Give Weight to Human Reactions: Optimizing Complementary AI in Practical Human-AI Teams

Hasan Amin [*1] Zhuoran Lu [*1] Ming Yin [1]

Abstract

With the rapid development of decision aids that are driven by AI models, the practice of human-AI joint 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 adjust the designs of the AI model underlying the decision aid by taking humans' reaction to AI into consideration. In particular, as humans are observed to accept AI advice more when their confidence in their own decision 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 of our proposed method in improving the human-AI joint decision making performance through systematic experimentation on both synthetic and real-world datasets.

1. Introduction

Systems leveraging Artificial Intelligence (AI) have seen wide-scale adoption in numerous application areas over the past few years (IBM, 2022). While many of them have had vast impact on their own, their independent utility is at times constrained by technical as well as socioethical limitations. This happens not only in high-stakes settings like criminal justice, where even a single wrong decision—by AI—has profound implications, but also in lower-stakes ones like sarcasm detection, where AI still struggles with fully addressing the task complexities (Abu Farha et al., 2022). In such scenarios, these AI systems can still be excellent aides to humans, who can make the overall decision making more efficient and effective by combining AI's assistance with their own formal and informal knowledge. This has motivated the formation of human-AI teams for joint decision making, which is being utilized in varied domains, from criminal justice (Green & Chen, 2019) and healthcare (Futoma et al., 2017) to credit lending (Kruppa et al., 2013) and content moderation (Link et al., 2016).

Complementarity, wherein the team members understand, and subsequently supplement, each other's strengths and weaknesses, is by definition central to effective collaboration in human-AI teams—or really any team in general. Effective collaboration here reflects the possibility of a human-AI team outperforming its individual counterparts. Earlier studies to enhance complementarity have emphasized on improving humans for the purpose, with particular stress on understanding and improving human reliance on AI (Bansal et al., 2019a; Lu & Yin, 2021). However, the AI systems—which are often more tunable, predictable and scalable than their human teammates—mostly continue to be designed for maximum individual accuracy. A recent effort to optimize team accuracy instead showed promising results but expected gains were not accompanied by empirical ones, likely because of having strong assumptions on human behavior (Bansal et al., 2021).

There is an evident need for better modeling of human behavior with respect to collaboration with AI in real-world scenarios, and integration of the same in (re)design of AI while accounting for humans' reaction to it. At a higher level, we want the AI teammate to perform better on instances where the human decision maker "needs" it more. These needs are related to both humans' actual as well as self-perceived strengths and weaknesses. Human confidence is thus one intuitive choice as indicator of such needs. In fact, a recent study (Chong et al., 2022) suggests that confidence of humans in their own decision, rather than in AI, dictates their decision to accept AI recommendation.

In this paper, we propose to train a complementary AI by

[*]Equal contribution [1]Department of Computer Science, Purdue University, West Lafayette, IN, USA. Correspondence to: Hasan Amin <hasanamin@purdue.edu>.

AI & HCI Workshop at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA.

using human-confidence-based instance weighting, instead of the standard empirical risk minimization where all instances are weighted equally. By upweighting instances where human decision maker has low self-confidence, our objective is to guide the AI towards regions of expertise that complement those of humans. The use of confidence, or perceived accuracy, rather than actual accuracy for instance weighting is particularly advantageous in mitigating the impact of cognitive biases exhibited by humans when interacting with AI, as these biases often stem from erroneous human perceptions and beliefs. More involved analysis later reveals that our proposed strategy provably optimizes for team performance under a suitable model for biased decision making. To validate the effectiveness of our approach, we conduct a systematic experimentation to determine the conditions under which our proposal yields maximum gains. Our results indicate that when distinct regions of expertise are present, the AI model trained using our proposed method effectively develops complementary expertise, with the greatest gains observed when the AI trained using the standard approach exhibits significant overlap in expertise with the human teammate. While factors such as confidence calibration and individual accuracy influence the degree of effectiveness, the instance-weighted AI generally remains a superior teammate even when these factors assume sub-optimal values, making it especially useful for real-world scenarios and practical human-AI teams.

