Designing Behavior-Aware AI to Improve the Human-AI Team Performance in AI-Assisted Decision Making

Syed Hasan Amin Mahmood, **Zhuoran Lu** and **Ming Yin**
Purdue University
{hasanamin, lu800, mingyin}@purdue.edu

Abstract

With the rapid development of decision aids that are driven by AI models, the practice of AI-assisted decision making has become increasingly prevalent. To improve the human-AI team performance in decision making, earlier studies mostly focus on enhancing humans' capability in better utilizing a *given* AI-driven decision aid. In this paper, we tackle this challenge through a complementary approach—we aim to train "*behavior-aware AI*" by adjusting the AI model underlying the decision aid to account for humans' behavior in adopting AI advice. In particular, as humans are observed to accept AI advice more when their confidence in their own judgement is low, we propose to train AI models with a *human-confidence-based instance weighting strategy*, instead of solving the standard empirical risk minimization problem. Under an assumed, threshold-based model characterizing when humans will adopt the AI advice, we first derive the optimal instance weighting strategy for training AI models. We then validate the efficacy and robustness of our proposed method in improving the human-AI joint decision making performance through systematic experimentation on synthetic datasets. Finally, via randomized experiments with real human subjects along with their actual behavior in adopting the AI advice, we demonstrate that our method can significantly improve the decision making performance of the human-AI team in practice.

1 Introduction

Artificial Intelligence (AI) technologies have been widely used to support decision making in many domains, leading to the paradigm of "*AI-assisted decision making*" where AI provides decision recommendations while humans integrate the AI advice with their own knowledge to arrive at the final decisions. However, the potential of such human-AI collaboration often falls short in practice, and "human-AI complementarity"—that is, the human-AI team outperforms either human or AI alone in decision making—is rarely achieved. This necessitates the exploration of novel approaches to improve the human-AI team performance in AI-assisted decision making.

Prior efforts have primarily focused on improving the human-AI collaborative decision making performance by "augmenting" humans, with particular emphasis on promoting humans' appropriate reliance on AI [Bansal *et al.*, 2021b; Buçinca *et al.*, 2021]. In these endeavors, the AI model underlying the decision aid is often assumed to be *given* and is designed to maximize its independent accuracy. This indicates a largely under-explored direction for improving AI-assisted decision making—can we quantitatively characterize how human decision makers would factor AI recommendations into their decisions, and utilize this to directly design AI models that optimize for the human-AI team accuracy in joint decision making? In other words, can we develop AI models that are aware of human behavior and complement humans by design? Compared to existing methods focusing on augmenting humans, designing behavior-aware AI can potentially be a more powerful and scalable approach to improve the human-AI team performance in AI-assisted decision making, as AI models are often more "tunable" than their human teammates.

In this paper, we take a first step towards designing behavior-aware AI for AI-assisted decision making, building upon recent empirical observations of the real-world human behavior in these scenarios. It is found that human decision makers' confidence in their own judgment (i.e., their "self-confidence") on a decision making case significantly influences their likelihood of adopting the AI's recommendation, with lower self-confidence associated with higher chance of adopting AI recommendation [Chong *et al.*, 2022]. We thus create a threshold-based team decision making model to characterize such human behavior, and propose to train the AI models to account for this behavior by following a *human-confidence-based instance weighting* method rather than solving the standard empirical risk minimization problem. This method effectively shifts the AI model's attention to those cases where decision makers have low self-confidence and have higher "needs" for accurate AI recommendations.

To validate the effectiveness of the proposed approach, we conduct comprehensive experiments that encompass both simulation-based evaluations and real-world human-subject studies. Our real-world human-subject experiment results show that when human decision makers are assisted by the AI model trained using our proposed method, the human-AI team accuracy in decision making is increased significantly compared to when they are assisted by a standard AI model that is trained to maximize its independent accuracy; this performance increase primarily comes from task instances on which humans are less confident about their own judgments. Moreover, our simulation results suggest that the human-AI team performance gain brought up by the human-confidence-aware AI is the largest over the standard AI when the expertise of human decision makers exhibits significant overlaps with the standard AI model, and is the largest over the human-accuracy-aware AI when humans' confidence is highly uncalibrated.

2 Related Work

As AI-driven decision aids are increasingly used to support decision making, research on how to improve human-AI collaboration in AI-assisted decision making has surged recently. A key objective of this research is to explore novel approaches to improve the human-AI team accuracy in joint decision making. To this end, researchers have been mostly focusing on helping humans better utilize the given AI, including assisting them to form better mental models of AI [Bansal *et al.*, 2019; Mozannar *et al.*, 2022], providing additional model information to enable calibrated trust in AI [Zhang *et al.*, 2020; Yang *et al.*, 2020], and forcing them to engage with AI's advice cognitively [Buçinca *et al.*, 2021].

In contrast, very limited studies take the approach of redesigning the AI models to account for humans' behavior in adopting AI recommendations and directly optimizing for the human-AI team accuracy in AI-assisted decision making. A notable exception is Bansal *et al.* [2021a], although the study assumes humans to be rational and uniformly accurate across all decision cases. In the real world, however, humans often exhibit irrational behavior. Indeed, empirical studies have shown that humans' adoption of AI recommendations is often influenced by their cognitive biases [Lu and Yin, 2021; Rastogi *et al.*, 2022; Bertrand *et al.*, 2022]. Thus, optimizing human-AI team decision making in practice requires us to model the empirically-grounded, realistic human behavior in AI-assisted decision making [Li *et al.*, 2023; Li *et al.*, 2024], and incorporate such behavior into the development of AI models. In this study, we focus on designing behavior-aware AI to account for one particular aspect of the real-world human behavior in adopting AI recommendation: humans' confidence in their own judgment is indicative of their inclination to accept AI recommendation [Chong *et al.*, 2022; Wang *et al.*, 2022].

