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

Chun-Wei Chiang
chiang80@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Zhuoran Lu
lu800@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Zhuoyan Li
li4178@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Ming Yin
mingyin@purdue.edu
Purdue University
West Lafayette, Indiana, USA

ABSTRACT

With the prevalence of AI assistance in decision making, a more relevant question to ask than the classical question of "are two heads better than one?" is how groups' behavior and performance in AI-assisted decision making compare with those of individuals'. In this paper, we conduct a case study to compare groups and individuals in human-AI collaborative recidivism risk assessment along six aspects, including decision accuracy and confidence, appropriateness of reliance on AI, understanding of AI, decision-making fairness, and willingness to take accountability. Our results highlight that compared to individuals, groups rely on AI models more regardless of their correctness, but they are more confident when they overturn incorrect AI recommendations. We also find that groups make fairer decisions than individuals according to the accuracy equality criterion, and groups are willing to give AI more credit when they make correct decisions. We conclude by discussing the implications of our work.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Human-AI interaction, Group-AI interaction, AI-assisted decision making

ACM Reference Format:
Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. 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 . In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581015

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581015

1 INTRODUCTION

AI-driven decision aids have been widely used to assist people in making decisions in diverse domains, including criminal justice [74], financial investment [113], medical diagnosis [22, 42], and more. In these AI-assisted decision making settings, AI plays the role of an assistant to provide decision recommendations, while humans can choose to accept or reject these recommendations and make the final decisions. With the increasing prevalence of AI-assisted decision making, a growing line of empirical research has been carried out in the HCI community to understand how the end-users interact with and utilize AI assistance in decision making. These studies have looked into many different aspects of the decision making processes and outcomes, such as whether and when people rely on AI recommendations [26, 56, 125], how accurate people's final decisions are [7, 8, 46, 62], and whether people's final decisions are in line with key societal values [2, 23, 49, 74].

Interestingly, much of the current empirical research on AI-assisted decision making focuses on examining how an *individual* decision maker behaves and performs when assisted by an AI-driven decision aid. However, real-world decision making often involves a *group* of decision makers—decisions on whether a defendant is guilty or not are made by juries, college admissions are decided by committees, and campaign strategies are finalized by a team of strategists. In fact, the folk knowledge that "two heads are better than one" reflects the common belief in collective intelligence and may explain why many decisions are made by groups. Meanwhile, it also inspires decades of research in social science to rigorously compare how individuals and groups behave and perform differently in decision making [1, 5, 16, 63, 69, 78, 89]. In this sense, in a world where AI-assisted decision making may become a primary paradigm of decision making in a foreseeable future, one critical question that needs to be answered is how groups' behavior and performance in *AI-assisted decision making* are different from those of the individuals'. Compared to the case where an individual interacts with an AI-driven decision aid, when there is a group of people, they can not only interact with the AI model but also with

each other, which may result in much more sophisticated interactions. Obtaining a systematic knowledge of the differences between individuals and groups in AI-assisted decision making, thus, can not only advance the scientific understandings in both human-AI interaction and decision making, but also provide insights into the potentially different requirements for designing AI to best support individuals or groups.

Therefore, in this paper, we contribute to the human-AI interaction research by providing an initial comparative case study on how humans utilize AI assistance and make decisions differently in AI-assisted decision making when they make decisions individually or in groups. We start by reviewing the existing empirical research on human-AI interaction to identify the key aspects of people's behavior and performance in AI-assisted decision making that have attracted the most attention within the research community in the past. We summarize from our literature review six aspects of the AI-assisted decision making processes and outcomes that have been commonly studied—decision accuracy, reliance on AI, decision confidence, understanding of AI, fairness in decision making, and willingness to take accountability. Although these aspects are certainly not exhaustive, they provide an initial set of concrete perspectives to compare how groups and individuals differ in AI-assisted decision making.

We then choose recidivism risk assessment as the domain of our case study, because it represents a family of real-world decision making domains where decisions can be made by either individuals or groups (e.g., juries), decision makers may have their own biases when making decisions, and AI-assisted decision making has become increasingly prevalent. Given this domain, we carry out a pre-registered, randomized human-subject experiment ($N = 326$) on Amazon Mechanical Turk, aiming to compare the differences between individuals and groups in AI-assisted recidivism risk assessment along the six aspects we have identified. In our experiment, subjects were asked to complete a series of recidivism risk prediction tasks with the assistance of an AI model (i.e., the original COMPAS algorithm). We created two treatments by varying whether subjects made their AI-assisted decisions on their own or in a group. In particular, when subjects needed to make their decisions in a group, they were asked to use a chatroom embedded in the task interface to discuss the decision making task with other members in their group in order to reach a *consensus* decision.

Our results suggest that in AI-assisted recidivism risk assessment, people who make decisions individually and those who make decisions in groups do not exhibit significant differences in their decision accuracy or their understandings of the AI model. However, we also find a few important distinctions between groups and individuals in AI-assisted decision making through our experiment—compared to individuals, groups are more likely to rely on the AI model's recommendations regardless of their correctness, but they also have higher confidence when they overturn the AI model's incorrect recommendations, appear to make fairer decisions according to the accuracy equality criterion, and are willing to give more credits to the AI model when they make correct decisions with the assistance of the model's correct recommendations. As many previous studies in group research have highlighted the importance of group dynamics and compositions on influencing collective behavior and performance [11, 35, 44, 67, 77, 79, 123], we further conduct a few exploratory analyses to gain more insights into the interactions between groups and AI in human-AI collaborative recidivism risk assessment. First, by analyzing the chat logs of groups in our experiment, we identify a few representative ways for the AI model to influence group dynamics, including serving as the reference point for people's initial decisions and the tiebreaker for people to reach a consensus when they hold conflicting opinions. In addition, by categorizing groups into homogeneous groups (i.e., low diversity) and heterogeneous groups (i.e., high diversity) based on the composition of group members' cognitive styles, we find that groups with a higher level of cognitive diversity significantly decrease their reliance on AI than groups with low diversity, which, however, does not appear to result in an improved level of decision making accuracy or fairness.