Related Work. Recent advancements in AI technologies have sparked a surge of research in the field of human-AI collaboration, exploring various aspects of interaction and cooperation between humans and AI systems (Ong et al., 2012; Nguyen et al., 2022; Siemon, 2022). Studies have sought to understand and foster human-AI complementarity, with efforts focused on delegability problem, i.e., to identify when should each individual's expertise be leveraged for enhanced team performance (Steyvers et al., 2022; Lubars & Tan, 2019; Holstein & Aleven, 2021). Some researchers have tackled the challenge of improving human reliance on AI by developing mental models for AI and AI trust (Bansal et al., 2019a;b; Zhang et al., 2022), while others have explored more direct approaches like exemplar-based teaching (Mozannar et al., 2022). Additionally, there is a growing body of work investigating human behavior patterns when interacting with AI, particularly in relation to how human cognitive biases influence their reliance on AI (Zhang et al., 2020; Schemmer et al., 2023). Such work highlights the importance of human factors for effective human-AI collaboration, especially confidence as an indicator of human inclination to accept AI recommendation (Chong et al., 2022; Wang et al., 2022; Lu & Yin, 2021). However, so far, only a few studies attempt to take these into consideration when designing AI, by identifying challenging instances for individual agents (Wilder et al., 2020) or directly optimizing for team utility (Bansal et al., 2021).

2. Problem Setup

In a human-AI joint decision making setting, given the decision making case characterized by features $\mathbf{x} \in \mathcal{X}$, the human-AI team needs to make a decision $y \in \mathcal{Y}$. In this study, we focus on a popular human-AI joint decision making setting which is often referred to as "*AI-assisted decision making*", where an AI model provides a decision recommendation $y_m = m(\mathbf{x}; \theta_m)$ to a human decision maker—who may have their own independent judgement $y_h = h(\mathbf{x}; \theta_h)$ on this case—and the human decision maker needs to make the final team decision d. Without loss of generality, we focus on multiclass classification tasks in this study (i.e., $\mathcal{Y} = \{1, 2, \ldots, K\}$).

To obtain the AI model, we have a training dataset which comprises N feature-label pairs, i.e., $\mathcal{D} = \{I_1, I_2, \ldots, I_N\}$ where $I_i = (\mathbf{x_i}, y_i)$. A common practice adopted to train the AI model is to learn the model parameters θ_m that minimize the empirical risks over the entire training dataset:

$$\theta_m = \arg\min_{\theta'_m} \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x_i}, y_i) \in \mathcal{D}} \ell(m(\mathbf{x_i}; \theta'_m), y_i) \tag{1}$$

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 decision maker's contribution to the decision making process. Assuming that the human decision maker'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 decision maker'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:

$$\begin{aligned} \theta_m &= \arg\min_{\theta'_m} L_{team} \\ &= \arg\min_{\theta'_m} \tfrac{1}{|\mathcal{D}|} \textstyle\sum_{(\mathbf{x_i}, y_i) \in \mathcal{D}} \ell\big(f(\mathbf{x_i}, m(\mathbf{x_i}; \theta'_m), h(\mathbf{x_i}; \theta_h)), y_i\big) \end{aligned} \tag{2}$$

Interestingly, recent empirical studies suggest that when collaborating with an AI model in decision making, human decision makers 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 & Du, 2018; Schemmer et al., 2023; Wang et al., 2022). Thus, when a human confidence oracle C that provides us with human self-confidence on each decision making instance (i.e., $C : H(X) \mapsto [0, 1]$) is available, this empirical insight can be reflected by a threshold-based team decision making model:

$$f(\mathbf{x_i}; m(\mathbf{x_i};\theta_m); h(\mathbf{x_i};\theta_h)) = \begin{cases} h(\mathbf{x_i};\theta_h) & \text{if } C_i > \tau \\ m(\mathbf{x_i};\theta_m) & \text{otherwise} \end{cases} \quad (3)$$

where $C_i := C(h(\mathbf{x_i};\theta_h))$ is the human decision maker's self-confidence on instance i, and τ is the self-confidence threshold for the human decision maker to rely on or ignore the AI recommendation—humans will rely on the AI recommendation if their self-confidence is below τ, thus a higher value of τ is associated with a higher frequency for humans to rely on the AI recommendation.

In this paper, as an initial step to better factor the human decision maker'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.

3. Human-Confidence-Based Instance Weighting

When the human-AI team uses the threshold-based model to determine their joint decisions, humans 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 recommendation more* in order to optimize for the human-AI team performance. To operationalize this idea, we propose to train a 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 decision maker's self-confidence on it (C_i):

$$\theta_c = \arg\min_{\theta_c'} \sum_{(\mathbf{x_i};y_i)\in D} w_i \cdot \ell(m_c(\mathbf{x_i};\theta_c'); y_i) \quad (4)$$

Intuitvely, the standard AI model $y_m = m(\mathbf{x};\theta_m)$ weighs all instances equally (i.e., $w_i = 1\ \forall I_i \in D$). In general, without additional information about the value of the self-confidence threshold τ, we have the following proposition:

