UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Effective but untrustworthy: How artificial intelligence bias opposing human bias affects judgments

Permalink
https://escholarship.org/uc/item/74g6w6fw

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Shirasuna, Masaru
Honda, Hidehito
Kagawa, Rina

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Effective but untrustworthy: How artificial intelligence bias opposing human bias affects judgments

Masaru Shirasuna (m.shirasuna1392@gmail.com)
Faculty of Informatics, Shizuoka University, 3-5-1, Johoku, Chuo-ku, Hamamatsu-shi, Shizuoka 432-8011, Japan

Hidehito Honda (hitohonda.02@gmail.com)
Department of Artificial Intelligence and Cognitive Science, Otemon Gakuin University,
2-1-15, Nishiai, Ibaraki-shi, Osaka, 567-8502, Japan

Rina Kagawa (kagawa-r@aist.go.jp)
Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST),
1-1-1, Umezono, Tsukuba-shi, Ibaraki, 305-8560 Japan

Abstract

Today, people make judgments with the help of artificial intelligence (AI) assistance in many situations, such as medical diagnoses. Although many studies have examined the effects of AI assistance, they have mainly focused on aspects of AI (e.g., AI's accuracy). Here, we emphasize the importance of interactions between AI and human biases. A highly accurate AI may not always be a promising intervention; rather, AI with biases (especially in the direction opposite to individuals' biases) may work effectively because AI's biases may cancel out individuals' biases (e.g., individuals' overestimation bias may be corrected by AI's underestimation bias). We investigated these issues using a simple perceptual task assuming medical judgments. First, computer simulations showed that appropriate AI assistance would differ depending on individuals' prior beliefs. Behavioral experiments demonstrated that AI with biases in the direction opposite to participants' biases could effectively reduce their biases. However, participants tended to evaluate AI with biases in the same direction as their own and considered it more trustworthy. Our theoretical and empirical results raise questions about conventional beliefs that more accurate, trustworthy AI should be better. Our findings will provide practical implications for designing AI as a collaborator of people.

Keywords: AI-assisted judgments; trustworthy; computer simulation; behavioral experiment

Introduction

With the rapid development of the digital age, artificial intelligence (AI) has become a collaborator for people. Today, people make judgments with the help of AI assistance in many situations (e.g., Akata et al., 2020). For example, in a medical situation where doctors judge whether a patient has a disease by observing their medical images (e.g., X-ray images; magnified photographs of a cell), AI or diagnosis assistance systems sometimes help the doctors' judgments (e.g., "positive" or "negative") by providing its judgment (Reverberi et al., 2022; Topol, 2019). In such situations of human-AI collaboration, there is a growing need for effective AI assistance in enhancing individuals' cognitive competence for making accurate judgments.

AI-assisted judgments and the wisdom of crowds

Regarding decision-making support, the development of AI itself, including AI performance and technology, is generally focused on. However, we focus on the interactions between AI's biases and individuals' biases. Given the age of human-AI collaboration, it is not sufficient if only AI's (or human's) accuracy is high. Instead, we should consider the accuracy of judgments made by humans *with* AI together (e.g., Steyvers & Kumar, 2024). When two or more individuals make judgments, individuals' biases tend to be canceled out if there are diverse (not uniform) judgments in the group by simply aggregating individuals' judgments (e.g., averaging). This effect is known as the wisdom of crowds in cognitive science (e.g., Navajas et al., 2018; Surowiecki, 2004). An essential factor in achieving the wisdom of crowds is the diversity of individuals' judgments in a group (e.g., Herzog et al., 2019; Lorenz et al., 2011). Even if some individuals in a group have biases and make errors, their biases are sometimes canceled out if they make different and diverse judgments. By contrast, if individuals in a group make similar and uniform judgments, their biases are unlikely to be corrected, and the group judgment sometimes goes in the wrong direction.

Based on the wisdom of crowds framework, highly accurate AI assistance may not always be effective. Rather, AI with some biases, especially biases that are in the inverse direction to human biases, can sometimes be more effective. It is because AI is likely to cancel out the cognitive biases that humans have (see also Chiang et al., 2024). For example, if a person has overestimation biases, AI with underestimation biases may effectively help improve her/his judgment accuracy compared to AI with no biases.

