Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment

Supplementary Materials

CHUN-WEI CHIANG, Purdue University, USA
ZHUORAN LU, Purdue University, USA
ZHUOYAN LI, Purdue University, USA
MING YIN, Purdue University, USA

1 RESULTS OF THE PILOT STUDY (WHEN HUMANS DO NOT HAVE AI)

To get some context on humans' decision making behavior and performance in recidivism risk assessment when they do *not* have access to an AI model, we first conducted a pilot study. Specifically, we recruited US-based workers from Amazon Mechanical Turk (MTurk) as our subjects to complete a set of 6 recidivism risk assessment tasks *on their own*—the six defendant profiles we presented to them on these tasks were selected from the same pool of 30 profiles that we prepared for the formal tasks of our Phase 2 HIT. In addition, same as that in the Phase 2 HIT, the 6 defendant profiles each subject saw contained 4 profiles that represented different combinations of defendant race and true recidivism status (i.e., Black reoffending, Black non-reoffending, White reoffending, White non-reoffending) and a pair of "twin" defendant profiles. In each task, the subject made a binary prediction on whether the defendant would reoffend within two years and indicated her confidence in the prediction on a 5-point Likert scale. We also included an attention check question in the pilot study, for which the subject was asked to select the pre-specified option.

A total of 67 workers completed the pilot study and passed the attention check. In the following, we report these subjects' independent decision making behavior and performance in the recidivism risk assessment, with respect to their decision accuracy, confidence, fairness in decision-making, and how often their decisions agree with each other and with the COMPAS algorithm (i.e., the AI model used in our formal study). We adopted the same metrics as those used in our formal experiment to quantify subject's decision making behavior and performance whenever applicable.

Decision Accuracy. In our pilot study, an average subject's decision accuracy on the recidivism risk assessment tasks without AI assistance was 47.5%.

Decision Confidence. For the correct decisions, subjects' average confidence on them was 3.61 (3 represents "neither confident nor unconfident" and 4 represents "confident" in our scale). Meanwhile, subjects' average confidence on their incorrect decisions was 3.62.

Decision Making Fairness. Table 1 reports how fair subjects' decisions in recidivism risk assessment tasks are when they do not receive any decision recommendation from an AI model. Note that since subjects in the pilot study made the recidivism risk assessment *without* the AI assistance, we did not include the metrics reflecting how fair subjects were when they interacted with the AI model (i.e., ΔREL-POS and ΔREL-NEG). To understand if subjects' decisions were substantially "unfair" according to some definition, for each metric, we conducted a two-tailed t-test to exam

Authors' addresses: Chun-Wei Chiang, chiang80@purdue.edu, Purdue University, West Lafayette, Indiana, USA, 47907; Zhuoran Lu, lu800@purdue.edu, Purdue University, West Lafayette, Indiana, USA, 47907; Zhuoyan Li, li4178@purdue.edu, Purdue University, West Lafayette, Indiana, USA, 47907; Ming Yin, mingyin@purdue.edu, Purdue University, West Lafayette, Indiana, USA, 47907.

	ΔPOS	ΔTwin	ΔACC	ΔFPR	ΔFNR
Individual (without AI)	0.045	-0.090	0.035	0.015	-0.127
(adjusted) p-value	0.201	0.135	0.374	0.741	0.075

Table 1. The fairness level of human subjects' decisions in recidivism risk assessment task when they have no access to AI models. Two-tailed t-tests are used to examine if the value of each metric is statistically different from zero, and p-values are reported (Bonferroni corrections are used for ΔFPR and ΔFNR).

whether the value on that metric was significantly different from zero. Our test results suggest that none of the values are reliably different from zero at the level of $p = 0.05$, which implies that without AI assistance, humans' own recidivism decisions are relatively fair and not biased towards any particular demographic groups.

Agreement in subjects' decisions. To get a sense of how often subjects agreed with each other in their recidivism risk assessment, for each of the 30 defendant profiles that we used in the pilot study, we computed the fraction of subjects who agreed with the majority decision—intuitively, the closer the fraction was to 100%, the more subjects agreed with one another in their decisions. We found that the number of profiles with 50-60%, 60-70%, 70-80%, 80-90%, or 90-100% of subjects agreeing with the majority decision was 6, 14, 4, 5, 1, respectively. In other words, for 24 out of the 30 defendant profiles (i.e., 80% of the profiles), at least 20% of the subjects disagreed with the majority decision. This suggests that the level of disagreement between subjects was quite high for most of the recidivism risk assessment tasks.

Subjects' agreement with the COMPAS algorithm. To obtain a better understanding of how frequently subjects' independent decisions may naturally agree with the AI recommendation, we computed the agreement rate between the decisions that subjects in this pilot study made and the COMPAS algorithm's decision recommendation. We found that on average, 42.3% of a subject's independent predictions were the same as the AI model's. This suggests that humans' independent decisions are sufficiently different from the AI recommendations, which implies that humans may be complementary to the AI model. To formally test this, we defined the "crowd's decision" on each defendant's profile as the majority decision made on that profile, and we looked into the accuracy of the crowd and the AI model for each profile. Among the 30 defendant profiles, the crowd and the AI model were both correct on 6 profiles, and both wrong on 5 profiles. In addition, there were 8 profiles where the crowd was correct but the AI model was wrong, and 11 profiles that the AI model was correct but the crowd was wrong. In other words, humans and the AI model indeed exhibit a high level of complementary strengths—if a human-AI team could make correct decisions on all profiles where either the AI model or the crowd was correct, they would achieve an accuracy of 83.3% on these 30 defendant profiles, which was much higher than either the average individual (47.5%) or the AI model (56.7%) alone.