Together, our study provides important experimental evidence that both the process and the outcome of people's interactions with AI-driven decision aids can be affected by whether the decisions are made individually or collectively. We conclude by providing the design implications and limitations of our study, as well as discussing the future directions in group-AI interaction research.

2 LITERATURE REVIEW

2.1 Decision Making in Groups

Group decision making refers to the scenarios where a group of decision makers make a decision collectively. Group decision making is often not a simple aggregation of individual's decisions. It involves many other important aspects such as collaborations within the group [75], communication between group members [52, 108], leadership [76, 116, 122], and more. For example, collaborations within a group of people during the decision making may bring about many benefits, such as reducing the impacts of individuals' biases and blind spots on the decision making outcome [10, 110] and generating synergy between group members so that the group can produce better collective decision making performance than individuals acting on their own [75]. However, it is found that groups sometimes can also perform poorly in decision making due to phenomena like groupthink [60] and polarization inside the group [95]. Previous empirical studies in management and psychology have shown that the performance of a group in decision making may depend on many different factors, such as the intra-group trust [32, 41, 86] and the levels of understanding between group members [57, 59]. Another critical aspect that will influence the decision making performance of a group is its composition. A wide range of predictors of collective intelligence have been identified, including the group's average level of skill [35], social perceptiveness [123], demographic diversity [55], and cognitive diversity [54, 55, 61, 85, 97, 97, 107].

In addition, the question of whether groups outperform individuals in decision making has attracted great interests within the research community, although the results are largely mixed—Some studies find dramatic advantages of groups over even the best-performing individuals in the groups [92, 110], suggesting a "process gain" from group interactions, while other studies report an opposite phenomenon of "process loss" [65]. The comparisons between the decision making performance of the individuals and groups are also found to be moderated by many factors, including task complexity [1], the performance gap between individuals

in the group [5], and the degree to which individual's subjective confidence can reflect their accuracy [69].

2.2 Human-AI Interaction: A Brief Review of Commonly Studied Aspects

Research in human-AI interaction is surging in recent years. The interaction structure between humans and AI can take many different formats, such as having humans and AI act as teammates (i.e., they each work on interdependent tasks but share a common goal) [30, 34, 90, 94, 130], or allow the AI to take leadership in a human-AI team [122]. In this paper, we focus on the AI-assisted decision making paradigm as our human-AI interaction structure [20, 115]—In this paradigm, the AI agent plays the role of an assistant to provide decision recommendations to humans, while humans make the final decisions. Over the past few years, a growing line of experimental research have been carried out to empirically understand how humans (mostly individuals) interact and work with AI assistance to make decisions [70]. By reviewing this literature, we highlight a set of six commonly studied aspects of AI-assisted decision making processes and outcomes as follows.

Aspect 1: Decision Accuracy. Early empirical studies in AI-assisted decision making often investigate into the accuracy of humans' final decisions, and aim to leverage the complementary strength of humans and AI to improve the decision accuracy of the human-AI team [7, 46, 62]. While researchers find that the usage of AI assistance often helps people increase their decision accuracy [50], it is also noted that enabling the human-AI team to achieve a decision making accuracy higher than either party alone is generally challenging [8]. In addition, researchers have also focused on studying people's decision accuracy in AI-assisted decision making when the AI recommendations are wrong [99], and it is shown that people may lack the capability to accurately evaluate the accuracy of AI recommendations; this often results in people's poor decision accuracy when the AI recommendations are wrong [26, 50, 72]. It is also suggested that humans' decision accuracy in AI-assisted decision making can largely depend on the complexity of the AI model's error boundary [6], as well as whether changes in the AI model's error boundary caused by model updates are compatible with humans' mental models about the AI [7].

Aspect 2: Reliance on AI. How accurate humans' decisions are in AI-assisted decision making is largely influenced by the ways that people decide to rely on the AI model [6, 8]. Researchers adopt different methods to measure people's willingness to rely on AI models, including asking people to report their confidence in the AI [8, 13, 20, 36, 64, 66, 101, 112, 125, 132], and computing the actual frequency that people request assistance from the AI model or adopt the AI model's recommendation [4, 6, 20, 26, 82, 84, 101, 104, 109, 125–128]. For regression tasks, people's reliance on AI can be calculated as the weight they assign to the AI recommendation when determining their final decisions [27, 56, 83, 99]. Interestingly, previous research identifies two opposite behavior patterns regarding how much humans are willing to rely on AI assistance in decision making—algorithm aversion [13, 36] and algorithm appreciation [83]. Recent research shows that humans' reliance on AI models can be influenced by a wide range of factors, such as the AI model's accuracy [101, 127, 128], the expert power of the AI model [56], the level of agreement between the AI model's predictions or rationale and their owns [84, 132], the first impression of AI model [112], the timing and type of errors made by the AI model [36, 66, 104], and more. Most recently, researchers have started to separate different types of reliance, and look into people's over-reliance, under-reliance, and appropriate reliance on AI to understand whether people rely on AI models in an appropriate way, and how to promote more appropriate reliance [26, 36, 109, 120, 125, 126].