Furthermore, another aspect is needed to be investigated: trustworthy AI (Krügel et al., 2022; Zhang et al., 2020). Generally, people are likely to accept similar opinions to their own (e.g., Yaniv, 2004). Thus, it is predicted that individuals may feel less trustworthy when they observe an AI with opposite biases, even if accepting such AI improves their judgment accuracy. For example, a person with an overestimation bias may evaluate AI with an underestimation bias as untrustworthy. We also need to examine such a paradox.

Study outline

Investigating these issues will provide practical implications for designing more effective interventions; in other words, better ways to collaborate with AI). We examine these issues

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

using basic experimental materials. Specifically, we place medical classification judgment situations into simple perceptual tasks (Vicente & Matute, 2023). We first conduct theoretical analyses and show that appropriate assistance that can lead to accurate estimation depends on humans' biases (i.e., prior beliefs for a target task). We then conduct experimental studies and demonstrate whether our predictions can be supported based on actual human behaviors, in terms not only of effects of AI assistance but also of trustworthy AI. Data, code, and Appendix are available at https://osf.io/3ka74/?view_only=c22254855ac4429ebb9ae6e7a89948a6

Theoretical analyses

First, using computer simulations, we theoretically show that appropriate interventions to enhance accurate judgments differ depending on individuals' prior beliefs about tasks.

Method

Task and materials We used a hypothetical medical task based on Vicente and Matute (2023). The assumptions for this task were as follows (Figure 1). A person was presented with an image comprising dark-pink and light-yellow tiles, which simulated a hypothetical patient's cell. The patient has Disease X (i.e., "positive") if the proportion of dark areas was more than 50% of the whole, while does not have X (i.e., "negative") if that of dark areas was less than 50%. At the same time, the person was also presented with AI assistance, which judged the patient as "positive" or "negative". For each image, the person was asked to estimate the proportion of dark areas based on the presented AI assistance.

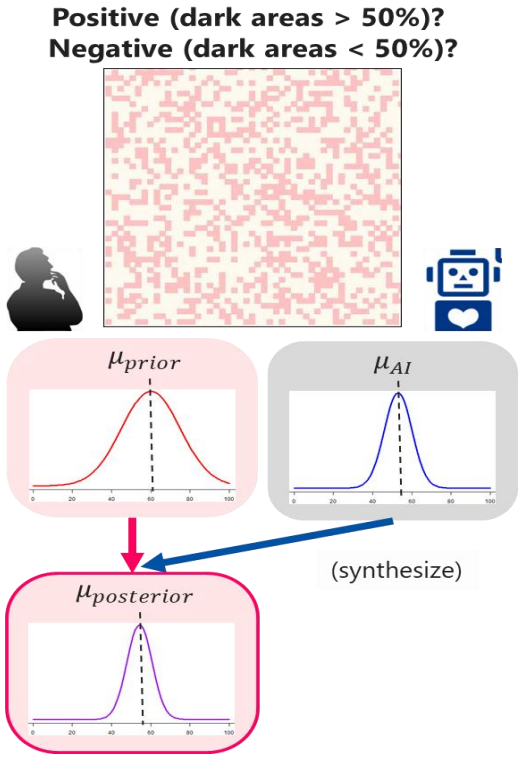


Figure 1 Schematics of computer simulation of the task (estimate of the proportion of dark areas). It was assumed that distributions of a person's prior belief and AI assist were synthesized and then the person's distribution of posterior belief was generated.

Model assumption and parameter settings It is assumed that a person first observed a stimulus and estimated the proportion, which is a prior belief. Then, the person observed AI assistance and then updated the initial estimation, which is a posterior belief. Formally, we assume that individuals' prior belief and AI's accuracy can be described as a form of probability distribution, and individuals' posterior belief after observing AI can be described as a synthesized distribution of these two distributions, like a Bayesian update framework (Honda et al., 2024; Turner & Schley, 2016). Based on Honda et al. (2024), an individual had a prior belief following a normal distribution with mean, μ_{prior}, and SD, SD_{prior}. AI made its estimation following a normal distribution with mean, μ_{AI}, and SD, SD_{AI}. Then, the individual's final, posterior distribution of estimation was defined as a synthesized normal distribution of them, which follows the mean $\mu_{posterior}$, and SD, $SD_{posterior}$:

$$\mu_{posterior} = \left.\frac{\frac{\mu_{prior}}{SD^2{}_{prior}} + \frac{AI_{prior}}{SD^2{}_{AI}}}{\frac{1}{SD^2{}_{prior}} + \frac{1}{SD^2{}_{AI}}}\right/$$

$$SD_{posterior} = \left(\frac{1}{SD^2{}_{prior}} + \frac{1}{SD^2{}_{AI}}\right)^{-1/2}$$

The SD was interpreted as the strength of an individual's belief. A smaller SD meant a stronger belief, which could be interpreted as having stronger confidence and being likely to persist in the individual's own estimation; and vice versa.

