UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

The Importance of Metacognitive Sensitivity in Human-AI Decision-Making

Permalink

https://escholarship.org/uc/item/0fg7g94k

Journal

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

Authors

Li, ZhaoBin

Steyvers, Mark

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

The Importance of Metacognitive Sensitivity in Human-AI Decision-Making

ZhaoBin Li (zhaobin.li@uci.edu)
Department of Cognitive Sciences
University of California, Irvine

Mark Steyvers (mark.steyvers@uci.edu)
Department of Cognitive Sciences
University of California, Irvine

Abstract

In human-AI decision-making, understanding the factors that maximize overall accuracy remains a critical challenge. This study highlights the role of metacognitive sensitivity—the agent's ability to assign confidence scores that reliably distinguish between correct and incorrect predictions. We propose a theoretical framework to evaluate the impact of accuracy and metacognitive sensitivity in hybrid decision-making contexts. Our analytical results establish conditions under which an agent with lower accuracy but higher metacognitive sensitivity can enhance overall decision accuracy when paired with another agent. Empirical analyses on a real-world image classification dataset confirm that stronger metacognitive sensitivity—whether in AI or human agents—can improve joint decision outcomes. These findings advocate for a more comprehensive approach to evaluating AI and human collaborators, emphasizing the joint optimization of accuracy and metacognitive sensitivity for enhanced decision-making.

Keywords: Human-AI collaboration; AI-Assisted Decision-Making; Model Accuracy; Metacognitive Sensitivity

Introduction

Artificial Intelligence (AI) is advancing rapidly in fields such as computer vision (Khan, Laghari, & Awan, 2018), speech recognition (Alharbi et al., 2021), and natural language processing (Raparthi et al., 2021). Humans are increasingly reliant on AI to make better decisions in a variety of fields, including AI-assisted driving, clinical diagnosis, and judicial advisory services. There is substantial potential for AI to assist humans, as the strengths and weaknesses of humans and AI are complementary. AI trained on large datasets can match or surpass human capabilities in specialized, narrow domains, whereas humans excel at learning quickly, generalizing from limited examples (Lake, Salakhutdinov, Gross, & Tenenbaum, 2011; S. W. Lee, O'Doherty, & Shimojo, 2015), adapting to novel situations (Franklin & Frank, 2020; Wu, Meder, & Schulz, 2024), and handling diverse tasks (Goertzel, 2014). Consequently, effective AI assistance is expected to become increasingly common in everyday life.

However, achieving human-AI complementarity—where human-AI teams outperform either humans or AI working independently—is challenging (Steyvers & Kumar, 2024). For instance, Vaccaro, Almaatouq, and Malone (2024) reported that nearly half of the reviewed studies found that human-AI collaboration led to worse performance. Similarly, a comprehensive review by Hemmer, Schemmer, Vössing, and Kühl (2021) of 53 studies on human-AI performance revealed that only 16 (30%) demonstrated complementarity. These findings highlight the need to better understand the factors driving successful human-AI collaboration.

One key factor determining the performance of a human-AI team is the accuracy of the human and AI working independently. If a human has a choice to work with one of two AI's, the human will be better off working with the more accurate AI. However, there are determinants other than accuracy that influence the performance of the human-AI team. In this paper, we focus on how confidence scores affect collaborative performance, with an emphasis on metacognitive sensitivity—the degree to which an agent consistently assigns higher confidence to correct decisions than to incorrect ones. Specifically, we explore how this aspect of metacognition interacts with accuracy to influence overall human-AI performance.

Confidence Scores in Human-AI Complementarity

Both humans and AI can provide confidence scores—estimates of how likely their decisions are to be correct. Humans routinely engage in metacognitive processes to evaluate their decision accuracy (Peters, 2022; Grimaldi, Lau, & Basso, 2015) and many AI models generate confidence estimates as part of their training objectives. For open-ended tasks, such as querying a large language model, confidence can be inferred through repeated evaluation or token likelihoods (Jiang, Araki, Ding, & Neubig, 2020; Kadavath et al., 2022). These confidence scores are essential for determining when human-AI teams achieve complementarity. For instance, even if an accuracy gap exists, complementarity can still occur if humans are confident in their decisions when the AI is not, and vice versa. Steyvers, Tejeda, Kerrigan, and Smyth (2022) demonstrated using Bayesian analysis that complementarity is achievable only when the accuracy difference between humans and AI falls within a specific range, dependent on the correlation between their confidence scores. These findings are supported by empirical evidence.