Aspect 3: Decision Confidence. Three kinds of confidence are often measured in AI-assisted decision making studies: people's confidence in the AI model, their self-confidence [29], and their confidence in their own final decisions [83]. As discussed earlier, people's confidence in the AI model is often used as a metric to approximate their trust in the AI model or their willingness to rely on it. Self-confidence represents how much people believe in their own ability on the decision making task [29], and recent research shows that people' self-confidence impacts their intention to use the AI assistance as well as their ability to appropriately utilize the AI assistance [28, 29, 48, 83, 119]. From the perspective of evaluating the decision making outcome in AI-assisted decision making, people's confidence in their *final* decisions is the most relevant measure. However, we note that when people decide to reject the AI model's decision recommendation and make the final decision themselves, their confidence in their final decision largely reflect their self-confidence. Since it is challenging to observe one's confidence externally, the main method adopted to measure people's confidence is asking people to directly report their confidence on a numeric scale [29, 83, 119, 125].

Aspect 4: Understanding of AI. Researchers have advocated for the needs of increasing people's understanding of AI models to promote their appropriate reliance on the models, and eventually improve people's decision accuracy in AI-assisted decision making [3, 71, 72, 102]. To evaluate the effectiveness of various explainable AI methods, especially on improving people's understandings of an AI model, a large number of experimental studies have been carried out [8, 19, 24, 25, 71, 82, 99, 120, 121, 125, 131]. In these studies, people's understanding of the AI model can be evaluated in both subjective and objective ways. Subjective understanding can be measured by asking people to report their perceived understanding of the AI model [14, 25, 27, 84]. Objective understanding can be evaluated by asking people to simulate the AI model's prediction [81, 99, 120], examining people's capability in recognizing the error of the AI [99, 120], having people describe the cause of an AI model's prediction [14, 80, 100], or having people analyze the importance and contributions of different features to an AI model's prediction [25, 53, 120].

Aspect 5: Fairness in decision making. Recent research in fairness in AI has highlighted that many AI models underlying decision aids may have a tendency to amplify the existing societal biases and discrimination, and make unfair decisions on the underrepresented population [2, 23, 31, 74, 93]. In light of this, researchers have started to look into the implications of humans' collaboration with an AI model on the fairness in their decision making. While early studies look into people's perceptions of AI model's fairness [37, 47, 118], most recently, studies have carried out to examine the fairness of the decisions that humans make in AI-assisted

decision making [39], as well as whether humans interact with the AI model in a biased way [49, 50].

Aspect 6: Accountability. With the occurrence of catastrophic failure of AI systems [114, 117], researchers have emphasized on the urging needs to rethink the relationship and accountability between humans and AIs in human-AI collaborations [73, 111]. It is argued that increasing people's responsibility to their final decision have the potential to reduce the decision bias [98, 114] and increase the social justice [21]. Despite its clear importance, empirical research on understanding how accountability is assigned among different parties in decision making is relatively few, and self-reports are again used to solicit people's accountability perceptions [105].

3 STUDY DESIGN

As a case study, we conduct a pre-registered[1], randomized human-subject experiment on Amazon Mechanical Turk (MTurk) to compare the behavior and performance of groups and individuals in AI-assisted decision making along the six aspects that we have identified in Section 2, in the domain of recidivism risk assessment.

3.1 Experimental Task

We used the recidivism risk assessment task in our experiment. Specifically, in each task, subjects were presented with the profile of a criminal defendant containing 8 features, including their basic demographics (e.g., gender, age, race), criminal history (e.g., the count of prior non-juvenile crimes, juvenile misdemeanor crimes, juvenile felony crimes committed), and information related to their current charge (e.g., charge issue, charge degree). Subjects were also given an AI model's prediction on whether this defendant would reoffend in the next two years. After reviewing all this information, subjects needed to make their decisions regarding whether they believed the defendant would reoffend. The defendant profiles we presented to subjects were selected from the COMPAS dataset, a public dataset containing the profiles of defendants from Broward County, Florida, between 2013 and 2014, and we used the version shared by Dressel and Farid [38]. In addition to defendant profiles, for each profile, the COMPAS dataset also contains a recidivism risk score on a scale from 1 to 10, which is given by the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm, a commercial decision support tool to assess the likelihood of a defendant becoming a recidivist. In Dressel and Farid [38], defendants with a recidivism risk score of 5 or above are considered to have a high risk of reoffending. Correspondingly, in our experiment, the AI model's binary prediction that subjects saw on each task was directly taken from the COMPAS algorithm, and defendants with a risk score of 5 or above would be predicted as "will reoffend" in the next two years (i.e., the AI model we used in our experiment was the COMPAS algorithm).

