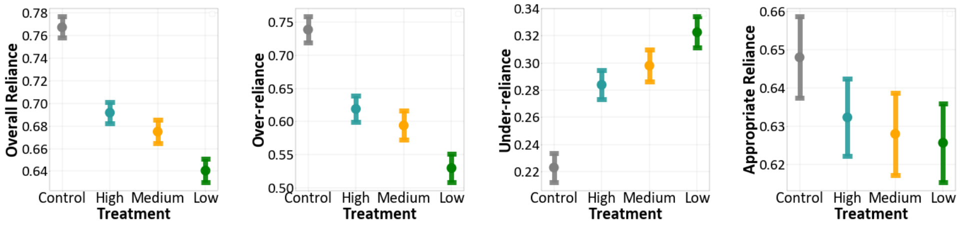
2.5 Experimental Results

In total, 428 subjects took our experiment HIT and passed the attention check (56.8% self-identified as male, 41.1% self-identified as female, and the most frequent age group reported by subjects was 25-34). To begin with, for the three treatments with peer judgements (i.e., Treatments 2–4), we checked the level of agreement between the AI model and the actual peer judgements presented to subjects. The average fraction of tasks in which the peer worker's judgement agreed with the AI model was 0.77, 0.51, 0.30 for treatments with high, medium, and low agreement peers, respectively, and a one-way ANOVA test confirms that the level of agreement between peers and the AI model across these three treatments is significantly different ($F(3, 424) = 790.34, p < 0.001$). This indicates that we successfully varied the peer-AI agreement level through our experimental design.

We then look into how the presence of second opinions from peers, which may show different levels of agreement with the AI, affects people's reliance on the AI model in AI-assisted decision-making.

Second opinions from peers decrease people's overall reliance on the AI model. Figure 2a shows subjects' average level of overall reliance on the AI model across the four treatments. Visually, it is clear that the presence of second opinions from peers results in a *decrease* in people's overall reliance on AI, and the more the peers disagree with the AI model, the more the reliance decreases. A one-way ANOVA test confirms that the difference in subjects' overall

[3] The pre-registration document can be found at: https://aspredicted.org/36J_RHS. All of our experiments were approved by the IRB of the authors' institution.



(a) Overall reliance (b) Over-reliance (c) Under-reliance (d) Appropriate reliance

Fig. 2. The effects of second opinions from peers on subjects' overall reliance (2a), over-reliance (2b), under-reliance (2c), and appropriate reliance (2d) on the AI model across treatments. Error bars represent the standard errors of the mean.

reliance across different treatments is statistically significant ($F(3, 8556) = 28.31, p < 0.001$). Furthermore, the post-hoc Tukey HSD test suggests that subjects in all treatments with peer judgements are less likely to rely on the AI model than those in the control treatment (i.e., control vs. with high agreement peers: $p < 0.001$, Cohen's $d = 0.17$; control vs. with medium agreement peers: $p < 0.001$, Cohen's $d = 0.21$; control vs. with low agreement peers: $p < 0.001$, Cohen's $d = 0.28$). In addition, the overall reliance difference shown between the two treatments with high agreement peers and low agreement treatment peers is also found to be significant (with high agreement peers vs. with low agreement peers: $p < 0.001$, Cohen's $d = 0.11$).

Second opinions from peers lead to lower over-reliance but higher under-reliance, and they do not significantly change the appropriate reliance. To understand whether the decrease in people's overall reliance on the AI model brought up by second opinions from peers is desirable, we further examine people's over-reliance, under-reliance, and appropriate reliance on AI separately. First, Figure 2b shows the comparison of subjects' over-reliance on the AI model across the four experimental treatments. It suggests that the presence of second opinions from peers helps subjects *reduce* their over-reliance on the AI model, especially when the peers' judgements have a relatively low level of agreement with the AI. The one-way ANOVA test indicates that the differences in subjects' over-reliance are significant across treatments ($F(3, 2136) = 17.27, p < 0.001$). Post-hoc Tukey HSD tests further show that these significant differences exist between the control treatment and every experimental treatment with peer judgements (control vs. with high agreement peers: $p < 0.001$, Cohen's $d = 0.26$; control vs. with medium agreement peers: $p < 0.001$, Cohen's $d = 0.31$; control vs. with low agreement peers: $p < 0.001$, Cohen's $d = 0.44$). At the same time, a significant difference also lies between the treatment with high agreement peers and the one with low agreement peers ($p = 0.010$, Cohen's $d = 0.18$).