The reliability of confidence scores—how well they reflect actual correctness—also plays a pivotal role in determining human-AI complementarity. When confidence scores are reliable, decisions can often be optimized using a Bayes-optimal approach, weighting decisions according to confidence levels and accuracy (Dietterich, 2000). This principle

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).

is evident in human-human collaboration. For example, in a perceptual experiment by Bahrami et al. (2010), pairs of participants identified which of six visual stimuli had higher contrast. When they disagreed, they discussed and made joint decisions. Consistent with Steyvers et al. (2022)'s analysis, the results showed that pairs achieved complementarity only when their individual accuracies were sufficiently similar and their joint decisions aligned with a confidence-weighted model.

Confidence Discrimination versus Calibration

The reliability of confidence scores can be evaluated through two key aspects: calibration and discrimination, both of which are critical for effective human-AI collaboration. Calibration measures the alignment between confidence and accuracy, capturing tendencies toward overconfidence or underconfidence. In psychology, this is referred to as metacognitive bias (Fleming & Lau, 2014). Better calibration—whether in humans or AI—has been shown to improve joint performance (Ma et al., 2024; Benz & Rodriguez, 2023; Zhang, Liao, & Bellamy, 2020), though some exceptions exist (Vodrahalli, Gerstenberg, & Zou, 2022).

Discrimination, also known as metacognitive sensitivity, assesses an agent's ability to distinguish between correct and incorrect responses based on its confidence levels (Fleming & Lau, 2014). Recent work by Steyvers et al. (2025) demonstrated that large language models capable of accurately communicating their confidence through verbal explanations helped humans better assess the model's accuracy and distinguish between its correct and incorrect responses.

Although related, calibration and discrimination are distinct. Calibration reflects how well confidence scores match empirical accuracy and can often be corrected post hoc. Discrimination, by contrast, measures the ability of the model to rank the confidence of correct predictions higher than incorrect ones and is not easily adjusted without modifying the model itself. To date, no prior work has analytically formalized or empirically investigated the role of discrimination in the context of human-AI teaming.

The Need for a Theoretical Framework

Although prior research has examined how confidence and metacognitive processes influence human-AI decision-making, no study has explicitly isolated the impact of metacognitive sensitivity—whether human or AI—on joint human-AI decision-making or explored its trade-offs with accuracy. To address this, we propose a theoretical framework that specifies how both accuracy and metacognitive sensitivity jointly affect the performance of human-AI teams.

Using a signal-theoretic approach (Maniscalco & Lau, 2012), we model the effect of an agent's discrimination and accuracy parameters on the accuracy of a human-AI team and apply calculus to derive the team's combined accuracy under reasonable assumptions. Our analysis quantifies the impact of accuracy and metacognitive sensitivity in human-AI decision-making and reveals conditions where lower-accuracy agents with superior metacognitive sensitivity enhance overall team performance, a result we refer to as an *inversion scenario*. We further validate this framework using an empirical dataset where humans and AI classify images. The results confirm the existence of inversion scenarios and show the importance of balancing both accuracy and metacognitive sensitivity for optimal human-AI collaboration.

Problem Setup

We aim to study the impact of accuracy and metacognitive sensitivity when combining the predictions of two agents. These agents can both be humans similar to aggregation problems studied in the context of collective intelligence (Kameda, Toyokawa, & Tindale, 2022), both AIs similar to ensemble learning problems (Dietterich, 2000), or a combination of one human and one AI. The agents' predictions can be combined by one agent integrating their own prediction with the prediction from the other agent (e.g., AI-assisted human decision-making) or by a statistical rule.

Our theoretical model applies to all these settings, but we will frame the model in terms of a hybrid human-AI decision-making problem where a statistical rule is used to combine the predictions. In this setting, let M and H represent the AI model and human agent, respectively. The accuracy of agents M and H is denoted as a_m and a_h, respectively, representing the proportion of correct decisions. Additionally, both agents provide confidence scores associated with their decisions, denoted as c_m and c_h, where $0 \leq c_m, c_h \leq 1$.