We chose to use the recidivism risk assessment task in our case study for several reasons. First, utilizing collective intelligence in criminal justice is not uncommon (e.g., whether a defendant is guilty is often decided by juries), which makes recidivism risk assessment a realistic domain where decision making can be carried out both by groups and by individuals. Second, AI-assisted decision making has become increasingly prevalent in criminal justice. In fact, the COMPAS algorithm—the precise AI model for recidivism risk assessment that we used in our experiment—has been previously used by many states in the U.S., including Florida, New York, and California [17, 43, 68]. Another critical reason is that in the recidivism risk assessment task, by comparing people's behavior and performance when making decisions for defendants of different races, we can investigate decision makers' "fairness" in AI-assisted decision making both in terms of biases in their decisions and biases in the ways that they interact with the AI model. In particular, previous research has reported that despite its relatively high accuracy [18], the COMPAS algorithm is unfair itself as it tends to incorrectly assign higher recidivism risk scores to black defendants [74]. Conducting our experiment on the recidivism risk assessment task with the COMPAS algorithm, thus, allows us to compare whether and how groups and individuals utilize an unfair AI model differently.

[1]The pre-registration document can be found at https://aspredicted.org/bd2jj.pdf.

3.2 Experimental Treatment

Depending on whether subjects were asked to complete the AI-assisted recidivism risk assessment tasks on their own or in a group, we created two treatments in our experiment:

- **Individual-AI collaboration treatment (Individual-AI)**: Subjects completed the recidivism risk assessment task without any interactions with other people. In particular, on each task, the subject reviewed the defendant profile and the risk prediction produced by the COMPAS algorithm. Then, the subject predicted whether the defendant would reoffend within the next two years.
- **Group-AI collaboration treatment (Group-AI)**: Subjects completed the recidivism risk assessment task within a group of 3 people. Figure 1 shows an example of the task interface for subjects in this treatment. In particular, on each task, all members in the group saw the same defendant profile and the same risk prediction produced by the COMPAS algorithm (Figure 1A). Subjects were asked to predict whether the defendant would reoffend within the next two years, and they needed to utilize the chatroom function embedded in the task interface to discuss the case with other group members in order to reach a consensus prediction (Figure 1B). Once all subjects in the group selected the same prediction, they were prompted to confirm it as their final decision.

3.3 Experimental Procedure

Our experiment consisted of two phases conducted on different days: Phase 1 was the recruiting phase, and Phase 2 was the real experiment phase. Figure 2 shows the flow of our experiment.

Phase 1. To implement the "Group-AI" treatment in our experiment, we need to coordinate the experiment participation time for a large number of subjects. Thus, following the best practice for conducting synchronous experiments on MTurk [1, 87, 88, 103, 133], we used Phase 1 of our experiment to recruit a large panel of potential human subjects for our experiment. Specifically, in Phase 1, we posted a recruitment HIT on MTurk for subjects to sign up for our experiment. In this HIT, subjects first needed to fill out a demographic survey. Then, subjects were asked to complete the same set of 9 recidivism risk assessment tasks (in a randomized order) *without* the assistance of the AI model (i.e., the COMPAS

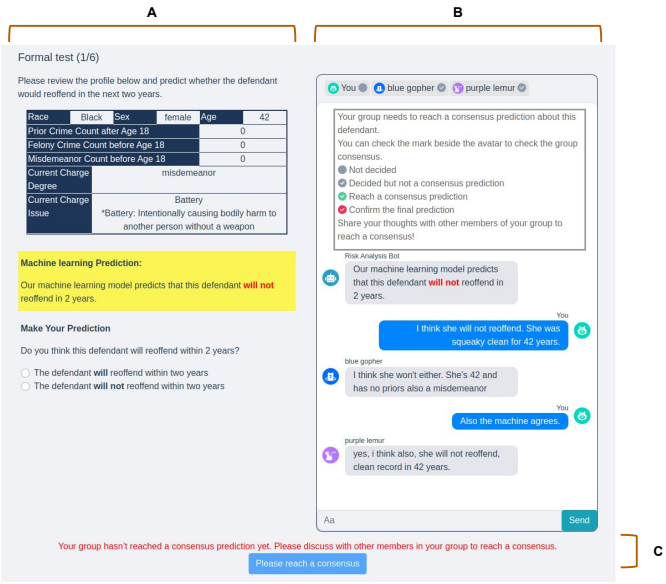


Figure 1: An example of the formal task interface for the Group-AI treatment. Subjects in the Individual-AI treatment could only see Part A. A: Subjects were presented with a defendant profile as well as the COMPAS algorithm's recidivism prediction. Based on this information, subjects were asked to make a binary prediction on whether the defendant would reoffend. B: Subjects of the Group-AI treatment could discuss the task with other members in their group via the chatroom. C: Subjects of the Group-AI treatment were required to reach a consensus prediction within the group on a task before moving on to the next task.

algorithm). On each task, we presented to subjects a criminal defendant's profile, and subjects were required to predict whether the defendant would reoffend within the next two years on their own. No feedback was provided to subjects on the correctness of their predictions. At the end of the HIT, subjects were asked to indicate in an exit survey if they would like to be notified of future experiment sessions of similar tasks (i.e., our Phase 2 HIT).