The idea of taking humans' real-world behavior into account in designing AI has been explored in other human-AI collaboration settings. For example, there is a line of literature on *learning to defer* that highlights the *division of labor* between humans and AI [Madras *et al.*, 2018; Wilder *et al.*, 2020; Bondi *et al.*, 2022; Dvijotham *et al.*, 2023]. In these studies, the AI model is designed to decide whether to make the decision itself or ask for a human to make the decision, taking humans' and AI's capabilities into account. In *human-robot co-planning* settings, where human and AI agents each make a sequence of decisions while coordinating with each other to complete a joint goal, researchers have demonstrated the advantage of training the AI agent using a human model rather than through self-play [Carroll *et al.*, 2019; Kwon *et al.*, 2020]. Our work differs from these prior studies as we focus on training behavior-aware AI in the *AI-assisted decision making* setting, where AI only provides recommendations and humans are always the final decision maker.

3 Problem Setup

In an AI-assisted decision making setting, given a decision making case characterized by features $\mathbf{x} \in \mathcal{X}$, an AI model first provides a decision recommendation $y_m = m(\mathbf{x}; \theta_m)$ to a human decision maker (DM)—who has their own independent judgment $y_h = h(\mathbf{x}; \theta_h)$ on this case—and then the human DM needs to make the final team decision $d \in \mathcal{Y}$. Unlike some other human-AI collaboration paradigms, humans always retain the role of final decision maker here, which is ubiquitous especially in contexts involving high-stake decisions. Without loss of generality, we focus on multiclass classification tasks in this study (i.e., $\mathcal{Y} = \{1, 2, \ldots, K\}$).

The AI model is typically learned from a training dataset which comprises N data instances, i.e., $\mathcal{D} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_N\}$ where $\mathcal{I}_i = (\mathbf{x}_i, y_i)$. A common practice adopted to train the AI model is to learn the model parameters θ_m to minimize the empirical risks over $\mathcal{D}$:

$$\theta_m = \arg\min_{\theta'_m} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \ell\left(m(\mathbf{x}_i;\, \theta'_m), y_i\right) \tag{1}$$

where $\ell(\cdot)$ is a loss function of interest (e.g., 0-1 loss). However, this training process effectively optimizes for the AI model's *independent* performance rather than the performance of the *human-AI team*. In other words, this optimization process neglects the human DM's contribution to the decision making process. Assuming that the human DM's final team decision $d = f(\mathbf{x}, y_m = m(\mathbf{x}; \theta_m), y_h = h(\mathbf{x}; \theta_h))$, i.e., d is influenced by the decision making case $\mathbf{x}$, the AI model's decision recommendation y_m, and the human DM's own independent judgment y_h, training an AI model that optimizes for the human-AI team performance requires us to solve a new empirical risk minimization problem focusing on team loss:

$$\arg\min_{\theta'_m} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \ell\left(f(\mathbf{x}_i, m(\mathbf{x}_i; \theta'_m), h(\mathbf{x}_i; \theta_h)), y_i\right) \tag{2}$$

It is therefore critical to understand the form of the human-AI team decision making model $f(\cdot)$ to accurately reflect how human DMs factor the AI model's decision recommendations into their final decisions. Interestingly, recent empirical studies suggest that when assisted by an AI model in decision making, human DMs are more inclined to accept the AI recommendation when they have low "self-confidence", that is, their confidence in their own independent judgment is low [Chong *et al.*, 2022; Wang and Du, 2018; Schemmer *et al.*, 2023; Wang *et al.*, 2022]. Thus, when a human confidence oracle $\mathcal{C}$ that provides us with human self-confidence on each decision making instance (i.e., $\mathcal{C} : \mathcal{H}(\mathcal{X}) \mapsto [0, 1]$) is available, this empirical insight can be reflected by a threshold-based team decision making model:

$$f(\mathbf{x}_i, m(\mathbf{x}_i; \theta_m), h(\mathbf{x}_i; \theta_h)) = \begin{cases} h(\mathbf{x}_i; \theta_h) & \text{if } \mathcal{C}_i > \tau \\ m(\mathbf{x}_i; \theta_m) & \text{otherwise} \end{cases} \tag{3}$$

where $\mathcal{C}_i := \mathcal{C}(h(\mathbf{x}_i; \theta_h))$ is the human DM's self-confidence on instance i, and τ is the self-confidence threshold for the human DM to adopt or ignore the AI recommendation. Since humans will rely on the AI recommendation if their self-confidence is below τ, a higher value of τ is associated with a higher frequency for humans to rely on the AI recommendation. Note that the human DM's self-confidence does *not* necessarily reflect the accuracy of their own judgment. In fact, humans can often overestimate (e.g., "Dunning-Kruger effect" [Dunning, 2011]) or underestimate (e.g., "impostor syndrome" [Langford and Clance, 1993]) their abilities.

In this paper, as an initial step to better factor the human DM's behavior in AI-assisted decision making into the training of the AI model, we explore how the AI model should be trained to optimize for the human-AI team performance when the team uses the threshold-based model (i.e., Equation 3) to make the joint decisions.