We will consider the effect of an agent's metacognitive sensitivity, the ability to assign higher confidence scores to correct predictions than to incorrect ones, on combined human-AI accuracy. Our results show that when H's accuracy a_h is low, M's predictions dominate the final decision, making the combined accuracy primarily dependent on M's accuracy a_m. However, as a_h increases, M's metacognitive sensitivity significantly influences overall accuracy. Notably, we demonstrate that in certain cases, higher accuracy can be achieved when H is paired with an agent M_1 that has lower accuracy but higher metacognitive sensitivity, compared to an agent M_2 with higher accuracy but lower metacognitive sensitivity.

Generative Model

To analyze the interplay between accuracy and metacognitive sensitivity in AI models, we adopt a signal detection theory framework to model the generation of confidence scores (Galvin, Podd, Drga, & Whitmore, 2003). This framework allows us to control and simulate the accuracy and metacognitive sensitivity of an AI model in a systematic and interpretable manner. We then derive the analytic solution for combined accuracies of the human and AI under assumptions of normality.

AI Confidence Distribution

Let y_m be a Bernoulli random variable representing the correctness of a prediction, with latent variable θ_m representing

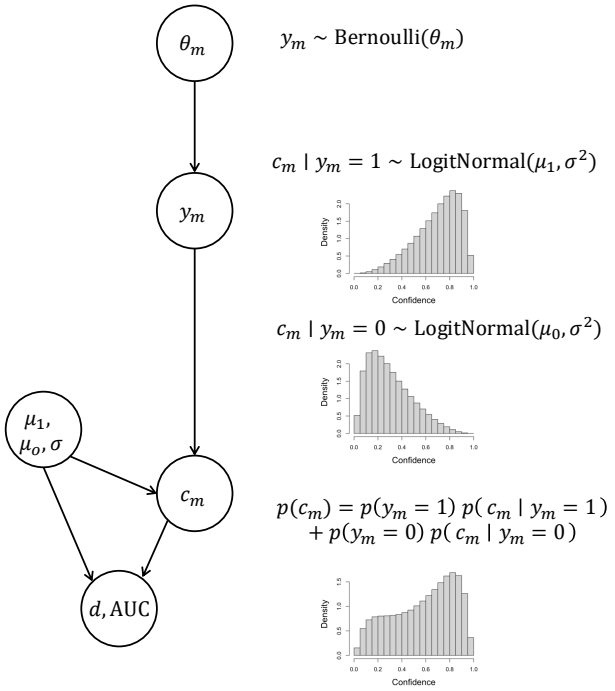


Figure 1: Illustration of the signal detection model to generate a model's confidence scores c_m by mixing confidence distributions for correct and incorrect decisions. The illustration is based on $\mu_0 = -1, \mu_1 = 1, \theta_m = 0.7$

the probability of a correct model prediction, i.e. $y_m = 1$:

$$y_m \sim \text{Bernoulli}(\theta_m) \tag{1}$$

The AI model generates a confidence score c_m conditioned on y_m. We assume the confidence distributions for correct and incorrect predictions follow logit-normal distributions:

$$c_m \mid y_m \sim \begin{cases} \text{LogitNormal}(\mu_0, \sigma^2) & \text{if } y_m = 0, \\ \text{LogitNormal}(\mu_1, \sigma^2) & \text{if } y_m = 1. \end{cases} \tag{2}$$

where $\mu_1 > \mu_0$ reflects the model's ability to assign higher confidence to correct predictions. The marginal confidence distribution is given by: $p(c_m) = p(y_m = 1)p(c_m \mid y_m = 1) + p(y_m = 0)p(c_m \mid y_m = 0)$ where $p(y_m = 1)$ is the model accuracy θ_m and $p(y_m = 0) = 1 - \theta_m$. Figure 1 illustrates the model confidence distributions for correct and incorrect predictions as well as the marginal confidence distribution.

Metacognitive Sensitivity

Metacognitive sensitivity reflects an agent's ability to assign higher confidence scores to correct predictions than to incorrect ones. We quantify this using Cohen's d, which measures the standardized mean difference between the confidence distributions for correct and incorrect predictions:

$$d = \frac{\mu_1 - \mu_0}{\sigma}. \tag{3}$$

As an alternative metric, we can also convert d to the area under the curve (AUC) for ease of interpretation. The AUC in our context is the probability that a randomly selected correct prediction will have a higher confidence score than an incorrect one. For normal distributions, this type of AUC can be directly derived from d (Kraemer, 2014):

$$\text{AUC} = \Phi\left(\frac{d}{\sqrt{2}}\right), \tag{4}$$

where Φ is the standard normal cumulative distribution function. The AUC metric for metacognitive sensitivity is **not** the same as the typical ROC AUC used in machine learning. The ROC AUC assesses the discrimination for positive and negative labels, whereas in the signal detection framework we apply here, the AUC assesses the discrimination for correct and incorrect decisions.