Generally, people sometimes have cognitive biases such as overestimation (i.e., classifying as "positive" even if dark pink areas were less than half of the whole) or underestimation (i.e., classifying as "negative" even if they were more than half). Furthermore, not all AI is accurate; some AI will show biases. In our simulations, to investigate interactions between individuals' and AI's biases, the values of both μ_{prior} and μ_{AI} were manipulated from 40 to 60 in increments of 0.2 (i.e., 40.0, 40.2, 40.4, …, 59.8, 60.0; 101 * 101 = 10,201 patterns). SD_{prior} was set at 2, 5, or 10 (three patterns), and SD_{AI} was fixed 5 (one pattern). Thus, we conducted simulations for 101 * 101 * 3 * 1 = 30,603 parameter patterns. Then, for each SD_{prior}, we investigated how $\mu_{posterior}$ and would be affected by the μ_{prior} and μ_{AI}. In this task, because a criterion of classification for positive or negative was 50 (i.e., the proportion of dark areas was over or under 50%), it was ideal that $\mu_{posterior}$ became 50.

Results and discussion

The simulation results are shown in Figure 2. The colors of the heatmap denote $\mu_{posterior}$ (colored by the extent of deviations from 50). It was theoretically clarified that individuals

tended to persist in their initial estimations when they had stronger prior beliefs. It was because $\mu_{posterior}$ was not likely changed from μ_{prior} in SD_{prior} = 2 (i.e., white areas were observed vertically), while was likely changed in SD_{prior} = 10 (i.e., white areas were observed horizontally). More importantly, it was also shown that accurate AI (i.e., μ_{AI} = 50) were not always optimal. For example, in μ_{prior} = 60 with SD_{prior} = 5, the $\mu_{posterior}$ became closer to 50 when μ_{AI} was 40, not 50. In other words, the optimal AI that could make a prior estimation closer to the ideal value depended on individuals' prior belief distributions. In short, these results suggest that the relationship between human and AI biases is important for improving judgment accuracy.

Behavioral experiment

How about actual human behaviors? Can biased AI work effectively to improve judgment accuracy? Furthermore, how much can people trust such a biased AI? Next, we investigated whether the theoretical findings were consistent with actual human behaviors and aimed to address these issues, through behavioral experiments.

Method

Participants A total of 512 Japanese people participated in this study (M_{age} = 42.71, SD_{age} = 10.36; n_{men} = 277, n_{women} = 227, n_{others} = 8). Because this experiment had five conditions as described later and we decided to assign approximately 100 participants to each condition based on our preliminary experiments and our study resources ($n_{positive+}$ = 100, $n_{positive}$ = 104, $n_{accurate}$ = 103, $n_{negative}$= 103, and $n_{negative+}$ = 102), we recruited approximately 500 participants. Note that one participant was omitted from the subsequent analyses because s/he answered "negative" to all questions. All participants were recruited via a Japanese crowdsourcing service, Crowdworks. The experimental protocols conformed to the Declaration of Helsinki and were approved by the Ethics Review Committee for Experimental Research at Otemon Gakuin University.

Tasks and materials We conducted a classification task simulating medical judgments, as in the above theoretical analyses. Participants were presented with a hypothetical medical image simulating a patient's cell. This image consisted of a 50*50 matrix. Dark-pink and light-yellow tiles were randomly distributed in this matrix. It was assumed that if the dark pink areas were over 50% of a whole, a patient has Disease X. Participants were asked to classify whether a person had Disease X ("positive") or not ("negative"). Regarding the percentage of dark areas, we prepared 41 types: 30%, 31%, 32%, …, 69%, and 70% (i.e., from 30 to 70 in increments of 1). There were two patterns for each percentage; thus, we used 41 * 2 = 82 images as experimental stimuli.