4 Human-Confidence-Based Instance Weighting

When humans use the threshold-based model to determine their final decisions in AI-assisted decision making, they will only adopt the AI recommendation when their self-confidence is sufficiently low (i.e., below τ). Intuitively, this implies that an AI model needs to be as accurate as possible on those decision making instances where humans are less confident about their own judgments and thus "need" the AI advice more. To operationalize this idea, we propose to train a behavior-aware, complementary AI model $y_c = m_c(\mathbf{x}; \theta_c)$ that minimizes the *weighted* empirical risks over the entire training dataset, where the weight of each instance (w_i) is a function of the human DM's self-confidence on it ($\mathcal{C}_i$):

$$\theta_c = \arg\min_{\theta'_c} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} w_i \cdot \ell(m_c(\mathbf{x}_i; \theta'_c), y_i) \qquad (4)$$

Note that the standard AI model $y_m = m(\mathbf{x}; \theta_m)$ can be seen as weighing all instances equally (i.e., $w_i = 1 \; \forall \mathcal{I}_i \in \mathcal{D}$). In general, without additional information about the value of the self-confidence threshold, we have the following proposition:

Proposition 1. *If the human DM is less confident about $\mathcal{I}_i$ than $\mathcal{I}_j$, then $\mathcal{I}_i$ should be weighted at least as high as $\mathcal{I}_j$, i.e., $w_i \geq w_j$ if $\mathcal{C}_i < \mathcal{C}_j$.*

Proof. See supplemental materials (SM) for the proof.

Following this proposition, we may propose a few heuristic methods for setting the weight for each training data instance, e.g., $w_i = 1 - \mathcal{C}_i$ or $w_i = \frac{1}{\mathcal{C}_i}$. Below, we discuss how to derive the optimal weight of each training data instance in two different scenarios with different kinds of information about the self-confidence threshold τ.

Scenario 1: Optimization for Known Self-Confidence Threshold. First, we consider the simplest scenario where the human DM has a fixed self-confidence threshold τ to determine their reliance on the AI recommendation, and its value is known to the AI model developer. Let $\mathcal{D}_h := \{\mathcal{I}_i \mid \mathcal{C}_i > \tau\}$ and $\mathcal{D}_l := \mathcal{D} \setminus \mathcal{D}_h$ be the sets of instances where human DM has high and low self-confidence, respectively. Using the threshold-based team decision making model (Equation 3), the complementary AI should focus only, and equally, on instances in the low confidence set $\mathcal{D}_l$.

Proposition 2. *When the human DM uses a fixed and known self-confidence threshold τ to determine the human-AI team decision, the team loss is minimized when $w_i = \mathbb{1}[\mathcal{C}_i \leq \tau]$.*

Proof. See SM for the proof.

Scenario 2: Optimization for Expected Self-Confidence Thresholds. In practice, humans' self-confidence threshold τ may not only be unknown to the AI model developer, but may also vary across different DMs and across time. To reflect this, we consider a scenario where the human DM draws τ from a known distribution (i.e., $\tau \sim f_T(\tau)$) and then applies the threshold-based model to determine their final decision. In this case, the complementary AI model needs to be trained to minimize the expected team loss over all possible τ.

Proposition 3. *When the human DM draws a self-confidence threshold from a known distribution to determine the human-AI team decision, i.e., $\tau \sim f_T(\tau)$, the expected team loss is minimized when $w_i = 1 - F_T(\mathcal{C}_i)$, where $F_T(\cdot)$ is the cumulative distribution function (CDF) for τ.*

Proof. Given the threshold-based team decision making model, we decompose the expected team loss ($\mathbb{E}[\mathcal{L}_{team}]$) as follows (we use $h(\mathbf{x})$ and $m_c(\mathbf{x})$ to refer to $h(\mathbf{x}; \theta_h)$ and $m_c(\mathbf{x}; \theta_c)$, respectively, for convenience and readability):

$$\begin{aligned}
&\int_{\tau=0}^{1} f_T(\tau) \cdot \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \ell(f(\mathbf{x}_i, m_c(\mathbf{x}_i), h(\mathbf{x}_i)), y_i) \; d\tau \\
=&\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \int_{0}^{1} f_T(\tau) \cdot \ell(f(\mathbf{x}_i, m_c(\mathbf{x}_i), h(\mathbf{x}_i)), y_i) \; d\tau \\
=&\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \Bigg(\int_{0}^{\mathcal{C}_i} f_T(\tau) \cdot \ell(h(\mathbf{x}_i), y_i) \; d\tau \\
&+ \int_{\mathcal{C}_i}^{1} f_T(\tau) \cdot \ell(m_c(\mathbf{x}_i), y_i) \; d\tau \Bigg) \\
=&\underbrace{\frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} F_T(\mathcal{C}_i) \cdot \ell(h(\mathbf{x}_i), y_i)}_{\text{uncontrollable human loss}} \\
&+ \frac{1}{|\mathcal{D}|} \sum_{(x_i, y_i) \in \mathcal{D}} (1 - F_T(\mathcal{C}_i)) \cdot \ell(m_c(\mathbf{x}_i), y_i) \\
\propto &\sum_{(x_i, y_i) \in \mathcal{D}} (1 - F_T(\mathcal{C}_i)) \cdot \ell(m_c(\mathbf{x}_i), y_i)
\end{aligned}$$

Thus, minimizing $\mathbb{E}[\mathcal{L}_{team}]$ is equivalent to minimizing $\sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} (1 - F_T(\mathcal{C}_i)) \cdot \ell(m_c(\mathbf{x}_i; \theta_c), y_i)$, which implies $w_i = 1 - F_T(\mathcal{C}_i)$.

