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

Supplementary Material

A Datasets

A.1 College Admission

In this dataset, we mimic the college admission scenario, where decision makers need to determine whether to admit an applicant to college (i.e., $\mathcal{Y} = \{+1, -1\}$, $+1$ 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"). We assume that applicants may either belong to the privileged group or underprivileged group; we will later use these two groups to reflect that human decision makers may have different levels of confidence on different subsets of decision making instances. In addition, our synthetic dataset (i.e., a set of decision making instances in the form of (x_{GPA}, x_{Score}, y) tuple) is generated to reflect that SCORE is more predictive of the admission outcome for privileged applicants, while GPA is more predictive for underprivileged applicants. Privileged applicants with access to better schools and preparation material, and ability to retake the test multiple times are more likely to have a representative SCORE—and their admission decision could primarily be made based on that. On the other hand, underprivileged applicants better demonstrate their abilities via more school/context-specific GPA.

We start by generating a set of decision making task instances. For each of the n instances (i.e., applicants), the values of x_{GPA} and x_{Score} are uniformly randomly sampled between 0 and 1 without loss of generality; for both GPA and SCORE, we refer to a value that is above (below) a threshold t as *high* (*low*). The applicant is further assigned to the privileged group with probability r. Finally, we follow these steps 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 $x_{GPA} + x_{Score}$ 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].

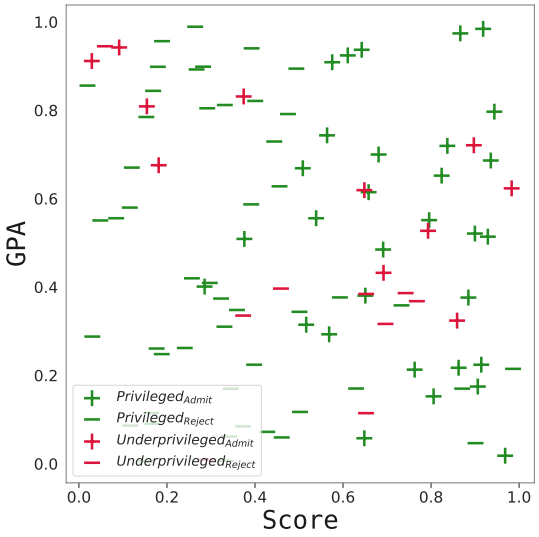


Figure A1: Visualization of decision making task instances from the synthetic College Admission dataset.

[1]We operationalize this by mapping the value of $x_{Score} + x_{GPA}$ to a p value in the interval between p_{min} (0.5 in our study) and p_{max} (1 in our study).

[2]We operationalize this by mapping the value of $x_{Score} + x_{GPA}$ to a q_0 value in the interval between 0 and q_{max} (0.1 in our study). Then, when $y = +1$, $q = 0.1 - q_0$, and when $y = -1$, $q = q_0$.

In the data we use for our experiments, we set number of instances $n = 100,000$ to have sufficiently large data, threshold $t = 0.5$ to refer an x value as high (low) if it is above (below) the mid point of the range, and $r = 0.75$ to have privileged applicants as the majority group. Sample instances from this dataset are provided in Figure A1.

A.2 WoofNette

The WoofNette dataset comprises images of 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 selection was motivated by two existing datasets, ImageWoof and ImageNette: github.com/fastai/imagenette. We select a reduced number of classes to avoid cognitive overload . It contains $9,446$ training images and $4,054$ test images, each of size $128 \times 128 \times 3$. Samples from the training data are shared in Figure A2.