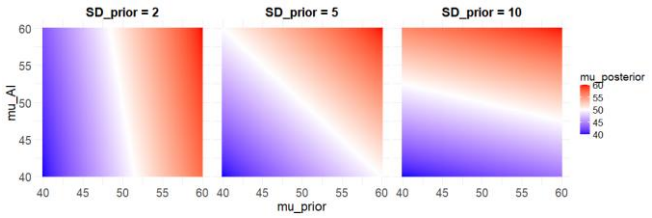


Figure 2 Results of computer simulations. The x- and y-axes show μ_{prior} and μ_{AI}, respectively. Left, middle, and right panels show the results under SD_{prior} = 2, 5, and 10, respectively. Colors show $\mu_{posterior}$. Red, blue, and white colors show overestimated, underestimated, and accurate estimations, respectively.

Procedure The experiment was conducted using an online platform, Qualtrics. In one question, a stimulus simulating a patient's cell, and "positive" and "negative" buttons appeared in the screen's center, and bottom-left, and bottom-right, respectively. A question sentence was also presented above the image. After participants clicked one of the two buttons and then clicked the "next" button, the next question was presented. They conducted 82 questions in one set (i.e., all the stimuli described above were presented one by one), which was repeated in three sets. The order of presentation of the stimuli was randomized. After completing all three sets, participants provided their age and gender.

The first and third sets were conducted using the above-mentioned procedure (Figure 3, left). Only in the second set, a (fictional) AI assistance was also presented for each question. Specifically, a picture of robot illustration and its judgment appeared next to a stimulus (Figure 3, right). Before starting the second set, participants were instructed: "In this set, an AI judgment is presented as well as a medical image. This AI judged the patient as 'positive' or 'negative' prior to this experiment. Please classify based on this AI judgment". By inserting such interventions in the second set, we intended to investigate effects of various AI assistance (see **Conditions** section). In addition, after the second set, participants asked to rate their trustworthy for the presented AI. This questionnaire was based on Hoffman et al. (2023) (partially modified to match the context of our experiment) and consisted of seven questions such as "I feel that this AI works well" (for all questions, see Appendix). Participants were asked to rate how much they agreed or disagreed using a visual analog scale ranging from 0 (not at all) to 100 (extremely).

Conditions In the second set, participants were randomly assigned to one of the following five conditions:

- Positive+: AI says "positive" if the proportion of dark areas was over 40% (i.e., extreme overestimation)

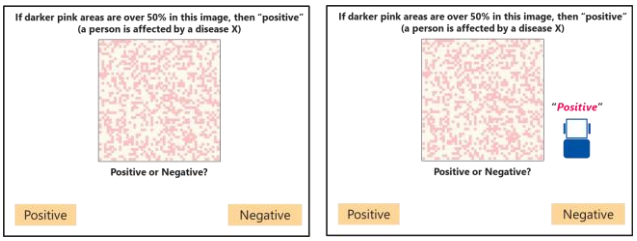


Figure 3 Examples of stimuli in behavioral experiments. For each question, only a fictional medical image was presented in the first and third sets (left), while AI assistance as well as the image were presented in the second set (right).

- Positive: AI says "positive" if the proportion of dark areas was over 45% (i.e., overestimation)
- Accurate: AI says "positive" if the proportion of dark areas was over 50% (i.e., no bias)
- Negative: AI says "positive" if the proportion of dark areas was over 55% (i.e., underestimation)
- Negative+: AI says "positive" if the proportion of dark areas was over 60% (i.e., extremely underestimation)

Estimation of discrimination point Based on the behavioral data, we estimated individuals' judgment biases as a participant's criterion for switching between negative and positive judgments, called "discrimination point". This was defined as an inflection point of a sigmoid function fitting to individuals' responses, with independent and dependent variables were the percentage of dark areas and the rate of "positive" responses, respectively (Figure 4). We conducted a logistic regression on individuals' responses for each set (82 responses per participant in one set). Then, we regarded the x-axis value at the inflection as the participant's discrimination point. If the discrimination point was less than 50, it could be interpreted that they had overestimation biases because they tended to judge a patient as "positive" even if the percentage of dark pink areas in a stimulus was less than 50%. By contrast, if the discrimination point was more than 50, it was interpreted that they had underestimation biases.