Remark. Following Proposition 3, we can see that when the human DM draws τ from a uniform distribution, i.e., $\tau \sim U[0, 1]$, the heuristic method of setting the weight of each training instance $w_i = 1 - \mathcal{C}_i$ is in fact the optimal.

5 Simulation Study

In this section, we conduct simulations to examine whether, and how, the joint decision making performance of human-AI teams improves when assisted by an AI model trained following our proposed human-confidence-based instance weighting (CBIW) method instead of the standard approach. This simulation is conducted on a synthetically generated college admission decision making dataset. Evaluation on this synthetic dataset is useful because it allows us to systematically control characteristics of the human DM's behavior, so that we can examine the robustness of the proposed method in improving the human-AI joint decision making performance.

5.1 Synthetic Dataset Generation

Generating the Ground Truth. We consider a decision making task where DMs need to determine whether to admit an applicant to college, given two features of the applicant—their Grade Point Average (i.e., "GPA") and standardized test scores (i.e., "SCORE"). Inspired by Haider *et al.* [2022], we assume that applicants belong to either the privileged group or the underprivileged group, and admission outcomes for applicants of different groups are primarily decided by distinct sets of features. More specifically, we generate a set of $100,000$ (x_{GPA}, x_{Score}, y) instances, where the values of x_{GPA} and x_{Score} are uniformly randomly sampled between 0 and 1 without loss of generality. The applicant is further assigned to the privileged group with probability r, and we use $r = 0.75$ in this simulation study. Finally, we follow the two steps below to generate the ground truth label y for each applicant: (1) we first set y for each applicant to reflect that SCORE is more predictive of the admission outcome for privileged applicants, while GPA is more predictive for underprivileged applicants; (2) to account for a degree of randomness in the admission process, we then flip the label y currently set for each applicant with a small probability, and this probability is either proportional (when flipping from "reject" to "admit") or inversely proportional (when flipping from "admit" to "reject") to the value of $x_{GPA} + x_{Score}$. Details of the generation process of the College Admission dataset can be found in SM.

Generating Human DMs' Behavior. Beyond generating the ground truth label for all instances in the synthetic dataset, we also need to simulate how humans will make their decisions on these instances. To reflect that humans have varying levels of accuracy on different subsets of decision making tasks, on a decision making instance that belongs to group g (i.e., privileged or underprivileged), we randomly generate a human DM's independent judgment y_h such that it is correct with a probability equal to their accuracy on this group (i.e., acc_g). Further, the human DM's confidence on this instance is randomly sampled from a range between $\hat{acc}_g - \Delta_u$ and $\hat{acc}_g + \Delta_o$ to reflect the DM's varying degree of confidence calibration ($\hat{acc}_g = acc_g$ is *not* guaranteed). Finally, the DM's self-confidence threshold τ on this instance is randomly sampled from a distribution $f_T(\tau)$, and we experiment with different $f_T(\tau)$ in our simulation.

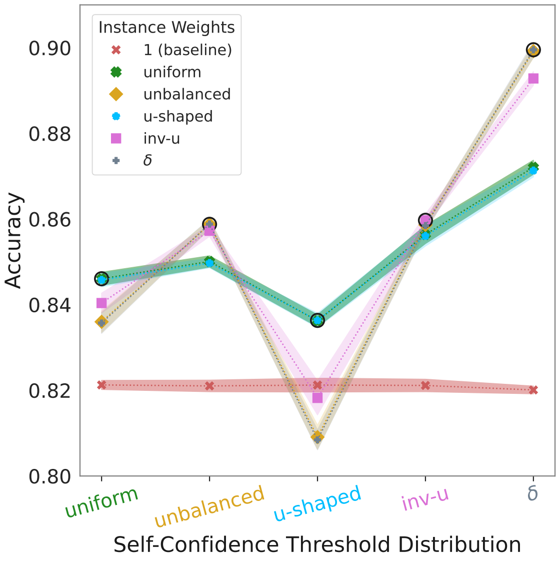

Figure 1: The human-AI team decision making accuracy (y-axis) when human DMs' self-confidence thresholds are drawn from different distributions (x-axis), and DMs collaborate with AI models trained using different human-confidence-based instance weighting strategies. Error shades represent the standard errors of the mean. Black circles are used to highlight the largest y-values for every self-confidence distribution on the x-axis.

5.2 Evaluating Varied Threshold Distributions

We first evaluate the effectiveness of the proposed CBIW-training method in improving the human-AI team performance when human DMs have different self-confidence threshold distributions in determining the team decisions (i.e., $f_T(\tau)$). Proposition 3 suggests that given a specific self-confidence threshold distribution $f_T(\tau)$, the optimal weighting function to be used to train the complementary AI model is $w_i = 1 - F_T(\mathcal{C}_i)$. However, knowing or being able to reliably estimate $f_T(\tau)$ can be unrealistic in practice. Thus, as a secondary goal of this evaluation, we aim to explore how critical using the exact optimal weighting function is to obtaining human-AI team performance gains through our CBIW-training method.