Proposition 3.1. *If the human decision maker is less confident about I_i than I_j, then I_i should be weighted at least as high as I_j, i.e., $w_i \geq w_j$ if $C_i < C_j$.*

Proof intuition. Given the unknown self-confidence threshold τ, if $C_i < C_j$, we have $C_j \leq \tau \implies C_i \leq \tau$ but $C_j > \tau \not\implies C_i > \tau$. In other words, I_i is always in "low confidence region" (i.e., below the self-confidence threshold τ) when I_j is in low confidence region, and I_i could be in low confidence region even when I_j is not. Since we aim to maximize the AI model's performance in the low confidence region where humans will adopt its recommendation, training data instances more likely to be in the low confidence region should be weighted at least as highly, i.e., $w_i \geq w_j$.

Following this proposition, we may propose a few heuristic methods for setting the weight for each training data instance, e.g., $w_i = 1 - C_i$ or $w_i = \frac{1}{C_i}$. 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.

3.1. Optimization for Known Self-Confidence Threshold

First, we consider the simplest scenario where the human decision maker 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 $D_h := \{I_i \mid C_i > \tau\}$ and $\bar{D}_h := D \setminus D_h$ be the sets of instances where human 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 region $\bar{D}_h$.

Proposition 3.2. *When the human decision maker uses a fixed and known self-confidence threshold τ to determine the human-AI team joint decision, the team loss is minimized when $w_i = \mathbb{1}[C_i \leq \tau]$.*

Proof. According to Equation 3, the team loss can be decomposed into "human loss" and "AI loss" as follows: $L_{team} = \frac{1}{|D|}\sum_{(\mathbf{x_i};y_i)\in D_h} \ell(h(\mathbf{x_i};\theta_h); y_i) + \frac{1}{|D|}\sum_{(\mathbf{x_i};y_i)\in \bar{D}_h} \ell(m_c(\mathbf{x_i};\theta_c); y_i)$. Since we can directly optimize AI only, the first term (i.e., the "human loss") is effectively a constant. This is equivalent to assigning a weight of 0 to instances in D_h and 1 to instances in $\bar{D}_h$, or setting $w_i = \mathbb{1}[C_i \leq \tau]$.

3.2. Optimization for Expected Self-Confidence Thresholds

In practice, however, humans' self-confidence threshold τ may not only be unknown to the AI model developer, but may also vary between individual human decision makers and across time. To reflect this, we consider a second scenario such that when facing a decision making instance, the human decision maker will need to draw a threshold value from a known distribution (i.e., $\tau \sim f_T(\tau)$) and then apply the threshold-based model to determine the final human-AI joint decision. In this case, the complementary AI model needs to be trained to minimize for the expected team loss over all possible self-confidence thresholds.

Proposition 3.3. *When the human decision maker draws a self-confidence threshold from a known distribution to determine the human-AI team joint decision, i.e., $\tau \sim f_T(\tau)$,* the expected team loss is minimized when $w_i = 1 - F_T(C_i)$, where $F_T(\cdot)$ is the cumulative distribution function (CDF) for threshold value τ.

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

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

Thus, minimizing $\mathbb{E}[L_{team}]$ is equivalent to minimizing $\sum_{(x_i,y_i)\in D} (1 - F_T(C_i))\, \ell(m_c(x_i;\theta_c), y_i)$, which implies $w_i = 1 - F_T(C_i)$.

Remarks. Following Proposition 3.3, we can see that the heuristic method of setting the weight of each training instance $w_i = 1 - C_i$ is in fact the optimal, when the human decision maker draws their self-confidence threshold from a uniform distribution, i.e., $\tau \sim U[0,1]$.

4. Evaluation

In this section, we conduct simulation studies on two datasets to evaluate that when human decision makers collaborate with an AI model trained following the proposed human-confidence-based instance weighting method, whether the human-AI team joint decision making performance improves compared to when they collaborate with an AI model trained following the standard method.

4.1. Simulation on Synthetic Data: College Admission

We conduct our first simulation study on a college admission decision making task, for which the dataset is generated entirely synthetically. Evaluation on this fully synthetic dataset is useful because: (1) we may artificially create a decision making scenario where human decision makers exhibit different levels of competence/confidence on different subsets of decision making tasks, so that the proposed method for training a complementary AI model is more likely to be beneficial; (2) we may systematically control characteristics of the human decision maker's behavior to examine the robustness of the proposed method in improving the human-AI joint decision making performance.