The purpose of including the recidivism risk assessment tasks in our Phase 1 HIT is three-fold. Firstly, these tasks enabled subjects to establish an expectation of what they would work on should they decide to participate in our future experiment sessions, so they could make an informed choice in the exit survey. Secondly, one of the 9 recidivism risk assessment tasks was an attention check task in which subjects were instructed to select a pre-specified option. To ensure the quality of our experimental data, we did not invite subjects who failed to pass the attention check in Phase 1 to participate in Phase 2 of our experiment. Finally, the set of recidivism risk assessment tasks we chose for Phase 1 HIT also allowed us to quantify each subject's own "*cognitive style*" in making recidivism risk predictions. In particular, in this study, a subject's cognitive style was characterized by the extent to which their recidivism predictions are influenced by the defendants' race (i.e., does the subject believe that a White defendant has higher recidivism risk than a Black defendant?) and charge degree (i.e., does the subject believe that a felony defendant has higher recidivism risk than a misdemeanor defendant?)[2]. Therefore, the 8 defendant profiles shown in all but the attention check task were carefully selected so that they were balanced on the defendant's race (Black or White), charge degree (felony or misdemeanor), and the true recidivism status (reoffend or not reoffend), while they had almost identical values on all other features in the profile. For more details on how a subject's cognitive style is computed using their recidivism predictions in Phase 1, and how we use this information to explore the impact of a group's cognitive diversity on the behavior and performance of the group, see our exploratory analysis in Section 4.7.2.

Phase 2. To formally evaluate the behavior and performance of individuals and groups in human-AI collaborative recidivism risk assessment. we posted our Phase 2 HIT a few days after Phase 1 was completed. For all Phase 1 subjects who passed the attention check and indicated that they'd like to receive notifications from us, we informed them about our Phase 2 HIT before it was posted online. In the Phase 2 HIT, subjects were asked to complete a total of 15 AI-assisted recidivism risk assessment tasks, which were composed of 9 practice tasks and 6 formal tasks. The purpose of the practice tasks was to enable subjects to obtain some understanding of how well the AI model performs on different cases, so that they could decide their best strategies for utilizing the assistance of the AI model in the later formal tasks. On each of the 9 practice tasks, a defendant profile was displayed along with the COMPAS algorithm's binary prediction on whether the defendant would reoffend within two years. Then, the subject was asked to make their predictions, before we revealed to them whether the defendant actually reoffended in reality. Among the 9 practice tasks, one was an attention check task in which subjects were asked to choose a pre-specified option. The other 8 tasks contained 8 defendant profiles that were balanced on the defendant's race (i.e., 4 Black defendants and 4 White defendants). The COMPAS algorithm's accuracy on these 8 profiles was 62.5%, which was close to its accuracy on the entire COMPAS dataset. Moreover, while we kept the fraction of defendants who actually reoffended to be the same within the 4 Black defendants and 4 White defendants in the practice tasks, the COMPAS algorithm's predictions on them showed a higher false positive rate on Black defendants, which is consistent with the COMPAS algorithm's behavior on the entire COMPAS dataset as reported in previous studies [74][3].

After the subject completed all 9 practice tasks, we presented a feedback page to them, which summarized the prediction accuracy of the COMPAS algorithm and themselves on each practice task. Then, the subject moved on to work on the 6 formal tasks, for which they would *not* receive immediate accuracy feedback after the task anymore. In particular, subjects would be randomly assigned to one of the two treatments, Individual-AI or Group-AI. If the

[2]We acknowledge that subjects' cognitive style can be defined with respect to how their recidivism predictions are influenced by values of other features in the defendant's profile. We chose to focus on the impact of defendants' race and charge degree because they are binary features, which makes it easier for us to enumerate all possible value combinations on these two features in our Phase 1 tasks and to see how subjects' predictions vary with values on these two features. Subject's belief on how the defendant's race impacts the recidivism risk also reflect their biases in these decisions.
[3]Throughout this paper, we consider predicting a defendant as "will reoffend" to be the "positive" prediction.

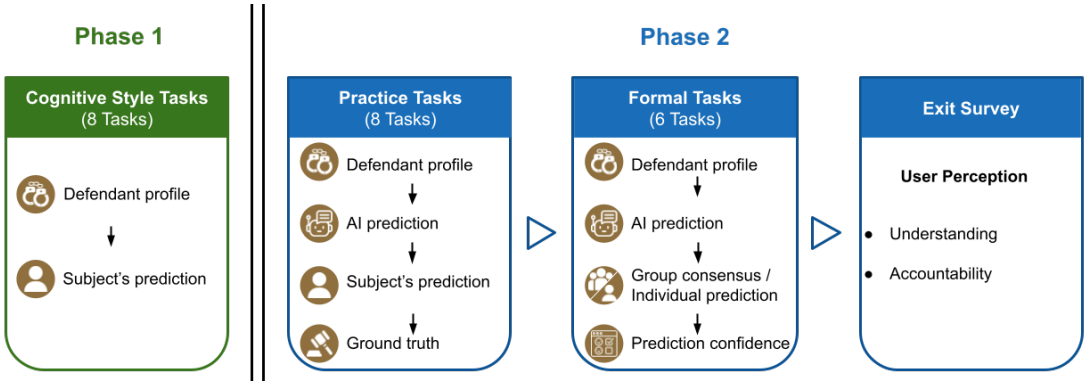


Figure 2: The overall flow of our experiment. Phase 1 was the recruiting phase, and Phase 2 was the real experiment phase. The HIT for Phase 2 was posted a few days after Phase 1.

subject was assigned to the Individual-AI treatment, they would complete each of the 6 formal tasks on their own as we've described in Section 3.2. On the other hand, if the subject was assigned to the Group-AI treatment, they would be sent to a lobby waiting for the other two members of their group to join, before starting working on the 6 formal tasks with them as a group. To protect subjects' anonymity, we asked subjects to pick an animal avatar to represent themselves throughout the rest of the experiment. If the subject waited for more than 5 minutes in the lobby, we would automatically redirect them to the Individual-AI treatment, so that they can complete the 6 formal tasks independently. For those subjects who successfully formed groups, as discussed in Section 3.2, on each formal task, they were asked to discuss the defendant of that task within their group to reach a consensus prediction[4]. Finally, for subjects in both treatments, after submitting their recidivism prediction on each formal task, they were asked to report their confidence in their final prediction (or their group's final prediction) on a 5-point Likert scale from 1 (not confident at all) to 5 (extremely confident); note that for subjects in the Group-AI treatment, each group member reported their confidence in the group's final prediction separately.