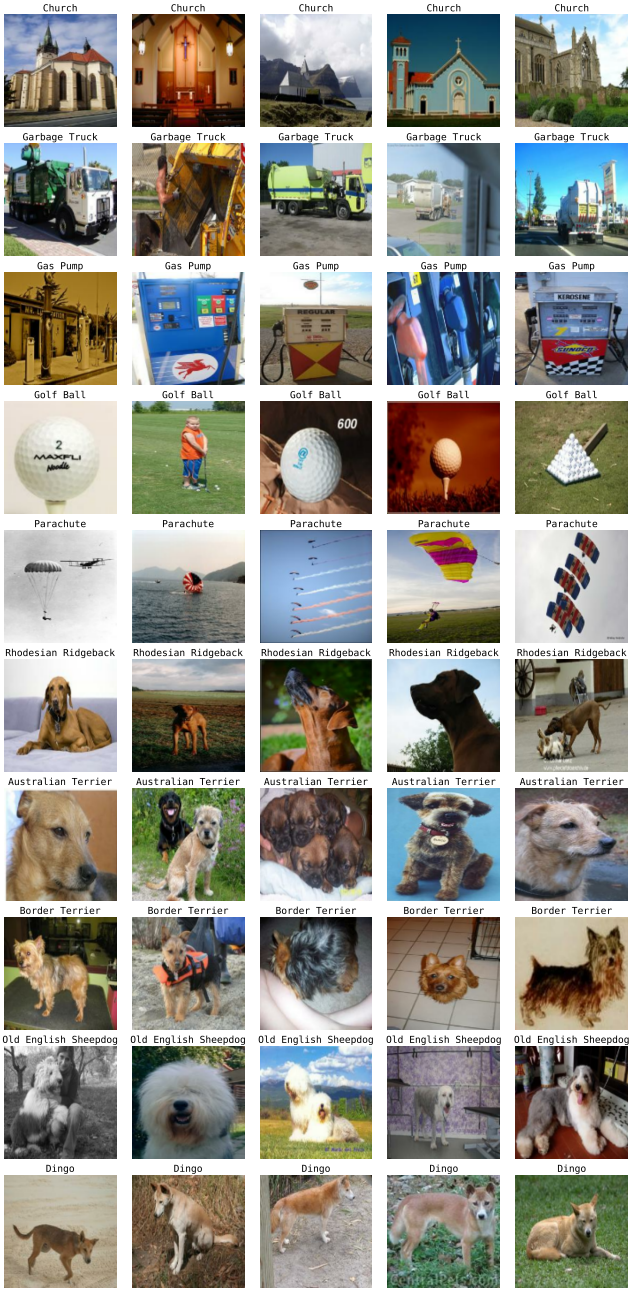


Figure A2: Sample images from the WoofNette dataset.

B Proofs for Propositions

We restate Propositions 1 and 2 from the main text, and provide the deferred proofs here. Recall that $\mathcal{D}_h := \{\mathcal{I}_i \mid \mathcal{C}_i > \tau\}$ and $\mathcal{D}_l := \mathcal{D} \setminus \mathcal{D}_h$ are the sets of instances where the human decision maker has high and low self-confidence respectively.

Proposition 1. *If the human decision maker 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. We aim to maximize the AI model's performance in the low confidence region $\mathcal{D}_l$, where humans will adopt its recommendation, so the training data instances more likely to be in $\mathcal{D}_l$ should be weighed higher. Thus to show that $w_i \geq w_j$ if $\mathcal{C}_i < \mathcal{C}_j$, we can show that expectation of $\mathcal{I}_i$ (instance i) being in $\mathcal{D}_l$ is higher than that of $\mathcal{I}_j$.

$$
\begin{aligned}
w_i &\propto \mathbb{E}[\mathcal{I}_i \in \mathcal{D}_l] \\
&= \int_0^1 f_T(\tau) \cdot \mathbb{P}[\mathcal{I}_i \in \mathcal{D}_l] \; d\tau \\
&= \int_0^1 f_T(\tau) \cdot \mathbb{1}[\mathcal{C}_i \leq \tau] \; d\tau \\
&= \int_{\mathcal{C}_i}^1 f_T(\tau) \; d\tau \\
&= 1 - F_T(\mathcal{C}_i)
\end{aligned}
$$

Since cumulative distribution function is non-decreasing, $\mathcal{C}_i < \mathcal{C}_j \implies F_T(\mathcal{C}_i) \leq F_T(\mathcal{C}_j) \implies w_i \geq w_j$.

Proposition 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 $w_i = \mathbb{1}[\mathcal{C}_i \leq \tau]$.*

Proof. According to Equation 3, the team loss can be decomposed as follows:

$$
\begin{aligned}
\mathcal{L}_{team} &= \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \ell(f(\mathbf{x}_i, m_c(\mathbf{x}_i; \theta_c), h(\mathbf{x}_i; \theta_h)), y_i) \\
&= \underbrace{\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_h} \ell(h(\mathbf{x}_i; \theta_h), y_i)}_{\text{human loss}} \\
&\quad + \underbrace{\frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_l} \ell(m_c(\mathbf{x}_i; \theta_c), y_i)}_{\text{AI loss}}
\end{aligned}
$$

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 $\mathcal{D}_h$ and 1 to instances in $\mathcal{D}_l$, or setting $w_i = \mathbb{1}[\mathcal{C}_i \leq \tau]$.

C Additional Results (College Admission)

C.1 Evaluating Varied Self-Confidence Thresholds

We are interested in investigating the impact of average self-confidence threshold on the human-AI team performance gains using our complementary AI training strategy. The human self-confidence threshold (τ) reflects the dependency of humans on AI, with a higher value indicating human DMs would adopt AI recommendation more frequently.