Synthetic Dataset Generation. Specifically, in this task, decision makers need to determine whether to admit an applicant to college (i.e., $Y = \{+1, -1\}$, $+1$ represents admitted while -1 represents rejected), given two features of the applicant—their Grade Point Average (i.e., "GPA") and their standardized test scores (i.e., "SCORE"). Inspired by Haider et al. (2022), we assume that applicants may either belong to the privileged group or underprivileged group. In addition, we assume that SCORE is more predictive of the admission outcome for privileged applicants, while GPA is more predictive for underprivileged applicants.

Generating decision making tasks. We start by generating a set of decision making task instances, where each instance is represented by a (x_{GPA}, x_{Score}, y) tuple. For each of the $n = 100,000$ instances (i.e., applicants), the values of x_{GPA} and x_{Score} are uniformly randomly sampled between 0 and 1; for both GPA and SCORE, we refer to a value that is above (below) a threshold t as high (low), and we use $t = 0.5$ in this study. The applicant is further assigned to the privileged group with probability r, and we use $r = 0.75$ in this study. Finally, we follow the steps below to determine the ground truth label y for each applicant:

1. If both x_{GPA} and x_{Score} are high, set $y = +1$ regardless of the group identity of the applicant;
2. For a privileged applicant, if x_{Score} is low, set $y = -1$; and if x_{Score} is high yet x_{GPA} is low, set $y = +1$ with a probability p that is proportional to the value of $x_{Score} + x_{GPA}$, i.e., the higher the $x_{Score} + x_{GPA}$ value is, the more likely the applicant will be admitted[1]. This reflects that SCORE is more predictive of the admission outcome for privileged applicants.
3. For an underprivileged applicant, if x_{GPA} is low, set $y = -1$; and if x_{GPA} is high yet x_{Score} is low, we again set $y = +1$ with a probability p that is proportional to the value of $x_{Score} + x_{GPA}$[1]. This reflects that GPA is more predictive of the admission outcome for underprivileged applicants.
4. Lastly, to account for a degree of randomness in the admission process, we will flip the label y currently set for the applicant with a small probability q. q is designed in a way such that when the current label $y = +1$, applicants with higher values of x_{GPA} +

[1] We operationalize this by mapping the value of x_{Score} + x_{GPA} to a p value in the interval between 0.5 and 1.

x_{Score} will have smaller q (thus less likely to be flipped to "rejected"), while when $y = 1$, applicants with smaller values of $x_{GPA} + x_{Score}$ will have smaller q (thus less likely to be flipped to "admitted").[2]

A visualization of the generated dataset is provided in Figure A.1 in the Appendix.

Generating human decision makers' behavior. To reflect that humans have varying levels of competence/confidence 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 decision maker's independent judgment y_H such that it is correct with a probability of acc_g. Further, the human decision maker's confidence on this instance is randomly sampled from a uniform distribution $U[acc_g \quad under; acc_g + over]$; we may systematically vary the values of $under$ and $over$ to control the human decision makers' confidence calibration degree. Finally, the decision maker's self-confidence threshold on this instance is randomly sampled from a distribution $f_T(\ )$, and we experiment with different distributions.

Evaluations with Different Threshold Distributions. Given our synthetic dataset, we first evaluate the effectiveness of the proposed AI training method in improving the human-AI team performance when human decision makers have different self-confidence threshold distributions in determining the team decisions (i.e., $f_T(\ )$). As shown in Proposition 3.3, given a specific self-confidence threshold distribution $f_T(\ )$, the optimal weighting function to be used to train the complementary AI model is $w_i = 1 \quad F_T(C_i)$. However, knowing or being able to reliably estimate the precise format of $f_T(\ )$ 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 complementary AI training method.

Evaluation Setup. In this evaluation, we assume human decision makers' independent judgments are more accurate on applicants from the privileged group. Thus, we set $acc_{priv} = 0{:}9$ and $acc_{unpriv} = 0{:}6$. We further set $under = 0{:}1$ and $over = 0{:}1$ (i.e., human decision makers' confidence is relatively well calibrated). Moreover, we consider 5 types of self-confidence threshold distributions: (1) UNIFORM: $(1;1)$[3], reflecting the case that decision makers' self-confidence threshold for relying on or ignoring the AI recommendation is uniformly spread over the spectrum; (2) UNBALANCED: $(1;2)$[3], reflecting the case that human decision makers' self-confidence threshold leans towards the lower end of the spectrum; (3) U-SHAPED: $(0{:}5;0{:}5)$[3], reflecting the case that decision makers' self-confidence threshold tends to be either very low or very high; (4) INV-U: $(2;2)$[3], reflecting the case that decision makers' self-confidence threshold leans towards the middle of the spectrum; (5): an impulse at $0{:}75$, reflecting the case that decision makers' self-confidence threshold is fixed at a single value.