Evaluation Setup. We assume DMs' independent judgments are more accurate for applicants from the privileged group. Thus, we set $acc_{priv} = 0.9$ and $acc_{unpriv} = 0.6$. We further set $\hat{acc}_g = acc_g, \Delta_u = \Delta_o = 0.1$ (i.e., DMs' confidence is well calibrated). We consider five types of self-confidence threshold distributions $f_T(\tau)$: (1) UNIFORM: $\tau \sim \beta(1,1)^1$, reflecting that DMs' self-confidence threshold for relying on or ignoring the AI recommendation is uniformly spread over the spectrum; (2) UNBALANCED: $\tau \sim \beta(1,2)^1$, reflecting that DMs' self-confidence threshold leans towards the lower end of the spectrum; (3) U-SHAPED: $\tau \sim \beta(0.5, 0.5)^1$, reflecting that DMs' self-confidence threshold tends to be either very low or very high; (4) INV-U: $\tau \sim \beta(2,2)^1$, reflecting that DMs' self-confidence threshold leans towards the middle of the spectrum; (5) δ: an impulse at 0.75, reflecting that DMs' self-confidence threshold is fixed.

[1] Distributions are rescaled to reflect that confidence on binary classification task varies between 0.5 and 1, instead of 0 and 1.

We randomly divide our synthetic dataset into the training and test folds based on a $80 : 20$ split. Given the training set, we train random forest classifiers with maximum tree depth of 5 as our AI models. The baseline model is trained using the standard loss (Equation 1), while the five other complementary AI models are trained using the team loss following the CBIW method (i.e., $w_i = 1 - F_T(\mathcal{C}_i)$), and each model corresponds to one threshold distribution listed above. Then, given each of the six AI models, we simulate the human-AI team decision on each test instance following the threshold-based model (Equation 3) and determine its accuracy by comparing the team decision against the ground truth label. We repeat this procedure for five times in total.

Evaluation Results. Figure 1 reports the comparison of the average human-AI team decision making accuracy on the test dataset, when human DMs are assisted by different AI models. We make three important observations: (1) Compared to the case when humans collaborate with the baseline AI model (red markers in Figure 1), for each of the 5 types of $f_T(\tau)$, when training the AI model using the corresponding optimal weighting function (markers with the same colors as the distribution names on the x-axis in Figure 1), we can see most significant increase in the human-AI joint decision making performance. (2) In most cases (except for when the true $f_T(\tau)$ is U-SHAPED), even if the instance weights are not optimal (i.e., computed based on incorrect assumptions about the threshold distribution), a notable human-AI team performance gain can still be found when humans collaborate with a complementary AI model rather than the baseline AI model. (3) The heuristic weighting function $w_i = 1 - \mathcal{C}_i$ (green markers in Figure 1), which does not rely on knowledge of $f_T(\tau)$, seems to be a good default choice that can lead to reasonable team performance gains in many cases. Based on these findings, we use this heuristic weighting function for convenience in the experiments below, unless stated otherwise[2].

5.3 Evaluating Varied Expertise Overlap

Next, we systematically vary human DMs' expertise overlap with the baseline AI model to identify under what conditions the proposed CBIW-training method may lead to the largest gain in the human-AI joint decision making performance.

We create five sets of human DMs' independent decision data to simulate that human DMs have varying levels of expertise overlap with the baseline AI (i.e., very high, high, medium, low, very low). In our setting, the baseline AI is more accurate on the privileged applicants as they are the majority group. Differing degrees of expertise overlap are achieved by adjusting the comparison between acc_{priv} and acc_{unpriv} (i.e., humans' independent decision accuracy) from being consistent with that of the baseline AI ($acc_{priv} > acc_{unpriv}$, high overlap) to being opposite to that of the baseline AI ($acc_{priv} < acc_{unpriv}$, low overlap)[3]. We concurrently try to keep the overall accuracy of humans' independent decisions relatively stable. We again set $\hat{acc}_g = acc_g, \Delta_u = \Delta_o = 0.1$, and assume the human self-confidence threshold is sampled from $\tau \sim U[0.8, 0.9]$.

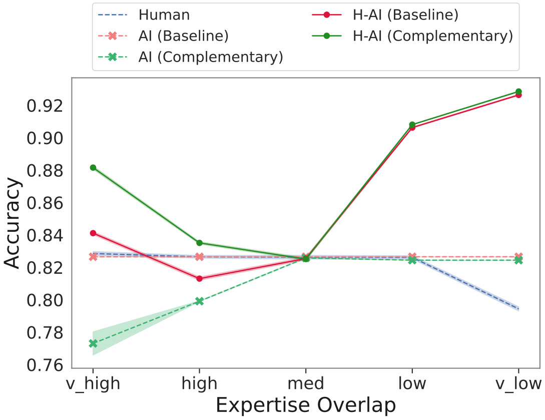

Figure 2: Impacts of the expertise overlap between humans and standard AI on human-AI team performance gains from the complementary AI (see differences between solid green and red lines).

Figure 2 shows the evaluation results. While our proposed method outperforms standard approach in both high and low expertise overlap settings, we find that it leads to considerably larger human-AI team performance gains when the baseline AI model has high expertise overlap with humans (i.e., it is not complementary already). This is because when the humans have low expertise overlap with the baseline AI, the baseline model due to being accurate yet dissimilar is "complementary" by itself, and becomes largely similar to the AI model obtained from using the proposed CBIW-training method.