Our default setting of $\tau \sim U[0.8, 0.9]$ (i.e., $\tau_{avg} = 0.85$) maps to a high self-confidence threshold on average, which reflects the human DM will be receptive to AI recommendation unless they are fairly confident about their own decision. We then change the sampling distribution to $U[0.5, 0.6], U[0.6, 0.7], U[0.7, 0.8]$ and $U[0.9, 1.0]$ to represent very low, low, medium, and very high average values of self-confidence threshold, respectively. We again set $acc_{priv} = 0.9$, $acc_{unpriv} = 0.6$, $\hat{acc}_g = acc_g$, $\Delta_u = \Delta_o = 0.1$, and we continue to the use the heuristic-based instance weighting function: $w_i = 1 - \mathcal{C}_i$.

As reflected in Figure A3, 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 because both humans and AI get frequent opportunities to contribute to the final team decision here, and our complementary model gets a chance to exhibit its complementary strengths. When τ_{avg} is very low, human DM 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 DM mostly adopts AI recommendation so team accuracy is close to AI accuracy. Here, the complementary AI model leads to negative gains; this is expected since complementary AI typically sacrifices individual accuracy to be able to focus on instances where human DMs need it more.

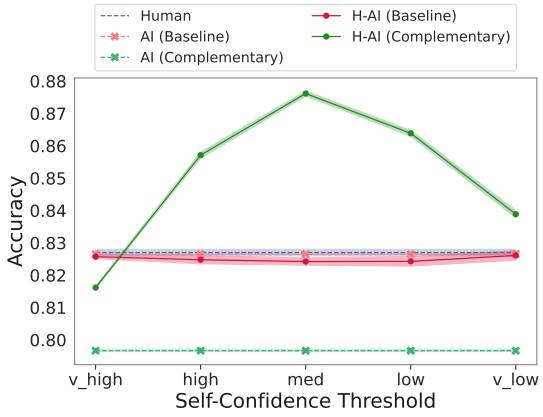


Figure A3: Impacts of the average human self-confidence threshold on human-AI team performance gains from the complementary AI (see differences between solid green and red lines).

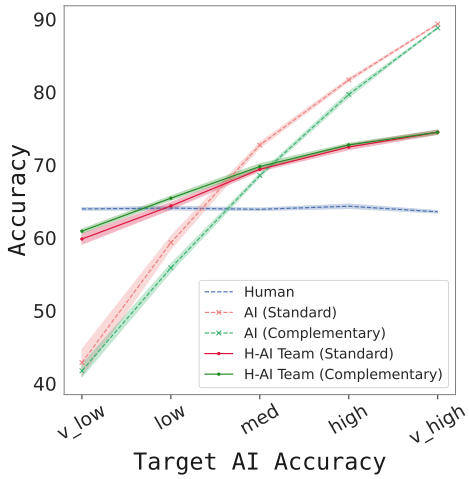
D Additional Results (WoofNette)

We discuss here some additional insights from experiments on WoofNette, the real-world dataset central to our human subject experiments.

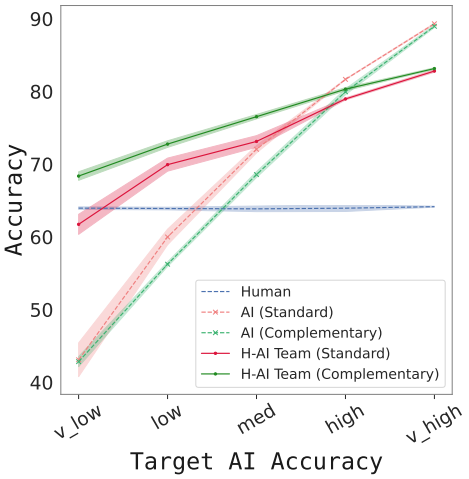
D.1 Evaluating Varied AI Accuracy

In our human-subject experiment, we employ an AI model that is trained after targeting for an accuracy that is comparable to that of independent human judgments. Here, we extend our investigation to understand how the benefits of our proposed human-confidence-based instance weighting training strategy may vary with the target accuracy of the AI model. To explore this, we conduct a simulation study.

We anticipate that our complementary training approach would yield improvements across scenarios, although the magnitude of these gains may differ. For instance, when the AI exhibits exceptionally high accuracy across all instances, there may be limited room for complementarity, resulting in modest improvements through any method, including ours. Nevertheless, even in such scenarios, the proposed strategy should not be detrimental, and should at least maintain team performance. Our evaluation results align well with these expectations.