We randomly divide our synthetic dataset into the training and test folds based on a $80:20$ split. Given the training dataset, 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 human-confidence-based instance weighting method (i.e., $w_i = 1 \quad F_T(C_i)$), and each model corresponds to one threshold distribution as listed above (i.e., UNIFORM, UNBALANCED, U-SHAPED, INV-U,). Then, on the test dataset, given each of the six AI models, we simulate the human-AI team decision on each instance following the threshold-based model (Equation 3) and determine its accuracy by comparing 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 decision makers are collaborating with different AI models. We make the following observations: (1) Compared to the case when humans collaborate with the baseline AI model, for each of the self-confidence threshold distributions we consider, when training the AI model using the corresponding optimal weighting function, we can see a significant increase in the human-AI joint decision making performance. (2) In most cases (except for when the true self-confidence threshold distribution 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 $w_i = 1 \quad C_i$, which does not rely on knowledge or estimation of the self-confidence threshold distribution, seems to be a good default choice that can lead to reasonable team performance gains.

Evaluation with Different Human Characteristics. In our second evaluation, we systematically vary a number of characteristics of the human decision makers, including their expertise overlap with the baseline AI model, their average self-confidence threshold for relying on or ignoring the AI recommendation, and their confidence calibration degree. We aim to use this evaluation to identify under what

---

[2] We operationalize this by mapping the value of $x_{Score} + x_{GPA}$ to a q_0 value in the interval between 0 and 0.1. Then, when $y = +1$, $q = 0{:}1 \quad q_0$, and when $y = 1$, $q = q_0$.

[3] Distributions are re-scaled to accommodate for the fact that confidence on binary classification task varies between 0.5 and 1, instead of 0 and 1.

Figure 1. The human-AI team decision making accuracy when human decision makers' self-confidence thresholds are drawn from different distributions (x-axis) and collaborate with AI models trained using different human-confidence-based instance-weighting strategies. Error shades represent the standard errors of the mean. Optimal weighting strategy as per Proposition 3.3 indeed results in the highest team performance (large, thick marker has the largest value on y-axis for every self-confidence distribution on x-axis), but other strategies also often lead to reasonable team performance gains against the baseline AI model.

conditions the proposed method may lead to the largest gain in the human-AI joint decision making performance. For simplicity, we adopt the heuristic weighting function $w_i = 1 - C_i$ for training the complementary AI model in this evaluation. On the other hand, human self-confidence threshold is sampled from $\tau \sim U[0.7, 0.8]$. Consistent with previous evaluation setup, we use $acc_{priv} = 0.9$ and $acc_{unpriv} = 0.6$, and set $\theta_{under} = 0.1$ and $\theta_{over} = 0.1$ in general.

Impact of Expertise Overlap between Humans and the Baseline AI model. We first examine the team performance gain brought up by the proposed instance-based weighting method to train complementary AI models when the human decision makers have varying levels of expertise overlap with the baseline AI model. In our setting, the baseline AI is found to be more accurate on the privileged applicants as they are the majority group. We then create 5 sets of human decision makers' independent decision data with varying levels of human-AI expertise overlap (i.e., very high, high, medium, low, very low) by controlling the humans' independent decision accuracy comparison on the two groups to change from being consistent with that of the baseline AI model (i.e., $acc_{priv} > acc_{unpriv}$, high overlap) to being opposite to that of the AI model (i.e., $acc_{priv} < acc_{unpriv}$, low overlap), while ensuring the overall accuracy of humans' independent decision does not change much[4].

Figure 2a shows the evaluation results. We find that the proposed method leads to the largest 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 understandable, as when the humans have low expertise overlap with the baseline AI model, the baseline model is "complementary" by itself and becomes largely similar to the AI model obtained from using the proposed human-confidence-based instance-weighting training method.

Impact of Average Self-Confidence Threshold. The human self-confidence threshold (from Equation 3) reflects the dependency of humans on AI, with a higher value indicating human decision makers would rely on AI recommendation more frequently. Beyond evaluating the impact of type of threshold distribution, as done earlier (Figure 1), we are also interested in evaluating how the (average) values of this threshold impact human-AI team performance gain. Our default setting of $\tau \sim U[0.7, 0.8]$ (i.e., $\tau_{avg} = 0.75$) maps to medium self-confidence on average. We change the sampling distribution to $U[0.5, 0.6], U[0.6, 0.7], U[0.8, 0.9]$ and $U[0.9, 1.0]$ to represent very low, low, high and very high values of average self-confidence respectively. The instance weighting function $w_i = 1 - C_i$ remains unchanged.