Combined Accuracy with Constant Human Confidence

When the human predictions are combined with the AI predictions, the goal is to maximize the combined decision accuracy. For a given input, a decision is made whether the final prediction is based on the human or AI prediction based on the AI's confidence score c_m and the human confidence score c_h. We first consider the case when human confidence is constant across all problems. For simplicity, we also assume perfect calibration for the human: when a human assigns a confidence c_h, their accuracy is also c_h on average.

Using Bayes' rule, the model calibration function that relates model accuracy to model confidence is

$$p(y_m = 1 \mid c_m) = \frac{p(y_m = 1)p(c_m \mid y_m = 1)}{p(c_m)}.$$

Since $\mu_1 > \mu_0$ for the model confidence distributions, the probability of a correct model prediction conditional on confidence, $p(y_m = 1 \mid c_m)$, increases monotonically with c_m.

The combined accuracy is determined by a switch-point $c_m = c^*$, where the human accuracy is aligned with AI accuracy:

$$c_h = p(y_m = 1 \mid c^*). \tag{5}$$

The probability of a correct final prediction for a particular level of human confidence c_h is given by:

$$p(y_{\text{combined}} = 1 \mid c_h) = \begin{cases} c_h & \text{if } c_m < c^*, \\ p(y_m = 1 \mid c_m) & \text{if } c_m \geq c^*. \end{cases} \tag{6}$$

In other words, if AI confidence is less than the switch-point confidence ($c_m < c^*$), the final prediction should be based on the human prediction because the human is more accurate. On the other hand, if $c_m \geq c^*$, the final prediction should be based on the AI because the AI is more accurate.

Thus, the combined accuracy can be expressed as follows:

$$p(y_{\text{combined}} = 1 \mid c_h) = \int_0^{c^*} c_h p(c_m)\, dc_m + \int_{c^*}^1 p(y_m = 1 \mid c_m) p(c_m)\, dc_m, \quad (7)$$

To find c^*, we invert f by setting:

$$c_h = f(c^*). \quad (8)$$

Solving this yields:

$$c^* = \text{sigmoid}\left(\frac{\sigma^2[\text{logit}(c_h) - \text{logit}(\theta_m)]}{\mu_1 - \mu_0} + \frac{\mu_1 + \mu_0}{2}\right), \quad (9)$$

Using properties of the logit-normal distribution and expanding c^*, the combined accuracy becomes:

$$p(y_{\text{combined}} = 1 \mid c_h) = c_h\left[\theta_m \Phi(r) + (1-\theta_m)\Phi(r+d)\right] + \theta_m\left[1 - \Phi(r)\right]. \quad (10)$$

where

$$r = \frac{\text{logit}(c_h) - \text{logit}(\theta_m)}{d} - \frac{d}{2} \quad (11)$$

Note that $p(y_{\text{combined}} = 1 \mid c_h)$ depends only on: the human confidence c_h, the model accuracy θ_m and the model's metacognitive sensitivity d.

Combined Accuracy with Variable Human Confidence

We now consider the more general case when human confidence varies across problems. Specifically, we will assume that human confidence follows a logit-normal distribution:

$$c_h \sim \text{LogitNormal}(\mu_h, \sigma_h). \quad (12)$$

As before, we assume that the human decision-maker is perfectly calibrated (i.e., the confidence c_h directly reflects the human's probability of correctness).

To derive the predicted combined accuracy, $p(y_{\text{combined}} = 1)$, we integrate the accuracy conditional on human confidence $p(y_{\text{combined}} = 1 \mid c_h)$ over c_h:

$$\begin{aligned} p(y_{\text{combined}} = 1) &= \int_0^1 p(c_h) p(y_{\text{combined}} = 1 \mid c_h) dc_h \\ &\approx \theta_m \,\text{BN}\left(\frac{a}{\sqrt{1+b^2}}, \frac{s}{\sqrt{1+t^2}}, \rho = \frac{bt}{\sqrt{1+b^2}\sqrt{1+t^2}}\right) \\ &+ (1-\theta_m)\,\text{BN}\left(\frac{a+d}{\sqrt{1+b^2}}, \frac{s}{\sqrt{1+t^2}}, \rho = \frac{bt}{\sqrt{1+b^2}\sqrt{1+t^2}}\right) \\ &+ \theta_m\left[1 - \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)\right] \end{aligned} \quad (13)$$

where

$$a = \frac{\mu_h - \text{logit}(\theta_m)}{d} - \frac{d}{2}, \quad b = \frac{\sigma_h}{d}$$
$$s = \lambda \mu_h, \quad t = \lambda \sigma_h \quad (14)$$