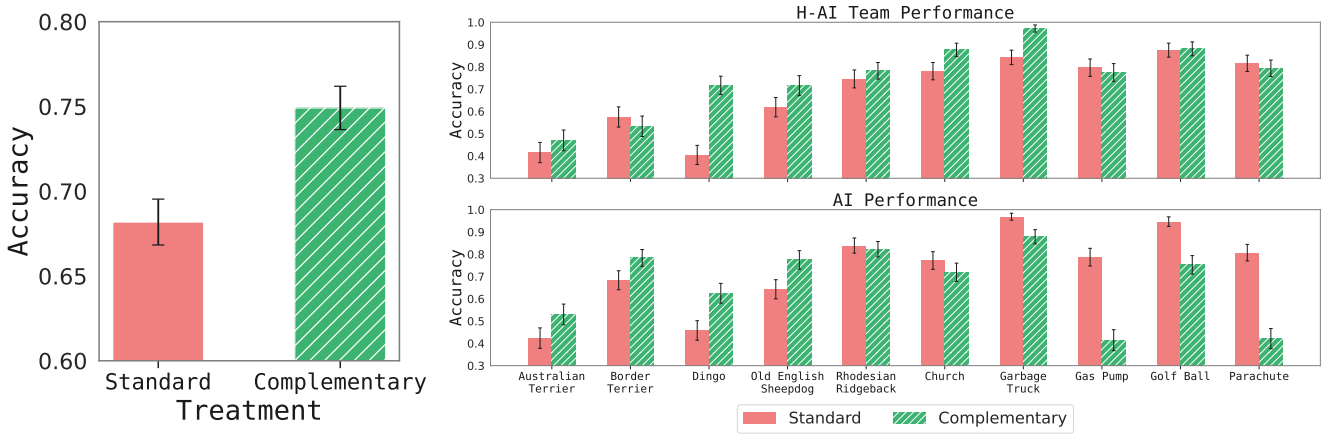


(a) UNIFORM Threshold Distribution



(b) δ Threshold Distribution

Figure A4: Human, AI and Human-AI team performance on WoofNette using standard and complementary AI training strategies (see differences between solid green and red lines).



(a) Overall Human-AI Team Performance

(b) Class-level Performance Breakdown

Figure A5: Comparing the decision accuracy of the human-AI team when human subjects collaborate with the standard (baseline) AI model or the complementary AI model. Error bars represent the standard errors of the mean.

Generating Human DMs' Behavior. For simulating human DMs' decisions on the WoofNette dataset, we utilize the data that we have already collected from our pilot study in which subjects completed image classification tasks on their own; we used this data originally to build the AI model for predicting humans' self-confidence on each image, i.e., $\hat{\mathcal{C}}_i = g(\mathbf{x}_i)$. This data allowed us to estimate the human subjects' own decision accuracy for classifying each class of object. Utilizing these accuracy estimates, 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. We further used $\hat{\mathcal{C}}_i$ as human confidence estimates for each image.

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-confidence-based instance-weighted categorical cross-entropy loss. We again adopt the simple $w_i = 1 - \hat{\mathcal{C}}_i$ weighting scheme. We intentionally restrict the AI's accuracy by training it till a "target accuracy" is reached. In the main text, this target accuracy was set to 65%, which is close to the expected human accuracy on our dataset. In our simulation setup here, we explore the impact of varying AI's target accuracy, ranging from 35% (very low or `v_low`) to 95% (very high or `v_high`) in evenly spaced intervals of 15%, on the observed gains of human-AI team performance in joint decision making.

Evaluation results. To simulate the human-AI team decision on each task, we consider two self-confidence threshold distributions: UNIFORM ($U[0.1, 1]$) and δ (impulse at 0.7). UNIFORM represents the most basic, uninformative scenario. On the other hand, δ may be more representative here as it would lead to two high and low confidence regions, which is what we expect with easy object images and difficult dog images. Figure A4 illustrates significant gains in the decision accuracy of the human-AI team when utilizing our proposed training method for both self-confidence threshold distributions, though the absolute improvement is more pronounced when human self-confidence follows a δ distribution. As expected, the human-AI team performance gains are higher when the target AI accuracy is lower since there is more room for contribution by human teammate (and for dissimilarity between baseline and complementary AI).

D.2 Performance Breakdown

We also conducted an analysis of the performance of our proposed complementary solution at the class level. The results of this breakdown for our human subject experiments are shared in Figure A5.

It is noteworthy that, although the individual performance of the complementary AI may experience significant drops for specific object classes, the overall human-AI team performance throughout remains comparable, if not higher, than AI trained through the standard approach. This observation suggests that even within easily recognizable object classes (or challenging dog classes), human decision-makers may encounter instances that pose difficulty (or are easier) and exhibit low (or high) confidence. Our confidence-based instance weighting strategy enables the complementary AI model to better support human decision-makers in handling these specific instances, resulting in nuanced instance-level support rather than generic class-level gains.