5.4 Evaluating Varied Confidence Distributions

Finally, we examine how the human-AI team performance gains brought up by the CBIW method vary with human DM's degree of confidence calibration. In this simulation, we set $acc_{priv} = 0.9$, $acc_{unpriv} = 0.6$, $\Delta_u = \Delta_o = 0.2$, and $\tau \sim U[0.8, 0.9]$. To reflect that DMs may be *under-confident* on instances where they are accurate, we assume that $\hat{acc}_{priv} = acc_{priv} = 0.9$ with probability λ, while $\hat{acc}_{priv} = 0.7$ with probability $1 - \lambda$. Similarly, to reflect that DMs may be *over-confident* on instances where they are inaccurate, we assume that $\hat{acc}_{unpriv} = acc_{unpriv} = 0.6$ with probability ω, while $\hat{acc}_{unpriv} = 0.8$ with probability $1 - \omega$. Intuitively, the smaller λ and ω are, the more uncalibrated DMs' confidence is. To further highlight the importance of incorporating human *confidence* rather than human *accuracy* in developing behavior-aware AI for AI-assisted decision making, in addition to training the complementary AI model following the proposed CBIW method, we also train another complementary AI model following an *accuracy-based instance weighting* (ABIW) method by assuming perfect confidence calibration ($w_i = 1 - acc_i$). Figure 3 shows our results,

[2] We also conduct another simulation in which $f_T(\tau) = U[\tau_{avg} - 0.05, \tau_{avg} + 0.05]$, $\tau_{avg} \in \{0.55, 0.65, 0.75, 0.85, 0.95\}$ to examine how the average self-confidence threshold τ_{avg} influences the human-AI team decision making accuracy gains when humans are assisted by a CBIW-trained AI versus the standard AI. Our results suggest that our proposed CBIW-training strategy is robust to the average self-confidence threshold changes, often leading to substantial gains over the standard training approach. See SM for details.

[3] The Pearson correlation between humans' and the baseline AI model's decisions decreases gradually from 0.52 to 0.31 as we go from "very high" to "very low" expertise overlap dataset.

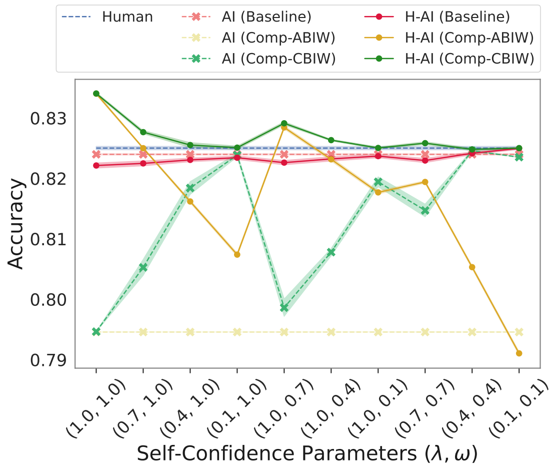

Figure 3: Impacts of humans' confidence calibration on human-AI team performance gains from the complementary AI obtained using the CBIW training method (see the solid green line).

suggesting: (1) DMs assisted by CBIW-trained AI consistently outperform DMs assisted by the baseline AI (see solid green and red lines), regardless of how uncalibrated their confidence is; and (2) DMs assisted by CBIW-trained AI outperform DMs assisted by ABIW-trained AI (see solid green and yellow lines), especially when DMs' confidence is uncalibrated.

6 Human Subject Experiments

To examine the effectiveness of our proposed method in improving the human-AI team performance in real-world AI-assisted decision making settings, we conduct a large-scale, randomized experiment with real human subjects.

Experimental Task. In this experiment, subjects are recruited to complete image classification tasks with the assistance of an AI model. We curate a subset of the widely used ImageNet dataset [Deng *et al.*, 2009], consisting of classes and instances that present varying levels of difficulty for humans. Specifically, we select 10 classes, comprising five easily recognizable objects (Church, Garbage Truck, Gas Pump, Golf Ball and Parachute) and five challenging dog breeds (Australian Terrier, Border Terrier, Dingo, Old English Sheepdog, and Rhodesian Ridgeback). The resulting dataset, which we name WoofNette, consists of a total of $9,446$ training images and $4,054$ test images, each of size $128 \times 128 \times 3$. Images that subjects are asked to classify in our experiment are randomly sampled from a 300-image subset of the WoofNette test set.

AI Training. We utilize the ResNet-50 architecture to train the AI models that we use in our experiment. The baseline AI model is established by fine-tuning the ResNet-50 network on the WoofNette training dataset to minimize the standard cross-entropy loss. On the other hand, we train the complementary AI model by minimizing the human-confidence-based, instance-weighted cross-entropy loss. For simplicity, we adopt the heuristic weighting function $w_i = 1 - \mathcal{C}_i$.

Training the complementary AI model requires the knowledge of human DMs' self-confidence $\mathcal{C}_i$ on different training data instances (i.e., different images). To estimate $\mathcal{C}_i$, we conducted a pilot study on Amazon Mechanical Turk (MTurk), in which subjects were asked to complete 18 image classification tasks *independently*, without any AI assistance. The images were randomly sampled from a 500-image subset from the WoofNette training dataset. In total, 206 subjects attended our pilot study, leading to 4644 image classifications, with about 9 classifications on each image. Based on the pilot study data, for each image, we used the *inter-annotator agreement*—the proportion of subjects whose classification on this image matches the majority classification—as a proxy for humans' self-confidence on it (i.e., higher agreement indicates greater confidence in humans' independent judgments)[4]. To generalize the human self-confidence estimation to other images outside of the 500-image subset used, we further leveraged the pre-trained ResNet-50 architecture to train an AI model for predicting humans' self-confidence on each image, i.e., $\hat{\mathcal{C}}_i = g(\mathbf{x}_i)$. Thus, when training the complementary AI model, the weight of each training instance is set as $w_i = 1 - \hat{\mathcal{C}}_i$ based on the predicted human self-confidence on the instance.