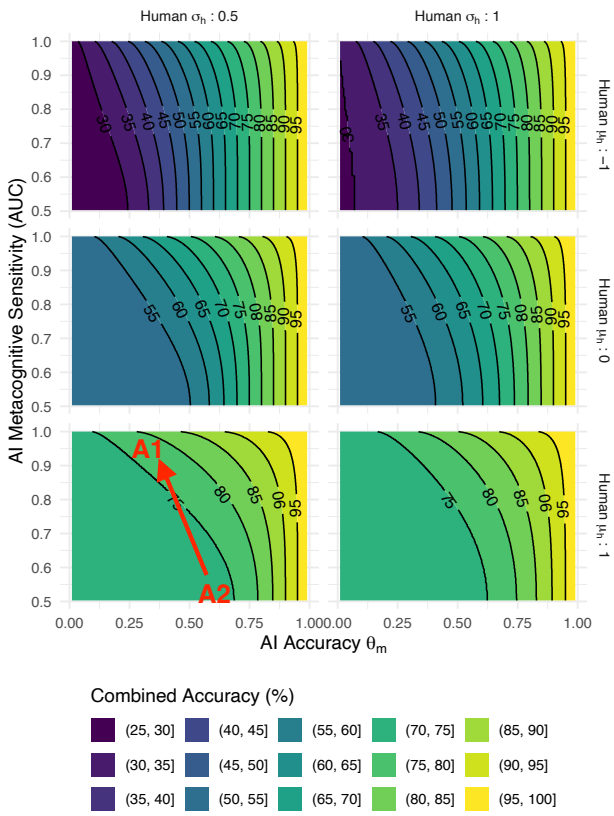


Figure 2: Predicted combined human and AI accuracy as a function of model accuracy (θ_m) and metacognitive sensitivity (assessed by AUC). Panels show results for different distributions of human confidence with mean μ_h and standard deviation σ_h. The red arrow highlights a scenario where AI model A1, having lower accuracy and higher metacognitive sensitivity than model A2, achieves higher accuracy when combined with a human.

The operator BN is the Bivariate-Normal CDF and $\lambda = \sqrt{\frac{\pi}{8}}$ is the constant for probit approximation. We tested various parameters and the approximation is highly accurate, within 1% difference of numerical integration.

The overall combined accuracy depends only on the mean μ_h and variance σ_h of the human confidence distribution, the model accuracy θ_m, and the model's metacognitive sensitivity d.

In Figure 2, we present a graph depicting combined accuracy as a function of these four variables. When human accuracy is low ($\mu_h = -1$), the combined accuracy primarily depends on AI accuracy (θ_m). However, as human accuracy increases ($\mu_h = 1$), AI metacognitive sensitivity plays a crucial role in determining overall accuracy. The figure also highlights an example where AI model A1, despite having lower accuracy but higher metacognitive sensitivity than model A2, achieves greater combined accuracy when paired with a human.

Experiments

To evaluate the impact of metacognitive sensitivity on the accuracy of combined decision-making in a real-world setting, we apply our theoretical framework to an image classification dataset comprising both human and machine predictions. We verify that the theoretical predictions align with observed outcomes through experiments that investigate the influence of human and AI metacognitive sensitivity.

Image Classification Dataset

We use the ImageNet-16H image classification dataset from Steyvers et al. (2022). This dataset comprises 1,200 images across 16 categories selected from the ImageNet challenge (Russakovsky et al., 2015). To increase task difficulty, four levels of phase noise distortion were applied to each image, yielding 4,800 unique images.

A total of 145 participants each labeled a subset of 200 images and provided confidence ratings (low, medium, high) for each decision, resulting in 28,997 human classifications with associated confidence scores. Additionally, the dataset includes predictions from five machine classifiers—AlexNet, DenseNet161, GoogleNet, ResNet152, and VGG19—pretrained on ImageNet. These models were fine-tuned for four different durations, producing 20 model variants in total.