While second opinions from peers bring about the benefit of decreased levels of over-reliance, these benefits also come with a cost. Specifically, Figure 2c shows subjects' average levels of under-reliance on the AI model in the four treatments. Here, we also find a significant difference across treatments (one-way ANOVA: $F(3, 6446) = 13.84, p < 0.001$). Post-hoc Tukey HSD tests show that compared to when the second opinions are absent, subjects significantly *increase* their under-reliance on the AI model when they receive the peers' judgements as the second opinions, regardless of how frequently the peers' judgements agree with the AI (i.e., control vs. with high agreement peers: $p < 0.001$, Cohen's $d = 0.14$; control vs. with medium agreement peers: $p < 0.001$, Cohen's $d = 0.17$; control vs. with low agreement peers: $p < 0.001$, Cohen's $d = 0.22$). This means that in AI-assisted decision-making when peer-generated second opinions are present, people decrease their reliance on the AI model no matter whether the AI model's prediction is correct or not.

Because of such behavior, as shown in Figure 2d, when examining subjects' appropriate reliance on the AI model across the experimental treatments, we even find a trend that the presence of second opinions from peers results in slight decreases in people's appropriate reliance/decision accuracy, although these decreases are not statistically significant ($p > 0.05$).

3 EXPERIMENT 2: PEER'S SECOND OPINIONS PRESENTED ONLY UPON DECISION MAKERS' REQUEST

Our Experiment 1 shows that providing second opinions from peers to people in AI-assisted decision-making can be an effective intervention to decrease their over-reliance on AI models, but it also comes with the undesirable side effect of increasing people's under-reliance on AI models. One reason for us to observe this side effect is the overly frequent presence of disagreeing second opinions on tasks where the AI model is correct, which sways subjects away from relying on the AI's correct decision recommendation. Thus, to mitigate this side effect, we are wondering instead of always providing peer-generated second opinions on all tasks, if these second opinions are presented only when the decision-makers actively *request* for them, can their presence help decrease over-reliance on AI models without increasing under-reliance?

To answer this question, we conducted our second pre-registered[4], randomized human-subject experiment, where subjects were again asked to complete the same set of 20 sentiment analysis tasks as those used in Experiment 1. Again, on each task, the prediction given by the same AI model as that used in Experiment 1 was presented to subjects as a decision recommendation. However, for those treatments where subjects *may* get access to second opinions generated by other peer workers, the second opinions are no longer presented on every task—subjects in these treatments were asked to decide whether they would like to see a second opinion from a peer worker and if so, they needed to click on a "Request" button to *solicit* the second opinion.

3.1 Experimental Design

Utilizing the same peer judgements as those collected from the pilot study of Experiment 1, we designed the following 3 treatments for Experiment 2[5]: Subjects in all treatments had access to the predictions of the AI model when completing each task. Moreover, subjects in **Treatment 1 (Control)** did not see any judgement made by other peer workers. In contrast, we would randomly select a worker from the three *high*/*low* agreement peers and present the selected worker's judgement on a task to the subject as a second opinion in **Treatment 2 (High agreement peers upon request)**/**Treatment 3 (Low agreement peers upon request)**, if subjects clicked on the "Request" button on a task. Figure 1b and 1c show an example of the task interface of Experiment 2 for treatments where subjects could solicit second opinions from peer workers (i.e., Treatment 2 or 3).

The procedure of Experiment 2 was identical to Experiment 1 except for the following differences: (1) Workers who participated in our Experiment 1 were not allowed to attend our Experiment 2; (2) To measure subjects' tendency to engage in deliberative thinking, we added a cognitive reflection test (CRT) [13, 38] in the exit-survey, which contained three mathematical questions that require people to utilize their cognitive reflection to override the intuitive, wrong answers (e.g., "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?"). In addition, the dependent variables and statistical analysis methods used in Experiment 2 are the same as those outlined in Section 2.4 for Experiment 1.

[4] The pre-registration document can be found at: https://aspredicted.org/MWC_PH1.

[5] As our results in Experiment 1 seem to suggest that the level of agreement between peers and the AI model mostly affect dependent variables in a linear way, we did not include a treatment with medium agreement peers in Experiment 2.