2 EXPLORATORY ANALYSIS ON THE IMPACTS OF COGNITIVE DIVERSITY OF A GROUP

In the following, we present the full set of results for our exploratory analysis on comparing the behavior and performance of HIGH-DIVERSITY groups and LOW-DIVERSITY groups in human-AI collaborative recidivism risk assessment, with respect to their decision accuracy and confidence, appropriateness of reliance on AI, understanding of AI, fairness in decision-making, and willingness to take accountability.

Decision Accuracy. The average decision accuracy was 53.8% for HIGH-DIVERSITY groups and 58.4% for LOW-DIVERSITY groups. A Welch's t-test indicates that the cognitive diversity of groups does not significantly affect their decision accuracy ($p = 0.465$).

Treatment	ΔPOS	ΔTwin	ΔACC	ΔFPR	ΔFNR	ΔREL-POS	ΔREL-NEG
HIGH-DIVERSITY	0.051	-0.231	0.103	0	-0.192	0.250	0.042
LOW-DIVERSITY	0.089	0.038	-0.038	0.139	-0.044	-0.109	-0.025
(adjusted) p-value	0.728	0.042*	0.191	0.614	0.918	0.048*	1.000

Table 2. Comparing the fairness of HIGH-DIVERSITY groups and LOW-DIVERSITY groups in AI-assisted decision making. * represents the statistical significance level of 0.05. Since ΔFPR, ΔFNR, ΔREL-POS, and ΔREL-NEG are computed on subsets of the data, Bonferroni corrections are used and adjusted p-values are reported.

Reliance on AI. Next, we compare how HIGH-DIVERSITY groups and LOW-DIVERSITY groups rely on the AI model differently. Interestingly, we found that LOW-DIVERSITY groups ($M = 0.75, SD = 0.28$) relied on the AI model significantly more than HIGH-DIVERSITY groups ($M = 0.60, SD = 0.27$; $t(91) = 2.16, p = 0.033$). However, we did not find any reliable evidence suggesting that HIGH-DIVERSITY groups and LOW-DIVERSITY groups had different levels of over-reliance (*adjusted-p* $= 0.222$) or under-reliance (*adjusted-p* $= 0.435$) on the AI model.

Decision Confidence. The average decision confidence for the HIGH-DIVERSITY groups ($M = 4.10, SD = 0.91$ on correct decisions; $M = 4.11, SD = 0.94$ on incorrect decisions) appeared to be slightly higher than that of the LOW-DIVERSITY groups ($M = 4.06, SD = 0.83$ on correct decisions; $M = 3.95, SD = 0.93$ on incorrect decisions), regardless of the correctness of the decisions. However, the results of Welch's t-tests with Bonferroni correction suggest that these differences are not statistically significant (*adjusted-p* $= 1.000$ for correct decisions and *adjusted-p* $= 0.355$ for incorrect decisions). We further compare the average decision confidence conditioned on both the decision correctness and the agreement between the decision and the AI recommendation, but we do not find any significant difference at the level of $p = 0.05$ after Bonferroni correction.

Understanding of AI. For subjects in the LOW-DIVERSITY groups, the average Pearson correlation coefficient between the ground truth of the feature importance and the subject's perceived feature importance reported in the exit survey was 0.086. Meanwhile, for subjects in the HIGH-DIVERSITY group, the average Pearson correlation coefficient value is 0.170. Our Welch's t-test suggests that the cognitive diversity of groups does not significantly change people's understanding of the AI model ($p = 0.233$).

Decision Making Fairness. Table 2 compares both how fair the groups' decisions are and how fair the groups interact with the AI model between HIGH-DIVERSITY groups and LOW-DIVERSITY groups. We found that HIGH-DIVERSITY groups and LOW-DIVERSITY groups had significant differences in how they make decisions for the twin cases ($p = 0.042$)—When given a pair of defendants who are identical on all features and their true recidivism status except for their race, the HIGH-DIVERSITY groups tended to believe that the Black defendant in the pair has a much lower risk to reoffend, while the LOW-DIVERSITY groups made roughly similar predictions for the Black defendant and the White defendant. In addition, we also found that in general, when the AI model made a positive prediction, HIGH-DIVERSITY groups were much more likely to follow it when the defendant was Black, which was significantly different from LOW-DIVERSITY groups who were more likely to follow it when the defendant was White ($p = 0.048$).

Accountability. We first compare the accountability that people in HIGH-DIVERSITY and LOW-DIVERSITY groups assign to themselves and the AI model conditioned on whether the final decision is correct or wrong, and we further examine their accountability assignment conditioned on both the correctness of their decisions and the AI recommendation. Overall, we do not find cognitive diversity of groups has any significant difference (at the level of $p = 0.05$ after Bonferroni correction) on their assignment of accountability to either themselves or the AI model.