Note that in addition to the discrimination point, individuals' judgment biases could be characterized by the slope of the sigmoid curve. The curve could be interpreted as a participant's sensitivity to discrimination. For example, a participant with a steep (gentle) slope was likely (unlikely) to drastically change the rate of "positive" responses when the percentage of dark areas changed slightly. Although we also estimated the slope for each participant (see Appendix), we omitted these analyses from the main text because of out of focuses of this study.

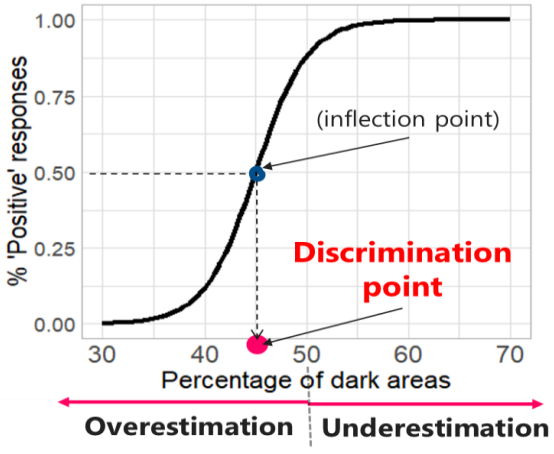


Figure 4 Discrimination point. This was a person's criterion for shifting negative/positive responses and was defined as an inflection of sigmoid curve (fitting to individual's behavioral data). If discrimination point was below (above) 50, the person was regarded as having over (under) estimation bias.

Results and discussion

In general, the accuracy (i.e., rate of correct judgments) was low for stimuli whose proportion of dark areas was close to 50% (difficult), and high for stimuli whose proportion was far from 50% (easy) (see Appendix). It can be considered that AI assistance will not be required if the questions are easy; rather, it is important to consider how AI assistance can effectively work in difficult questions. Therefore, hereafter, we defined the stimuli whose proportion of dark pink areas was 43-53% as "difficult questions" and analyzed behavioral data focusing on the difficult questions. (Note that the mean and 1st quantile of accuracy in all questions were .887 and .825, respectively, and the mean accuracy in 43-53% stimuli was below this 1st quantile value, .825. We regarded this as a criterion for splitting difficult and easy questions.)

Effects of AI assistance on human judgments Did participants' accuracy and discrimination points change by observing AI assistance? We investigated the changes in participants' judgments between sets. We used a two-way ANOVA model in which independent variables were set (first, second, third) and condition (positive+, positive, accurate, negative, negative+), and a dependent variable was discrimination point or accuracy using an R package "brms" (Bürkner, 2017). In each set, we first calculated dependent variables for each participant. Then, based on a Markov Chain Monte Carlo method with 4,000 iterations, 2,000 burn-in, and 4 chains, we estimated these dependent variables' mean and 95% credible intervals (CIs). We show the results in Figure 5 (for descriptive statistics, see Appendix).

We first focused on the discrimination point that participants might have before observing AI (i.e., in the first set; denoted as "set01", black bars of Figure 5). Overall, the discrimination points in many participants were below 50% (dotted horizontal lines), which indicates that participants tended to have overestimation biases in this task. Although there was no clear evidence as to why there were more over estimators than under estimators, it was speculated that darker tiles were more prominent and noticeable (and were likely to be perceived as existing more) than lighter tiles.

How did their prior discrimination points change when they observed AI? Because this task required participants to judge whether the darker areas were more than 50%, the ideal discrimination point was 50. As shown in "set02" (red bars in Figure 5), discrimination points tended to shift in the same directions as AI biases. Specifically, discrimination points (i) shifted above from 50 under positive+ and positive conditions (i.e., overestimation biases became greater); (ii) were not greatly shifted under accurate condition (i.e., biases were kept); and (iii) shifted closer to 50 under negative and negative+ conditions (i.e., overestimation biases were cancelled out) (in the second set, $M_{\text{positive+}} = 44.3$, $M_{\text{positive}} = 46.2$, $M_{\text{accurate}} = 48.0$, $M_{\text{negative}} = 49.4$, $M_{\text{negative+}} = 49.2$). In addition, the 95%CI did not overlap between the first and second sets, both under positive+ and positive conditions (positive+ CI_{set01} = [46.2, 48.3], CI_{set02} = [43.2, 45.3]; positive CI_{set01} = [47.3, 49.2], CI_{set02} = [45.2, 47.2]). These results suggest that the AI

with overestimation biases significantly changed participants' discrimination points.