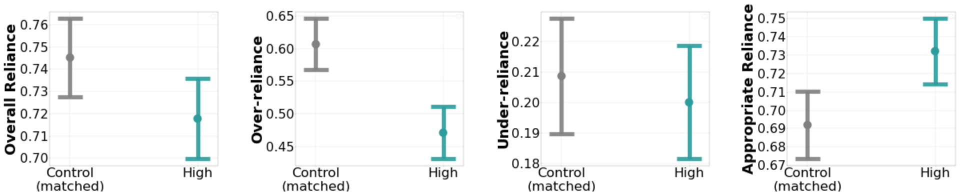
3.2 Experiment Results

In total, 336 subjects participated in Experiment 2 and passed the attention check (52.4% self-identified as male, 45.2% self-identified as female, and the most frequent age group reported by subjects was 25-34). Again, as a check of the effectiveness of our experimental manipulation, we confirmed that the actual peer judgements presented to subjects upon request in the treatment with high agreement peers agreed with the AI model significantly more than those peer judgements presented to subjects in the treatment with low agreement peers ($p < 0.001$).

Considering some subjects did not request second opinions at all during the experiment, we focus on only those subjects who requested second opinions from peers for at least once, and we aim to understand how their *active solicitations* of second opinions changed their reliance on the AI model in AI-assisted decision-making. Given the systematic differences in demographic backgrounds between subjects who had solicited or had never solicited second opinions, directly comparing the reliance behavior of those subjects who had solicited second opinions from high agreement peers (in Treatment 2) or low agreement peers (in Treatment 3) with that of all subjects in the control treatment can be misleading. To ensure the robustness of our analyses, we adopt *matching methods* to pair up subjects with similar demographic characteristics in the control treatment and the experimental treatment (i.e., Treatments 2 or 3), and then conduct comparisons between paired subjects.

We first conduct *propensity score matching* [30] for the 31 subjects in Treatment 2 who had solicited second opinions from high agreement peers for at least once. Specifically, given each subject in Treatments 1 (control) and 2 (high agreement peers upon request) of Experiment 2, we characterize them using all the demographic information that they self-reported in the exit-survey (e.g., age, gender, education, prior programming knowledge, CRT score, etc.), and we build a logistic regression model to predict a subject's treatment given their features (i.e., "covariates"). The predicted log-likelihood for a subject to belong to Treatment 2 is thus used as the subject's "propensity score." Then, for each of the 31 subjects in Treatment 2 who requested for second opinions at least once, we identify a subject in the control treatment with the closest propensity score (with replacement) to be their "*match*," hence these two subjects form a pair with very similar demographic characteristics, but one subject in the pair had the chance to solicit second opinions from high agreement peers while the other did not. After the matching, we find that between subjects who requested for second opinions at least once in Treatment 2 and their matches, paired t-tests suggest that there are no significant differences in the values for any of the covariates. Furthermore, between the requested subjects and their matches, the standard mean differences (SMD) for most of the covariates are less or equal to 0.1, which indicates that subjects in the pairs are comparable [2, 26].

Figures 3a–3d show the comparisons in subjects' overall reliance, over-reliance, under-reliance, and appropriate reliance on the AI model, respectively, between those subjects who requested for second opinions from high agreement peers for at least once and their matched subjects in the control treatment. Conducting paired t-tests on the 31 pairs of subjects, we find that subjects' active solicitations of second opinions from high agreement peers lead to a slight decrease in their overall reliance on the AI model (control: $M = 0.75, SD = 0.44$ vs. second opinions from high agreement peers solicited: $M = 0.72, SD = 0.45$, $p > 0.05$), and a significant decrease in their over-reliance on the AI model (control: $M = 0.61, SD = 0.49$ vs. second opinions from high agreement peers solicited: $M = 0.47, SD = 0.50$; $p = 0.009$). Importantly, we also find that the solicitation of second opinions from high agreement peers does *not* result in significant increases in subjects' under-reliance on the AI model ($p > 0.05$). As a result, as shown in Figure 3d, we observe a slight increase in subjects' appropriate reliance on the AI model when they requested for second opinions from high



(a) Overall reliance (b) Over-reliance (c) Under-reliance (d) Appropriate reliance

Fig. 3. The effects of active solicitations of second opinions from high agreement peers on subjects' overall reliance (3a), over-reliance (3b), under-reliance (3c), and appropriate reliance (3d) on the AI model. Data for the control treatment contains only the matched subjects after applying propensity score matching. Error bars represent the standard errors of the mean.

agreement peers at least once (control: $M = 0.69, SD = 0.46$ vs. second opinions from high agreement peers solicited: $M = 0.73, SD = 0.44$), although the difference is not statistically significant at the level of $p = 0.05$.