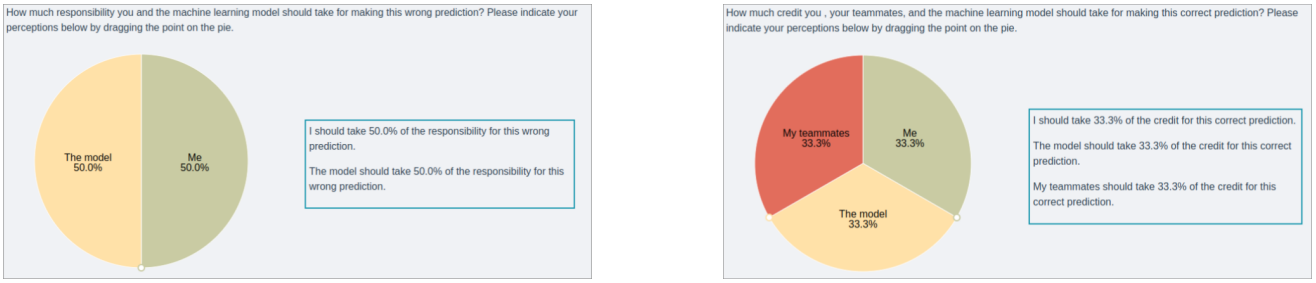
To facilitate our later comparisons on how fair the decisions of groups and individuals are in human-AI collaborative recidivism risk assessment, the defendant profiles for the 6 formal tasks were carefully selected. In particular, we first prepared 4 subsets of defendant profiles for different combinations of defendant race and true recidivism status (i.e., Black reoffending, Black non-reoffending, White reoffending, White non-reoffending) from the COMPAS dataset, with each subset containing 5 profiles. In addition, we also prepared 5 pairs of "twin" defendant profiles such that the two profiles in each pair had exactly the same features and true recidivism status except for the race. To compose the 6 formal tasks for a subject or a group of subjects, we randomly selected one profile from each of the 4 subsets, and we also randomly picked one pair of twin defendant profiles. In doing so, the 6 formal tasks that subjects worked on were balanced on defendant race, and the fraction of reoffending defendants were kept the same among the Black and White defendants. Note that across the entire pool of 30 profiles that we prepared (i.e., 20 from the 4 subsets and 10 from the 5 twin pairs), the COMPAS algorithm's accuracy was 56.7%, with a higher false positive rate on Black defendants (Black: 44.4% vs. White: 11.1%). Moreover, within the pool of 10 twin profiles, the COMPAS algorithm's accuracy was 60%, but it still had a higher false positive rate on Black defendants (Black: 50% vs. White: 25%)[5].

Finally, after completing all the formal tasks, the subject was asked to fill out an exit survey individually. In the survey, we first tested subject's understanding of the AI model by asking them to report their perceived importance of a feature in influencing the AI model's recidivism prediction on a 5-point Likert scale from 1 ("not influential at all") to 5 ("extremely influential"), for each of the 8 features. Then, for each of the 6 formal tasks that they (or their group) completed, we conducted a review with the subject—For each task, we first displayed the defendant's profile, the COMPAS algorithm's prediction, as well as the final prediction of the subject (or the subject's group) to the subject again, before we revealed to them the true recidivism status of the defendant. Then, depending on the correctness of the subject's final prediction, we asked the subject to determine how much *credit* each party should take for the correct prediction or how much *responsibility* each party should take for the incorrect prediction, for each party involved in the decision making—the subject themselves, the AI model, and the subject's teammates (only applicable for subjects in the Group-AI treatment). Subjects allocated the credit or responsibility via interacting with a piechart as shown in Figure 3.

We opened our experiment only to U.S. workers, and each worker was allowed to participate at most once. The base payment was $0.3 for Phase 1 and $1.0 for Phase 2. In addition, to motivate subjects to carefully deliberate (and discuss with other members in their group if applicable) about what predictions to make in the formal task, we further informed each subject at the beginning of the Phase 2 HIT that they could earn a $0.4 bonus for each correct final prediction made on the formal task. Thus, the maximum amount of bonuses a subject could receive in Phase 2 was $2.4.

[4]To ensure that subjects in the Group-AI treatment would actively engage in discussions, we sent a prompt message to subjects if they were idle on the interface for more than 1 minute; if they did not take any actions (e.g., enter chat messages, make a prediction) for more than 2 minutes, they would be removed from the group.

[5]To get a sense of subjects' own decision making behavior and performance on the selected 30 defendant profiles, we conducted a pilot study in which subjects were asked to make independent predictions on these defendant profiles *without* the assistance of the AI model. The results of this pilot study are reported in the supplemental materials.



(a) Individual-AI treatments **(b) Group-AI treatments**

Figure 3: Subjects allocated credit or responsibility for each party involved in the decision making on a pie chart after the correctness of the final prediction was revealed.