Methods

We paired each of the 20 model variants with each of the 200 human participants, yielding 4,000 human-AI pairs. Each pair classified a subset of 200 images. Within each pair, we generated calibration curves for both the human and AI models, which are essential for computing both theoretical and empirical combined accuracies.

For human participants, we constructed calibration curves by computing mean accuracy at three confidence levels. For AI models, we extracted top-1 confidence scores for each prediction, binned them into ten equal-width intervals between 0 and 1, and computed the mean accuracy within each bin to form the AI calibration curve.

Experiment 1: Effect of AI Metacognitive Sensitivity Since the theoretical framework assumes that agent H is perfectly calibrated, we empirically calibrated human confidence scores by replacing them with the mean accuracies of the corresponding bins in the human calibration curve. Additionally, we assumed that agent H's confidence scores follow a logit-normal distribution with mean μ_h and standard deviation σ_h. We estimated them via maximum likelihood by logit-transforming the human calibrated scores and computing their mean and standard deviation.

Agent M is characterized by latent accuracy θ_m and metacognitive sensitivity d. We estimated them using maximum likelihood on the AI confidence scores. After determining all four parameters—μ_h, σ_h, θ_m, and d—we computed the predicted combined accuracy using equation 13. We also obtained the empirical combined accuracy using the following procedure:

1. Bootstrap a sample of 2,000 images for each human-AI pair from their set of 200 images.
2. Retrieve the AI's top-1 confidence scores and the human's calibrated accuracy for each sampled image.
3. Use the AI's calibration curve to estimate the mean accuracy corresponding to its confidence score.
4. Choose the prediction with the higher mean accuracy: follow the human's prediction if their calibrated accuracy exceeds the AI's; otherwise, follow the AI's prediction.
5. Compare the selected prediction to the ground-truth label to assess correctness.

The proportion of correct predictions across the 2,000 images was recorded as the empirical combined accuracy.

Experiment 2: Effect of Human Metacognitive Sensitivity In the second experiment, we followed the same procedure but reversed the roles of the human and AI agents. Here, the AI was designated as agent H, and the human as agent M, allowing us to examine the impact of human metacognitive sensitivity on combined decision accuracy.

Specifically, we empirically calibrated the AI's confidence scores and estimated its logit-normal distribution parameters. We then computed the human's latent accuracy and metacognitive sensitivity, applying the same method to obtain the combined decision accuracy.

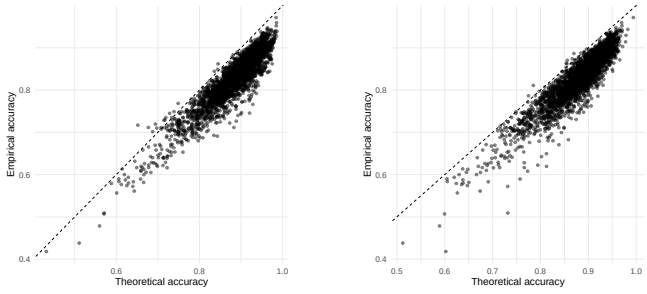


Figure 3: Theoretical vs. empirical accuracy of human-AI combinations for Experiment 1 (left) and Experiment 2 (right). The dashed line represents the points where the two accuracies are equal.

Results

Theoretical Accuracy Aligns with Empirical Accuracy Figure 3 illustrates that for both experiments, the combined accuracy predicted by our theoretical framework (equation 13) closely aligns with the accuracy obtained from empirical simulations. This yields a Pearson correlation of $R = 0.92$ with $p < 0.001$ for both experiments.