We then repeat the propensity score matching process for the 44 subjects in Treatment 3 who had solicited second opinions from low agreement peers for at least once. Here, we find that compared to their matched subjects in the control treatment, subjects in Treatment 3 who solicited second opinions from low agreement peers for at least once significantly reduced their overall reliance ($p < 0.001$) and over-reliance ($p = 0.003$) on the AI model, but they also exhibited significantly higher levels of under-reliance on the AI model ($p = 0.003$). Together, the active solicitations of second opinions from low agreement peers result in a slight decrease in subjects' appropriate reliance on the AI model (control: $M = 0.70, SD = 0.46$ vs. second opinions from low agreement peers solicited: $M = 0.68, SD = 0.47$), although the decrease is not statistically significant.

Together, these results show the promise of utilizing second opinions from peers to help people reduce their over-reliance on an AI model while not increasing their under-reliance—this goal can be achieved by enabling people to actively *solicit* second opinions from those peers who have a relatively high level of agreement with the AI model. We conjecture that this approach is effective because (1) people solicit second opinions more frequently on tasks where the AI model is wrong (i.e., disagreements between peers and the AI model on these tasks lead to lower over-reliance), and (2) the relatively high level of agreement between the peers and the AI model minimizes the chance that people get misled by incorrect peer judgements on tasks where the AI model is correct (i.e., under-reliance is not increased).

4 LIMITATIONS AND FUTURE WORK

Our study was conducted with laypeople (i.e., subjects recruited from MTurk) on a decision-making task that does not require much expertise yet still seems to be not easy for laypeople (e.g., in our pilot study, the average decision accuracy of the crowd workers on our selected sentiment analysis task is 64.34%) with AI-assistance having a specific level of performance (i.e., 75% accuracy). Cautions should be used when generalizing the results of this work to different settings, such as for a different population of people, for tasks that are much easier, or for tasks that require substantially more domain expertise. In addition, the second opinion in our experiment came from randomly selected crowd workers that our subjects did not know. In reality, the peers from whom people in AI-assisted decision making seek help can be someone that they are quite familiar with and naturally trust, which may significantly change how they respond to the agreements or disagreements between the AI recommendations and the peer's second opinions. Our current study only reveals the effects of providing peer-generated second opinions on decision-makers in AI-assisted decision making, but does not attempt to separate the effects brought up by the presence of the second opinion and the effects brought up by the source of the second opinions, and future studies can investigate into this separation in more details. In addition, the advice structure between the decision-makers and the "advisors" is also not limited to what we've studied in this work. For example, multiple second opinions can be provided instead of one, and the second opinions may come from not only human peers but also other AI models. Understanding how second opinions under these settings affect decision-makers' behavior and performance in AI-assisted decision-making is another exciting future work.

5 CONCLUSION

In this paper, we explore the effect of providing second opinions from human peers to people on their behavior and performance in AI-assisted decision-making. Via two pre-registered, randomized experiments, we show that always presenting second opinions from peers along with the AI recommendation can reduce decision-makers' over-reliance on AI, but it also increases decision-makers' under-reliance on AI. Nevertheless, by enabling decision-makers to actively solicit second opinions from peers as needed, we show that the active solicitations of second opinions have the promise to reduce decision-makers' over-reliance on the AI model without increasing the under-reliance, if the peer judgements agree with the AI model at a relatively high level. Our results highlight the potential benefits and risks of presenting second opinions from peers to people in AI-assisted decision making for promoting the human-AI team performance. We hope this work could open more discussions on understanding the effects of second opinions in AI-assisted decision-making and better utilizing them as an intervention to enhance human-AI collaboration in decision-making.

REFERENCES

[1] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. 2021. AI-Assisted Human Labeling: Batching for Efficiency without Overreliance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[2] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46, 3 (2011), 399–424.

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[4] André Betzer and Jan Philipp Harries. 2022. How online discussion board activity affects stock trading: the case of GameStop. *Financial markets and portfolio management* (2022), 1–30.

[5] Silvia Bonaccio and Reeshad S Dalal. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes* 101, 2 (2006), 127–151.