Next, the participants' accuracy was examined. In this task, the accuracy was expected to decrease if the discrimination points were far from 50. As a result, 95%CIs of accuracy between the first and second sets did not overlap under positive+ and positive conditions (positive+ CI_{set01} = [.615, .691], CI_{set02} = [.475, .551]; positive CI_{set01} = [.662, .738], CI_{set02} = [.577, 653]), which indicates the significant decrease in accuracy by observing AI with overestimation.

If we simply focus on the values of accuracy in the second set, the accurate condition had the highest accuracy among the five ($M_{positive+}$ = .513, $M_{positive}$ = 615, $M_{accurate}$ = 757, $M_{negative}$ = 738, $M_{negative+}$ = 702). That is, if the accurate AI was presented, participants were likely to make accurate judgments. However, and importantly, it could be equally effective even when AI with biased estimations was presented. Specifically, if AI had a bias in the (moderately) inverse direction to participants' biases, the AI was likely to cancel out (i.e., the cases where AI had underestimation, because most participants had overestimation in this experiment; see set02 in negative condition). However, even in the opposite direction, extremely biased assistance was unlikely to work effectively (i.e., $M_{negative+}$ = 702 was lower than $M_{negative}$ = 738).

Relationships between performance and trustworthiness

We also analyzed participants' trust in AI assistance. As an indicator of trust, we calculated the mean of rating scores in seven questions asked after the second set (0 as "not at all" - 100 as "extremely" in six questions; but 100 as "not at all" - 0 as "extremely" in one question as an inverse item) for each participant. We examined the relationships between participants' discrimination points in the first set and their ratings of trust. As a result, significantly negative correlations were observed under positive+ and positive conditions ($r_{positive+}$ = -.512, $r_{positive}$ = -.429), and positive correlations were observed under negative and negative+ conditions ($r_{negative}$ = .456, $r_{negative+}$ = .410) (all ps < .001; Figure 6). These correlations indicate that participants tended to evaluate AI with biases in the same direction as being more trustworthy. Based on the results in the previous sections, it is desirable for participants to accept AI with the opposite bias to their biases. However, based on the results of trustworthiness, such opposing AI tended to be evaluated as less trustworthy.

The integration of these results implies that simply accepting trustworthy AI is not always a good strategy. It is sometimes important to accept untrustworthy AI because such AI may have biases in the opposite direction to individuals' biases and are likely to improve their judgment accuracy.

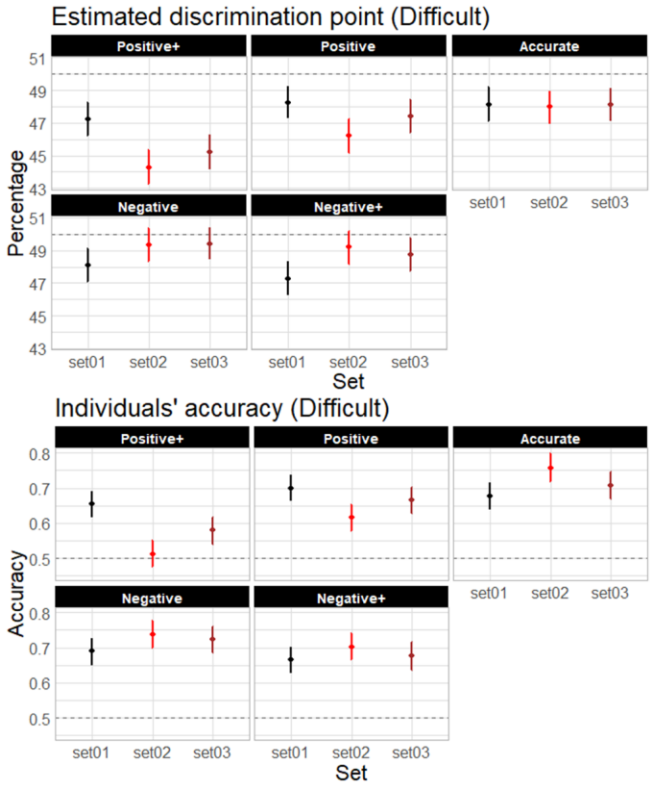


Figure 5 Discrimination point (upper) and accuracy (lower) in behavioral experiments. Each panel shows each condition. The term "set01", "set02", and "set03" in x-axis denote the first, second, and third sets, respectively. Error bars show 95%CI.