Figure 2b shows the evaluation results. We find that the proposed method leads to the largest human-AI team performance gains when the self-confidence threshold takes on moderate values on average. This is understandable because both humans and AI may often contribute to the final team decision here, and our complementary model gets a chance to leverage its complementary strengths. When τ_{avg} is very low, human decision maker mostly discards AI recommendation so team accuracy is close to human accuracy with limited gains from complementary AI model. When τ_{avg} is very high, human decision maker mostly accepts AI recommendation so team accuracy is close to AI accuracy with negative gains from complementary AI model; this is expected since complementary AI typically sacrifices individual accuracy on entire data to be able to focus on instances where human decision makers need it more.

Impact of Human Confidence Calibration. We assumed human self-confidence to be well-calibrated till now. While this is a common assumption, especially under popular rational decision making that intrinsically relies on it, we know that this seldom holds in practice. Therefore, we examine how gains by our proposed instance weighting method vary with varying levels of confidence calibration. Under our

[4] The Pearson correlation between humans' and the baseline AI model's decisions decreases gradually from 0.53 to 0.28 as we go from "very high" to "very low" expertise overlap dataset.

Figure 2. Impact of different human charactersitics on gains from complementary AI (difference between solid green and red lines).

default setup, the human decision maker's confidence on an instance from group g is randomly sampled from a uniform distribution $U[acc_g \; under; acc_g + over]$, and we have been using $over = 0.1$ and $under = 0.1$ so far. This represents the well-calibrated setting. We test out four additional settings here: $under = 0.2$ and $over = 0$ to represent very high degree of underconfidence, $under = 0.1$ and $over = 0$ to represent high degree of underconfidence, $under = 0$ and $over = 0.1$ to represent high degree of overconfidence, and $under = 0$ and $over = 0.2$ to represent very high degree of overconfidence.

Figure 2c shows the evaluation results. We find that our proposed method exhibits robustness to varying degrees of confidence calibration and consistently yields substantial gains. Since the same potentially miscalibrated confidence is used for both meta-decision to accept AI recommendation and instance weighting, the method mitigates the impact of calibration errors to a certain extent, contributing to its overall effectiveness. Maximum gains are attained when human is slightly underconfident here.

4.2. Simulation on Real World Data: WoofNette

For more realistic and interpretable experimental conditions, we sought to identify a vision dataset that ideally contains distinct "groups" of instances with room for complementarity (i.e., humans do not possess high and/or equal accuracy across all groups). With this objective in mind, we curated a subset of the widely used ImageNet dataset (Deng et al., 2009), consisting of classes and instances that present varying levels of difficulty for human classification. Ultimately, we selected 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), from ImageNet. The resulting dataset, named WoofNette, consists of a total of 9,446 training images and 4,054 test images, each of size 128 128 3. Sample images from the WoofNette dataset are provided in Figure A.2 in the Appendix.

Human behavior data. We conducted a pilot study on Amazon Mechanical Turk involving 200 images, with nearly 7 annotations per image. This allowed us to estimate the accuracy of human for each class. Human decision makers' independent judgment on images belonging to a certain class was then randomly simulated such that the probability that it was correct equals to humans' accuracy on that class. Moreover, for a given image, we take the proportion of workers in the pilot study whose annotation matches the majority annotation for this image as the proxy for humans' self-confidence on it (i.e., higher agreement with the majority indicated greater confidence in their independent judgments). However, since we only had this information for the 200 pilot study images, compared to the nearly 10,000 training images, we ended up using this data to develop a separate AI model for confidence prediction. More specifically, a ResNet-50 deep neural network, initialized with standard ImageNet weights, was trained to predict humans' self-confidence based on input images. This AI model provided self-confidence values for each task instance generated by our synthetic human. This model was then used to provide self-confidence values of our synthetic human on each task instance.

(a) Uniform Self-Con dence Threshold Distribution (b) Self-Con dence Threshold Distribution

Figure 3. Human, AI and Human-AI team performance on WoofNette using standard and complementary AI training strategies.

AI model training. We utilize the ResNet-50 architecture, which is pre-initialized with ImageNet weights, as the AI model. To establish a baseline AI, we train this model on the WoofNette dataset by minimizing the standard categorical cross-entropy loss. Additionally, to obtain a complementary AI, we train the AI model using human-con dence-based instance-weighted categorical cross-entropy loss. We again adopt the simple $(1 \; C_i)$ weighting scheme. However, training the AI model optimally leads to very high AI accuracy, limiting the potential for complementarity with humans. To create a more realistic scenario that aligns with practical human-AI interaction, we intentionally restrict the AI's accuracy by training it for fewer epochs. In fact, we explore the impact of AI accuracy on the observed gains by training both the baseline and complementary AI models until a speci ed “target accuracy” is reached (i.e., we consider the model as converged and stop training when the training accuracy surpasses the target accuracy).