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

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

[8] Chun-Wei Chiang and Ming Yin. 2021. You'd better stop! Understanding human reliance on machine learning models under covariate shift. In *13th ACM Web Science Conference 2021*. 120–129.

[9] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248* (2020).

[10] Maria De-Arteaga, Alexandra Chouldechova, and Artur Dubrawski. 2022. Doubting AI Predictions: Influence-Driven Second Opinion Recommendation. *arXiv preprint arXiv:2205.00072* (2022).

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

[12] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[13] Shane Frederick. 2005. Cognitive reflection and decision making. *Journal of Economic perspectives* 19, 4 (2005), 25–42.

[14] Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831* (2020).

[15] Yoyo Tsung-Yu Hou and Malte F Jung. 2021. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[16] Siu Cheung Hui and G Jha. 2000. Data mining for customer service support. *Information & Management* 38, 1 (2000), 1–13.

[17] Timothy R Johnson, David V Budescu, and Thomas S Wallsten. 2001. Averaging probability judgments: Monte Carlo analyses of asymptotic diagnostic value. *Journal of Behavioral Decision Making* 14, 2 (2001), 123–140.

[18] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. " Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[19] Pahulpreet Singh Kohli and Shriya Arora. 2018. Application of machine learning in disease prediction. In *2018 4th International conference on computing communication and automation (ICCCA)*. IEEE, 1–4.

[20] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* 13 (2015), 8–17.

[21] Mahinda Mailagaha Kumbure, Christoph Lohrmann, Pasi Luukka, and Jari Porras. 2022. Machine learning techniques and data for stock market forecasting: a literature review. *Expert Systems with Applications* (2022), 116659.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[23] Suwan Long, Brian Lucey, Ying Xie, and Larisa Yarovaya. 2022. "I just like the stock": The role of Reddit sentiment in the GameStop share rally. *Financial Review* (2022).

[24] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. http://www.aclweb.org/anthology/P11-1015

[25] Hoang Nguyen, Le-Minh Kieu, Tao Wen, and Chen Cai. 2018. Deep learning methods in transportation domain: a review. *IET Intelligent Transport Systems* 12, 9 (2018), 998–1004.

[26] Sharon-Lise T Normand, Mary Beth Landrum, Edward Guadagnoli, John Z Ayanian, Thomas J Ryan, Paul D Cleary, and Barbara J McNeil. 2001. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of clinical epidemiology* 54, 4 (2001), 387–398.

[27] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.

[28] Jigar Patel, Sahil Shah, Priyank Thakkar, and Ketan Kotecha. 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications* 42, 1 (2015), 259–268.

[29] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[30] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[31] Sahar F Sabbeh. 2018. Machine-learning techniques for customer retention: A comparative study. *International Journal of Advanced Computer Science and Applications* 9, 2 (2018).

[32] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making. *arXiv preprint arXiv:2204.06916* (2022).

[33] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A Meta-Analysis on the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. *arXiv preprint arXiv:2205.05126* (2022).

[34] Philipp Schmidt and Felix Biessmann. 2019. Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558* (2019).

[35] Thomas Schultze, Andreas Mojzisch, and Stefan Schulz-Hardt. 2017. On the inability to ignore useless advice: A case for anchoring in the judge-advisor-system. *Experimental Psychology* 64, 3 (2017), 170.

[36] Linda J Skitka, Kathleen L Mosier, Mark Burdick, and Bonnie Rosenblatt. 2000. Automation bias and errors: are crews better than individuals? *The International journal of aviation psychology* 10, 1 (2000), 85–97.

[37] Minhyang Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. Ai as social glue: Uncovering the roles of deep generative ai during social music composition. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–11.

[38] Maggie E Toplak, Richard F West, and Keith E Stanovich. 2011. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition* 39, 7 (2011), 1275–1289.

[39] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, and Arun Kumar. 2020. Stock closing price prediction using machine learning techniques. *Procedia computer science* 167 (2020), 599–606.

[40] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[41] Ilan Yaniv. 2004. The benefit of additional opinions. *Current directions in psychological science* 13, 2 (2004), 75–78.

[42] Ilan Yaniv and Eli Kleinberger. 2000. Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes* 83, 2 (2000), 260–281.

[43] Ilan Yaniv and Maxim Milyavsky. 2007. Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes* 103, 1 (2007), 104–120.