General discussion

Generally, AI with higher accuracy is expected to be more desirable for improving human judgment accuracy. However, we predict that AI with a bias, especially in the opposite direction of humans, can effectively counteract human biases. For example, based on the wisdom of crowds framework, humans' overestimation biases may be effectively cancelled out by AI's underestimation judgments. This study addressed such issues through both theoretical (computer simulations) and empirical (behavioral experiments) approaches. We found that (i) appropriate AI assistance for improving individuals' judgment accuracy depended on cognitive biases that individuals have; (ii) AI with a bias in the moderately (not extremely) inverse direction to individuals' biases could effectively reduce their biases; and (iii) individuals tended to evaluate AI with biases that were identical direction to their own as more trustworthy. We revealed the (negative) correlation between AI trustworthiness and improvement of human judgments. This study emphasizes the importance of interactions between humans and AI biases. Our study will serve as scaffolding for a better and more effective collaboration between humans and AI.

Practical implications

We believe that our study has some implications for the design of AI systems and AI-assisted judgments. When developing AI systems to collaborate with humans, it is generally thought that a more reliable and trustworthy AI is desirable (e.g., Jacovi et al., 2021). However, our findings raise a question to this. Given that individuals generally accept others whose opinions and social categories seem similar to themselves (e.g., Tajfel et al., 1971), AI that individuals feel as trustworthy may sometimes have same biases as they have. If so, the AI will lead the individuals in the wrong direction. Therefore, AI designers should consider biases that

individuals have and may sometimes need to "moderately" decrease AI's trustworthiness by, for example, making AI have slightly different opinions from individuals.

On the other hand, individuals should not reject AI assistance just because it is untrustworthy. Previous studies on advice-taking have shown that individuals tend to adhere to their initial judgment, and it is difficult to accept others' opinions (e.g., Yaniv & Kleinberger, 2000). Individuals tend to be more cooperative with those with similar opinions and beliefs (e.g., Balliet et al., 2014). However, if they rely only on AI with similar opinions to them, they may be unable to notice and correct their mistakes. Thus, it will sometimes be important to accept AI with different opinions.

Future research should further investigate some issues. The first issue is the duration of effects. Even if AI helped improve participants' judgment accuracy, could the effects be maintained after AI was removed? Our behavioral experiment briefly examined the duration by setting the third set. Additional analyses revealed that in the third set, the accuracy in negative condition was slightly higher than that in accurate condition (accurate .707; negative .723). Furthermore, the accuracy in the negative condition was likely to be maintained from the second to third sets compared to that in the accurate condition: The differences in the mean of accuracy (third set minus second set) were –.049 (= .707 – .756) and –.015 (= .723 – .738) under accurate and negative conditions, respectively. However, these results were observed right after AI was removed. The duration should be investigated with a longer time span such as after several hours or days.

The second issue is about the estimation of individuals' biases. If one tries to give individuals AI with a bias that is opposite to individuals' biases, it is firstly needed to know the direction of individuals' biases by collecting some data before giving AI to individuals (just like the first set in the experiment). However, how long should one go on such an initial data collection? For practical implementation of our findings, such issues will also need to be resolved.

The final issue is applicability and generalizability. This study used only a simple perceptual task simulating medical diagnoses. To examine whether and to what extent our results can be applied to the real world, we should conduct the same task in, for example, actual medical diagnoses situations for medical professionals (see also Croskerry et al., 2023; Kurvers et al., 2016). Furthermore, we may also need to conduct other tasks using the same experimental framework to examine the generalizability.