Evaluation results. To obtain the team decision, we use try two self-con dence threshold distributions UNIFORM ($U[0{:}1; 1]$) and (impulse at 0.7). UNIFORM represents the most basic, uninformative scenario. On the other hand, may be more representative here as it would lead to two high and low con dence regions, which is what we expect with easy object images and dif cult dog images. We get signi cant gains, especially for lower target accuracy, using our proposed training method in both cases, although the absolute improvement in accuracy is much higher in case of (Figure 3). As expected, the gains are higher when target AI accuracy is lower since there is more room for contribution by human teammate

5. Conclusion

In this paper, we address the challenge of improving human-AI joint decision making by designing AI-driven decision aids that take into account humans' reactions when interacting with it. Our approach focuses on adjusting the AI models based on humans' con dence in their own decisions. We rst formulated a threshold-based team decision making model that characterizes humans' willingness to adopt AI advice only when they have low con dence in their own decisions. We then proposed a human-con dence-based instance-weighting strategy for training complementary AI models. Under the assumed decision making model, we also derived optimal weighting strategies, and conducted experiments on both synthetic College Admission and real-world WoofNette datasets. The results of our experiments demonstrated that our proposed strategy can signi cantly improve the performance of human-AI joint decision making, even under suboptimal settings like when human con dence is not well-calibrated, making our solution particularly bene cial for use in practical setups. By considering the human factors and integrating them into the AI model design, we offer insights into how AI models can be tailored to better support humans in their decision-making processes. This could complement existing body of work that focuses on improving humans' capability to better utilize a given AI model. Future work will explore additional factors in uencing human acceptance of AI advice and investigate alternative methods for adjusting AI models based on human reactions, aiming to further enhance human-AI team performance and re ne the collaboration between humans and AI in decision-making processes across various domains.

References

Abu Farha, I., Oprea, S. V., Wilson, S., and Magdy, W. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 802–814. Association for Computational Linguistics, July 2022.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. 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, pp. 2–11, 2019a.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., and Horvitz, E. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2429–2437, 2019b.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11405–11414, 2021.

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., and Cagan, J. 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, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Futoma, J., Hariharan, S., Heller, K., Sendak, M., Brajer, N., Clement, M., Bedoya, A., and O'brien, C. An improved multi-output gaussian process rnn with real-time validation for early sepsis detection. In *Machine Learning for Healthcare Conference*, pp. 243–254. PMLR, 2017.

Green, B. and Chen, Y. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 90–99, 2019.

Haider, C. M. R., Clifton, C., and Zhou, Y. Unfair ai: It isn't just biased data. In *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 957–962. IEEE, 2022.

Holstein, K. and Aleven, V. Designing for human-ai complementarity in k-12 education. *arXiv preprint arXiv:2104.01266*, 2021.

IBM. Global ai adoption index 2022. https://www.ibm.com/downloads/cas/GVAGA3JP, 2022. Accessed: 2022-07-23.

Kruppa, J., Schwarz, A., Arminger, G., and Ziegler, A. Consumer credit risk: Individual probability estimates using machine learning. *Expert systems with applications*, 40(13):5125–5131, 2013.

Link, D., Hellingrath, B., and Ling, J. A human-is-the-loop approach for semi-automated content moderation. In *ISCRAM*, 2016.

Lu, Z. and Yin, M. 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*, pp. 1–16, 2021.

Lubars, B. and Tan, C. Ask not what ai can do, but what ai should do: Towards a framework of task delegability. In *Proceedings of NeurIPS*, 2019.

Mozannar, H., Satyanarayan, A., and Sontag, D. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5323–5331, 2022.

Nguyen, G., Taesiri, M. R., and Nguyen, A. Visual correspondence-based explanations improve ai robustness and human-ai team accuracy. *arXiv preprint arXiv:2208.00780*, 2022.

Ong, C., McGee, K., and Chuah, T. L. Closing the human-ai team-mate gap: how changes to displayed information impact player behavior towards computer teammates. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, pp. 433–439, 2012.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., and Satzger, G. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 410–422, 2023.

Siemon, D. Elaborating team roles for artificial intelligence-based teammates in human-ai collaboration. *Group Decision and Negotiation*, pp. 1–42, 2022.