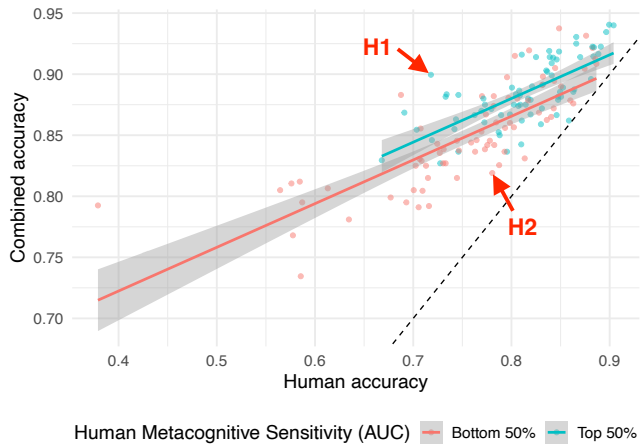


Figure 4: Empirical combined human-AI accuracy versus human accuracy for 200 participants paired with the VGG19 model fine-tuned for one epoch. Participants are divided into two groups based on a median split of their metacognitive sensitivity, measured by AUC. The trendline is fitted using linear regression, with confidence bands representing standard errors. The dashed line denotes points where human and combined accuracies are equal. Points H1 and H2 highlight potential inversions, where a participant with lower accuracy but higher metacognitive sensitivity achieves a higher combined accuracy (than a human with higher accuracy but lower metacognitive sensitivity).

Higher Metacognitive Sensitivity Increases Combined Accuracy Visualizing the results of the first experiment is challenging, as illustrating the impact of AI's metacognitive sensitivity would require a separate graph for each of the 200 participants. To analyze this effect, we employed a multilevel linear regression model predicting empirical combined accuracy. This model included human accuracy, AI accuracy, and AI metacognitive sensitivity as covariates, with human participants treated as random intercepts.

The findings indicate that AI metacognitive sensitivity positively influences combined accuracy, with significant effects observed for both empirical ($\beta = 0.07,\ SE = 0.013,\ t = 5.73,\ p < 0.001$) and theoretical ($\beta = 0.15,\ SE = 0.009,\ t = 16.49,\ p < 0.001$) combined accuracy.

In Experiment 2, Figure 4 plots empirical combined accuracy against human accuracy for the VGG19 model fine-tuned for one epoch and paired with 200 participants. Participants were divided into two groups based on a median split of their metacognitive sensitivity (AUC). The trendline for the top 50% lies above that of the bottom 50%, demonstrating that greater metacognitive sensitivity enhances combined accuracy. Notably, we observed instances where a participant with lower accuracy but higher metacognitive sensitivity achieved higher combined accuracy in collaboration with the AI.

To validate this effect, we applied a multilevel linear regression model predicting empirical combined accuracy across all 20 AI models. This model included human accuracy, AI accuracy, and human metacognitive sensitivity as covariates, with model variants treated as random intercepts. The results confirm that higher human metacognitive sensitivity significantly improves combined accuracy, with an effect size comparable to that observed in the first experiment ($\beta = 0.06,\ SE = 0.006,\ t = 10.59,\ p < 0.001$).

Discussion and Limitations

This study establishes a foundation for understanding the role of accuracy and metacognitive sensitivity in hybrid human-AI decision-making. Our findings demonstrate that both human and AI metacognitive sensitivity are crucial factors in shaping decision outcomes. The proposed analytical framework provides precise predictions on the interplay between metacognitive sensitivity and accuracy. Specifically, it predicts that when human accuracy is near chance, AI accuracy solely determines joint performance. However, when humans exhibit some level of expertise, optimizing both accuracy and metacognitive sensitivity becomes essential for enhancing overall performance.

Several avenues remain open for further exploration to refine both the theoretical framework and its practical applications. A key direction for future research is improving the modeling of human decision-makers. In this study, we assumed perfect calibration—where human confidence accurately reflects the probability of correctness. Future work could incorporate more realistic human behaviors, such as biases, overconfidence, and underconfidence. Computational modeling approaches from cognitive science, such as those based on signal detection theory, could provide a richer representation of human confidence and decision-making processes (Maniscalco & Lau, 2012).

Another essential direction for future research is conducting behavioral experiments to assess whether humans can effectively integrate an AI model's metacognitive sensitivity into their decision-making. While prior research (D. Lee, Pruitt, Zhou, Du, & Odegaard, 2024) suggests that humans are attuned to metacognitive sensitivity when receiving advice from other agents, empirical validation is needed to confirm this effect in human-AI interactions.

Conclusion

In this paper, we introduced a theoretical framework to evaluate the impact of accuracy and metacognitive sensitivity in hybrid human-AI decision-making tasks. Our analysis highlights that agents—whether human or AI—with lower accuracy but superior metacognitive sensitivity can significantly enhance overall performance in hybrid systems. These findings challenge the conventional emphasis on accuracy alone, underscoring the critical role of metacognitive sensitivity in optimizing human-AI collaboration. This work advances our understanding of human-AI complementarity, offering key insights into the metrics that drive effective cooperation.