3.4 Measurements

To conduct a comprehensive comparison between groups and individuals in AI-assisted decision making, we defined the following measurements to quantify the six aspects of behavior or performance commonly studied in AI-assisted decision making.

3.4.1 Decision accuracy. We measured the decision accuracy of an individual (or a group) as the fraction of formal tasks in which the individual's (or the group's) prediction was correct. Naturally, the higher the decision accuracy, the better the individual (or the group) performs.

3.4.2 Reliance on AI. As the primary metric of reliance, for each individual or group, we measured their *overall reliance* on the AI model using the fraction of formal tasks in which their final prediction was the same as the AI model's prediction, following similar methods used in previous studies on human reliance on AI [84, 132]. In addition, to determine if the individual or the group's reliance on the AI model was appropriate, we also considered two secondary metrics of reliance—subjects' over-reliance and under-reliance on the AI model [20, 120, 125]. Over-reliance was quantified by the fraction of formal tasks where the individual or the group's final prediction was the same as the AI model's prediction, among all formal tasks where the AI model's prediction was wrong. On the other hand, under-reliance was computed as the fraction of formal tasks where the individual or the group's final prediction was different from the AI model's prediction, among all formal tasks where the AI model's prediction was correct. For individuals or groups to perform better, they should lower both their over-reliance and under-reliance on the AI model.

3.4.3 Decision confidence. Subjects' confidence in their decision was measured by their self-reported confidence provided at the end of each formal task. We looked into subjects' decision confidence for their correct final decisions and incorrect final decisions separately. Since confidence reflects subjects' subjective belief in how likely their predictions are correct, it is most desirable if subject's confidence is *calibrated*, i.e., accurately reflects the correctness likelihood of the predictions. In other words, for correct decisions, the more confident subjects are, the better; for incorrect decisions, the less confident subjects are, the better.

3.4.4 Understandings of AI. To quantify how much a subject understood the AI model, we measured how well the subject could recognize the strength of the relationships between different features and the AI model's predictions (i.e., the "importance" of different features). Specifically, the true "importance score" of a feature (e.g., the defendant's age) was calculated as the *absolute value* of the Pearson correlation coefficient between the feature's value and the AI model's predictions—the larger the score for a feature, the more the feature relates to the AI model's predictions[6]. Then, a subject's understanding of the AI model was computed as the Pearson correlation coefficient between the true importance scores and the subject's perceived importance of all features, where the latter was taken from the subject's self-reports in the exit survey. Intuitively, the larger the correlation, the more the subject understood which features in the defendant's profiles had a stronger influence on the AI model's predictions.

3.4.5 Fairness in decision making. We considered two types of fairness in AI-assisted decision making in this study. First, we examined that in the formal tasks, whether individuals' or groups' decisions themselves were fair based on the following metrics [9]:

- **Positive prediction difference (ΔPOS)**: The likelihood of making positive predictions (i.e., predicting "will reoffend") on all Black defendants minus that on all White defendants.
- **Twin case prediction difference (ΔTwin)**: Within the pair of twin defendant profiles, the individual or the group's binary prediction made on the Black defendant in the pair minus that made on the White defendant in the pair.
- **Accuracy difference (ΔACC)**: The prediction accuracy on all Black defendants minus that on all White defendants.
- **False positive rate difference (ΔFPR)**: The prediction's false positive rate (FPR) on all Black defendants minus that on all White defendants.
- **False negative rate difference (ΔFNR)**: The prediction's false negative rate (FNR) on all Black defendants minus that on all White defendants.

For all these metrics, the closer the values are to zero, the fairer the decisions are. We considered ΔPOS, ΔTwin, and ΔACC as the three primary metrics for the fairness level of subjects' decisions, since each one of them directly maps to a unique type of classical fairness definition used in the literature. For example, when an individual or a group's decisions have ΔPOS = 0, their decisions satisfy the fairness definition of *demographic parity* [40, 91]. Decisions with

[6]By computing the *absolute* value of the correlation coefficient between the feature and the AI model's predictions, we quantified the strength of each feature's influence on the AI model's predictions without differentiating the sign of the influence.

ΔTwin = 0 satisfy the *individual fairness* definition [39, 40], that is, treating similar individuals similarly. Decisions with ΔACC = 0 satisfy the fairness definition of *accuracy equality* [12]. In addition, ΔFPR and ΔFNR were considered as the secondary metrics of the decision fairness which examine the accuracy equality across defendants of different races, conditioned on the ground truth is negative or positive respectively; when both ΔFPR = 0 and ΔFNR = 0, the fairness definition of *equalized odds* [51, 129] is satisfied.

Beyond the fairness of the decisions, we were also interested in the fairness in how individuals or groups interacted with the AI model, i.e., whether subjects exhibited any tendency of "*disparate interactions*" [49, 50]. Accordingly, we defined the following metrics based on the individual or the group's reliance on the AI model on the 6 formal tasks:

- **Positive prediction reliance difference (ΔREL-POS)**: The individual or the group's overall reliance on the AI model for all Black defendants where the AI model made a positive prediction minus that for all White defendants where the AI model made a positive prediction.
- **Negative prediction reliance difference (ΔREL-NEG)**: The individual or the group's overall reliance on the AI model for all Black defendants where the AI model made a negative prediction minus that for all White defendants where the AI model made a negative prediction.