Note that training AI models to reach the optimal performance leads to extremely high AI accuracy, which limits the potential for achieving human-AI complementarity in joint decision making. Thus, in our experiment, we train the AI models for fewer epochs—we stop the training once the AI model reaches a target accuracy of 65%, which is close to humans' independent accuracy that we observed in our pilot study on this image classification task[5]. As a result, the test accuracy of the baseline AI model and complementary AI model we use in the experiment is 69% and 65%, respectively.

Experimental Treatments and Procedure. We include two treatments in our experiment. In the *control* treatment, subjects are assisted by the baseline AI model (i.e., the "standard AI") to complete the image classification tasks, while subjects in the *experimental* treatment are assisted by the complementary AI model in the image classification tasks.

We post our experiment on MTurk as a human intelligence task (HIT) and recruit MTurk workers as our subjects. Upon arrival, each subject is randomly assigned to one of the two treatments. Subjects start the HIT by completing a tutorial, which describes the image classification tasks that they need to work on in the HIT. At the end of the tutorial, subjects are asked to complete an example task, and they could only proceed to the actual experiment after making correct classification in this example task. After completing the tutorial, subjects start to work on a set of 18 image identification tasks under the AI assistance, and the images used in these tasks are randomly sampled from the WoofNette test dataset.

Our experiment was only open to US workers who have completed more than 100 HITs on MTurk and with a 90+% approval rate. Each subject could participate in the experiment only once. We included two attention check questions in our

[4] High confidence here does not necessarily imply high accuracy; it is possible that humans tend to agree with each other (hence high confidence) on some tasks but they agree on a wrong decision.

[5] We conducted additional simulation experiments by varying the target AI accuracy. We found that the human-AI team equipped with complementary AI consistently outperformed the standard AI across a range of target AI accuracy values, though the target AI accuracy did affect the magnitude of the gains. See SM for additional details.

				AI Accuracy		H-AI Team Accuracy	
Data	# Instances	# Classifications	Human Accuracy	Standard	Complementary	Standard	Complementary
Objects	150	1144	0.86	0.86	0.63^{***} ↓	0.82	0.86 ↑
Dogs	150	1196	0.46	0.61	0.71^{***} ↑	0.55	$\mathbf{0.64}^{**}$ ↑
High Conf	150	1148	0.72	0.85	0.64^{***} ↓	0.81	0.85 ↑
Low Conf	150	1192	0.47	0.61	0.70^{**} ↑	0.57	$\mathbf{0.65}^{**}$ ↑
Overall	300	2340	0.66	0.73	0.67^{**} ↓	0.68	$\mathbf{0.75}^{***}$ ↑

Table 1: Comparing the decision accuracy of the human-AI team on different subsets of data when human subjects are assisted by the standard or the complementary AI model. In the "AI Accuracy" and "H-AI Team Accuracy" columns, we compare the values corresponding to the complementary AI model and the values corresponding to the standard AI model. We use ↑ (↓) to indicate that the value corresponding to the complementary AI model is larger (smaller). Moreover, *, ** and *** indicate the difference is statistically significant, with $p < 0.05, p < 0.01$ and $p < 0.001$, respectively. Human Accuracy values are for reference only; they are collected from a separate pilot study in which subjects complete classification tasks without AI assistance on a different subset of the WoofNette dataset than the one we used in our experiment. High Conf and Low Conf refer to the two sets of task instances where the predicted human self-confidence was above or below the median value.

experiment, asking subjects to choose a randomly specified option in these questions. Only data from subjects who answered both attention check questions correctly was considered as valid. The base payment of the task was \$1.2. In addition, we provided a performance-based bonus to encourage subjects to make decisions to the best of their abilities—if a subject's decision accuracy was higher than 70%, we provided them with extra 5 cents for each correct decision that they made.

Experimental Results. After filtering the data from inattentive subjects, we obtained valid data from 130 subjects. We find that when subjects are assisted by the complementary AI model, the resulting decision accuracy of the human-AI team is 75%, which is higher than those subjects who are assisted by the standard AI model and achieve an accuracy of 68%. A t-test suggests that the accuracy difference between subjects in the two treatments is statistically significant ($p < 0.001$).

We then take a closer look into our experimental data to gain insights into *why* the use of the complementary AI model leads to increased human-AI team accuracy in AI-assisted decision making. First, based on how the WoofNette dataset is prepared, we conjecture that the complementary AI model may lead to increased decision accuracy of the human-AI team because it better supports human DMs in classifying the challenging dog breeds, on which DMs may have low self-confidence. We thus compare the human-AI team's decision accuracy between the two treatments for the five classes of easily recognizable objects and the five classes of dog breeds separately, and results are reported in Table 1 (top two rows). Indeed, we find that on dog classes, the complementary AI model's independent accuracy is significantly higher than that of the standard AI model, which further results in a significant increase in the human-AI team accuracy on them. Interestingly, even on the easily recognizable object classes, while the complementary AI model's independent accuracy is significantly lower than that of the standard AI model, human DMs also achieve a slightly higher accuracy (although insignificant) on these classes when they are assisted by the complementary AI model rather than the standard AI model.