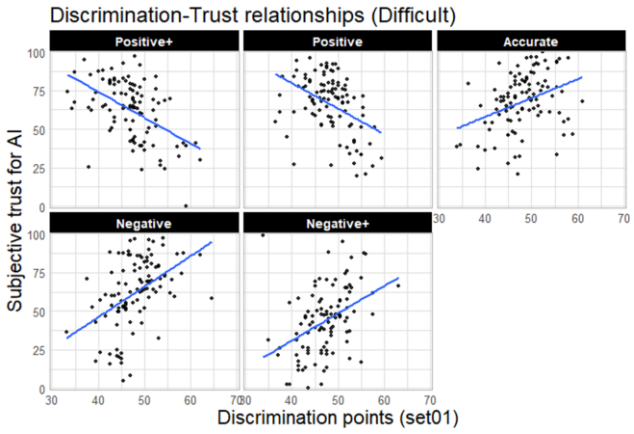


Figure 6 Relationships between discrimination point in the first set and ratings for AI trustworthy. Each point denotes each participant.

Acknowledgments

This study was supported by JST PRESTO (JPMJPR23I3) and JSPS KAKENHI (23K16249, 23K25169).

References

Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., Van Der Gaag, L., Van Harmelen, F., … Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect with Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, *53*(8), 18–28. https://doi.org/10.1109/MC.2020.2996587

Balliet, D., Wu, J., & De Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin*, *140*(6), 1556–1581. https://doi.org/10.1037/a0037737

Becker, F., Skirzyński, J., Opheusden, B. van, & Lieder, F. (2022). Boosting human decision-making with AI-generated decision aids. *PsyArXiv*, 1–30.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chiang, C.-W., Lu, Z., Li, Z., & Yin, M. (2024). Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 103–119. https://doi.org/10.1145/3640543.3645199

Croskerry, P., Campbell, S. G., & Petrie, D. A. (2023). The challenge of cognitive science for medical diagnosis. *Cognitive Research: Principles and Implications*, *8*(13). https://doi.org/10.1186/s41235-022-00460-z

Feufel, M. A., Keller, N., Kendel, F., & Spies, C. D. (2023). Boosting for insight and/or boosting for agency? How to maximize accurate test interpretation with natural frequencies. *BMC Medical Education*, *23*(75). https://doi.org/10.1186/s12909-023-04025-6

Herzog, S. M., Litvinova, A., Yahosseini, K. S., Tump, A. N., & Kurvers, R. H. J. M. (2019). The ecological rationality of the wisdom of crowds. In *Taming Uncertainty* (pp. 245–262). The MIT Press.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*. https://doi.org/10.3389/fcomp.2023.1096257

Honda, H., Kagawa, R., & Shirasuna, M. (2024). The nature of anchor-biased estimates and its application to the wisdom of crowds. *Cognition*, *246*, 105758. https://doi.org/10.1016/j.cognition.2024.105758

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. https://doi.org/10.1145/3442188.3445923

Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions. *Philosophy & Technology*, *35*(1), 17. https://doi.org/10.1007/s13347-022-00511-9

Kurvers, R. H. J. M., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Argenziano, G., Zalaudek, I., & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, *113*(31). https://doi.org/10.1073/pnas.1601827113

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(22), 9020–9025. https://doi.org/10.1073/pnas.1008636108

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, *2*, 126–132. https://doi.org/10.1038/s41562-017-0273-4

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., Antonelli, G., Awadie, H., Bernhofer, S., Carballal, S., Dinis-Ribeiro, M., Fernández-Clotett, A., Esparrach, G. F., Gralnek, I., Higasa, Y., Hirabayashi, T., Hirai, T., Iwatate, M., Kawano, M., Mader, M., … Cherubini, A. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports*, *12*(1), 14952. https://doi.org/10.1038/s41598-022-18751-2

Steyvers, M., & Kumar, A. (2024). Three Challenges for AI-Assisted Decision-Making. *Perspectives on Psychological Science*, *19*(5), 722–734. https://doi.org/10.1177/17456916231181102

Surowiecki, J. (2004). The wisdom of crowds. In *Anchor*. https://doi.org/10.1016/S0140-6736(16)31130-8

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Turner, B. M., & Schley, D. R. (2016). The anchor integration model: A descriptive model of anchoring effects. *Cognitive Psychology*, *90*, 1–47. https://doi.org/10.1016/j.cogpsych.2016.07.003

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, *13*, 15737. https://doi.org/10.1038/s41598-023-42384-8

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/10.1006/OBHD.2000.2909

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852