Steyvers, M., Tejeda, H., Kerrigan, G., and Smyth, P. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119, 2022.

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

Wang, X., Lu, Z., and Yin, M. Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making. In *Proceedings of the ACM Web Conference 2022*, pp. 1697–1708, 2022.

Wilder, B., Horvitz, E., and Kamar, E. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020.

Zhang, Q., Lee, M. L., and Carter, S. You complete me: Human-ai teams and complementary expertise. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–28, 2022.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. 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*, pp. 295–305, 2020.

A. Dataset

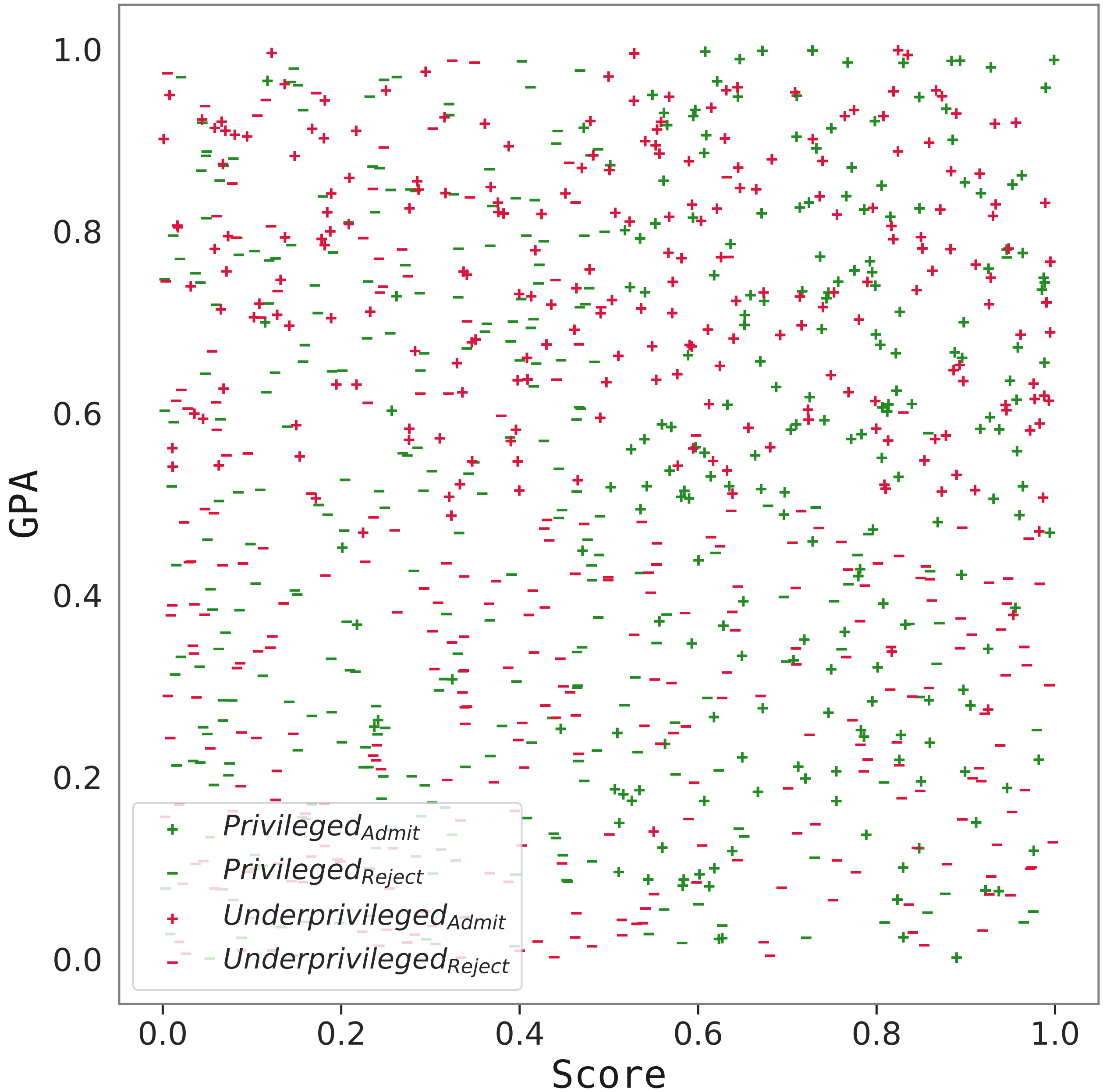

Figure A.1. Visualization of decision making task instances from the synthetic College Admission dataset.