One explanation for this observation is that even within easily recognizable object classes, human DMs may still find some task instances to be challenging and have low confidence in them, and our CBIW-training method allows the complementary AI model to better support human DMs on these instances. To test this explanation more directly, we split all images used in our human-subject experiments into two subsets based on the median value of the predicted human self-confidence on the images. We then compare the human-AI team's decision accuracy between the two treatments for the subset of images where human DMs have either high or low self-confidence, and results are reported in Table 1 (rows 3–4). As we expect, here, we find that the use of complementary AI model primarily results in increases in the human-AI team accuracy on those task instances where humans have low self-confidence, although on task instances where humans have high self-confidence, humans also seem to be able to avoid being misled by the less accurate recommendations made by the complementary AI model.

7 Conclusion

This paper contributes a novel behavior-aware AI design paradigm to enhance the human-AI team decision making performance in AI-assisted decision making. We address the challenge of improving human-AI joint decision making by designing AI-driven decision aids that take into account the real-world human behavior when interacting with AI. Our approach focuses on adjusting the training of AI models based on humans' confidence in their own decisions. We first formulate a threshold-based team decision making model that characterizes humans' willingness to adopt AI advice. We then propose a human-confidence-based instance weighting strategy for training complementary AI models. Extensive experiments are conducted on both the synthetic College Admission and the real-world WoofNette datasets to evaluate the effectiveness of the proposed behavior-aware AI training approach. Results of our experiments demonstrate that our proposed strategy can significantly improve the team performance, and such improvement is robust across a wide range of settings where human decision makers exhibit diverse behaviors. By considering the human factors and integrating them into the AI model design, we offer insights into how AI models can be tailored to human behavior to better support and complement humans in their decision-making processes.

Ethical Statement

This study was approved by our institute's IRB.

Acknowledgments

We thank the support of the National Science Foundation under grant IIS-2229876 and IIS-2340209 on this work. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

Contribution Statement

Syed Hasan Amin Mahmood and Zhuoran Lu contributed equally to this work.

References

[Bansal *et al.*, 2019] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019.

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

[Bansal *et al.*, 2021b] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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*, pages 1–16, 2021.

[Bertrand *et al.*, 2022] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. How cognitive biases affect xai-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society*, pages 78–91, 2022.

[Bondi *et al.*, 2022] Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. Role of human-ai interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5286–5294, 2022.

[Buçinca *et al.*, 2021] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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):1–21, 2021.

[Carroll *et al.*, 2019] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

[Chong *et al.*, 2022] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice. *Computers in Human Behavior*, 127:107018, 2022.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Dunning, 2011] David Dunning. The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*, volume 44, pages 247–296. Elsevier, 2011.

[Dvijotham *et al.*, 2023] Krishnamurthy Dvijotham, Jim Winkens, Melih Barsbey, Sumedh Ghaisas, Robert Stanforth, Nick Pawlowski, Patricia Strachan, Zahra Ahmed, Shekoofeh Azizi, Yoram Bachrach, et al. Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. *Nature Medicine*, 29(7):1814–1820, 2023.

[Haider *et al.*, 2022] Chowdhury Mohammad Rakin Haider, Chris Clifton, and Yan Zhou. Unfair ai: It isn't just biased data. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 957–962. IEEE, 2022.

[Kwon *et al.*, 2020] Minae Kwon, Erdem Biyik, Aditi Talati, Karan Bhasin, Dylan P Losey, and Dorsa Sadigh. When humans aren't optimal: Robots that collaborate with risk-aware humans. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pages 43–52, 2020.

[Langford and Clance, 1993] Joe Langford and Pauline Rose Clance. The imposter phenomenon: Recent research findings regarding dynamics, personality and family patterns and their implications for treatment. *Psychotherapy: theory, research, practice, training*, 30(3):495, 1993.

[Li *et al.*, 2023] Zhuoyan Li, Zhuoran Lu, and Ming Yin. Modeling human trust and reliance in ai-assisted decision making: A markovian approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6056–6064, 2023.

[Li *et al.*, 2024] Zhuoyan Li, Zhuoran Lu, and Ming Yin. Decoding ai's nudge: A unified framework to predict human behavior in ai-assisted decision making. *arXiv preprint arXiv:2401.05840*, 2024.

[Lu and Yin, 2021] Zhuoran Lu and Ming Yin. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[Madras *et al.*, 2018] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31, 2018.

[Mozannar *et al.*, 2022] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5323–5331, 2022.

[Rastogi *et al.*, 2022] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–22, 2022.

[Schemmer *et al.*, 2023] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 410–422, 2023.

[Wang and Du, 2018] Xiuxin Wang and Xiufang Du. Why does advice discounting occur? the combined roles of confidence and trust. *Frontiers in psychology*, 9:2381, 2018.

[Wang *et al.*, 2022] Xinru Wang, Zhuoran Lu, and Ming Yin. Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making. In *Proceedings of the ACM Web Conference 2022*, pages 1697–1708, 2022.

[Wilder *et al.*, 2020] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1526–1533, 2020.

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

[Zhang *et al.*, 2020] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 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*, pages 295–305, 2020.