References

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., ... Almojil, M. (2021). Automatic speech recognition: Systematic literature review. *IEEE Access*, *9*, 131858–131876.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010, August). Optimally interacting minds. *Science*, *329*(5995), 1081–1085.

Benz, N. L. C., & Rodriguez, M. G. (2023, May). Human-aligned calibration for AI-assisted decision making. *arXiv [cs.LG]*, 14609–14636.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1–15). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fleming, S. M., & Lau, H. C. (2014, July). How to measure metacognition. *Front. Hum. Neurosci.*, *8*, 443.

Franklin, N. T., & Frank, M. J. (2020, April). Generalizing to generalize: Humans flexibly switch between compositional and conjunctive structures during reinforcement learning. *PLoS Comput. Biol.*, *16*(4), e1007720.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003, December). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychon. Bull. Rev.*, *10*(4), 843–876.

Goertzel, B. (2014, December). Artificial general intelligence: Concept, state of the art, and future prospects. *J. Artif. Gen. Intell.*, *5*(1), 1–48.

Grimaldi, P., Lau, H., & Basso, M. A. (2015, August). There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. *Neurosci. Biobehav. Rev.*, *55*, 88–97.

Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-AI complementarity in hybrid intelligence systems: A structured literature review. *PACIS*, 78.

Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2020, December). How can we know when language models know? on the calibration of language models for question answering. *arXiv [cs.CL]*.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., ... Kaplan, J. (2022, July). Language models (mostly) know what they know. *arXiv [cs.CL]*.

Kameda, T., Toyokawa, W., & Tindale, R. S. (2022). Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology*, *1*(6), 345–357.

Khan, A., Laghari, A., & Awan, S. (2018, July). Machine learning in computer vision: A review. *ICST Trans. Scalable Inf. Syst.*, *8*(32), 169418.

Kraemer, H. C. (2014, May). Effect size. In *The encyclopedia of clinical psychology* (pp. 1–3). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Lake, B., Salakhutdinov, R., Gross, J., & Tenenbaum, J. (2011). One shot learning of simple visual concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33).

Lee, D., Pruitt, J., Zhou, T., Du, J., & Odegaard, B. (2024, November). Metacognitive sensitivity: the key to calibrating trust and optimal decision-making with AI. *PsyArXiv*.

Lee, S. W., O'Doherty, J. P., & Shimojo, S. (2015, April). Neural computations mediating one-shot learning in the human brain. *PLoS Biol.*, *13*(4), e1002137.

Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M., & Ma, X. (2024, May). "are you really sure?" understanding the effects of human self-confidence calibration in AI-assisted decision making. In *Proceedings of the chi conference on human factors in computing systems* (Vol. 63, pp. 1–20). New York, NY, USA: ACM.

Maniscalco, B., & Lau, H. (2012, March). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.*, *21*(1), 422–430.

Peters, M. A. K. (2022, March). Confidence in decision-making. In *Oxford research encyclopedia of neuroscience.* Oxford University Press.

Raparthi, M., Dodda, S. B., Reddy, S. R. B., Thunki, P., Maruthi, S., & Ravichandran, P. (2021, August). Advancements in natural language processing - a comprehensive review of AI techniques. *Journal of Bioinformatics and Artificial Intelligence*, *1*(1), 1–10.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015, December). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, *115*(3), 211–252.

Steyvers, M., & Kumar, A. (2024). Three challenges for ai-assisted decision-making. *Perspectives on Psychological Science*, *19*(5), 722–734.

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022, March). Bayesian modeling of human-AI complementarity. *Proc. Natl. Acad. Sci. U. S. A.*, *119*(11), e2111547119.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., ... Smyth, P. (2025, January). What large language models know and what people think they know. *Nat. Mach. Intell.*, 1–11.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024, December). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nat. Hum. Behav.*, *8*(12), 2293–2303.

Vodrahalli, K., Gerstenberg, T., & Zou, J. (2022, February). Uncalibrated models can improve human-AI collaboration. *arXiv [cs. AI]*.

Wu, C. M., Meder, B., & Schulz, E. (2024, October). Unifying principles of generalization: Past, present, and future. *Annu. Rev. Psychol.*.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency.* New York, NY, USA: ACM.