Again, the closer ΔREL-POS and ΔREL-NEG are to zero, the less the ways how an individual or a group interacts with the AI model is changed because of the defendant's race, and the fairer the interactions.

3.4.6 Accountability. The accountability of each party involved in the decision making was measured through the subject's self-reported accountability assignment for each task in the exit survey. We looked into subjects' credit assignments for their correct final decisions and responsibility assignment for their incorrect final decisions separately. Recall that subjects in the Individual-AI treatment only assigned the accountability between the AI model and themselves, while subjects in the Group-AI treatment needed to assign the accountability among the AI model, their teammate, and themselves. Thus, to make the accountability comparable between the two treatments, we computed the *normalized accountability* of each party, which is the raw accountability percentage assigned to the party by the subject, divided by the amount of accountability that would have been assigned to the party if it was equally shared among all parties in the decision making (i.e., 50% for each party in the Individual-AI treatment, 33.3% for each party in the Group-AI treatment). In this way, for a specific party, a normalized accountability value larger than 1 (or smaller than 1) means that the subject believed that that party needed to take more than (or less than) an equal share of the accountability. While there is no right or wrong assignment of accountability, ideally, one would hope that when people's final decision is wrong in AI-assisted decision making, they will not shift most of the blame to the AI model [106].

4 RESULTS

1444 workers from Amazon Mechanical Turk (MTurk) took our Phase 1 recruitment HIT and passed the attention check. Among them, 326 subjects participated in Phase 2 of our experiment (45% self-identified as female, and the majority age group was 25–34). By the end of our experiment, we collected valid experimental data from 93 individual subjects for the Individual-AI treatment and 233 subjects (92 groups) for the Group-AI treatment[7]. As a sanity check, we found no significant difference in subjects' decision accuracy or their reliance on the AI model in the 9 practice tasks between subjects of the two treatments, which suggests our randomization of subjects was successful.

In the following, we report our comparisons on the behavior and performance of groups and individuals 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. As an exploratory analysis, in the end, we also look into how the presence of the AI model impacts group dynamics, and how the behavior and performance of a group in AI-assisted decision making vary with the cognitive diversity of the group. Unless otherwise specified, we use two-tailed Welch's t-tests (an adaptation of the Student's t-tests for samples with unequal variances) [33] to examine whether the differences observed between individuals and groups (or different types of groups) are significant. For some metrics, computing their values requires us to divide the entire dataset into several subsets (e.g., over-reliance and under-reliance, the two secondary metrics of reliance, are measured based on two disjoint subsets of the data after conditioning on the correctness of the AI model's predictions). In this case, we consider the comparisons between the two treatments in different subsets to belong to the same family, and we use Bonferroni corrections to correct the p-values for multiple comparisons. For clarity, we use *adjusted-p* in the following to indicate the corrected p-values whenever Bonferroni corrections are used.

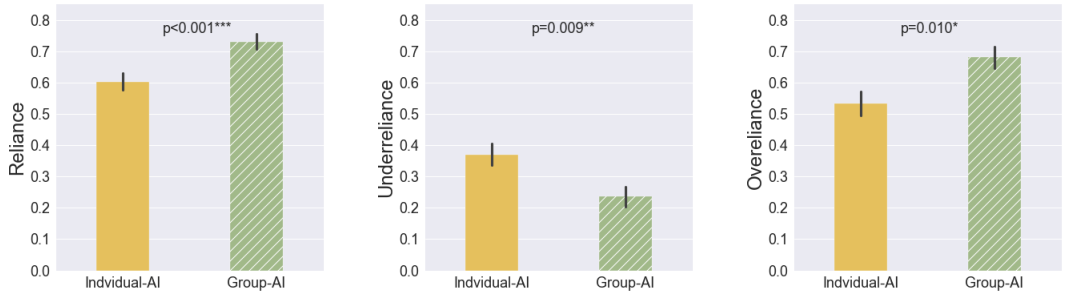
4.1 Comparison on Decision Accuracy

The average decision accuracy for groups in the Group-AI treatment and individuals in the Individual-AI treatment was 57.8% and 55.3%, respectively. The Welch's t-test result suggests that the difference in the decision accuracy between groups and individuals is not statistically significant ($t(184) = -0.75, p = 0.452$).

4.2 Comparison on Reliance on AI

We now move on to examine whether groups and individuals rely on the AI model differently in AI-assisted decision making. Figure 4a compares the overall reliance on the AI model between groups in the Group-AI treatment and individuals in the Individual-AI treatment. Visually, it is clear that when subjects made decisions in groups, they were more likely to rely on the AI model than when they made decisions alone. The Welch's t-test result confirms that the difference is statistically significant (Individual-AI: $M = 0.60, SD = 0.26$; Group-AI: $M = 0.73, SD = 0.24$; $t(184) = -3.47, p < 0.001$).

[7] For subjects of the Individual-AI treatment, 71 were initially assigned to this treatment while 22 were transferred from the Group-AI treatment as they did not successfully form a group. Moreover, among the 233 subjects of the Group-AI treatment, 92 groups were formed, with 49 of them containing three subjects and 43 of them containing two subjects—For each subject assigned to the Group-AI treatment, we made the attempt to form a 3-person group. However, after the 3-person group was formed, some members might drop out of the experiment or be removed from the group due to their inactivity, resulting in a number of 2-person groups.



(a) Overall reliance **(b) Under-reliance** **(c) Over-reliance**

Figure 4: Comparing individuals and groups' overall reliance, under-reliance, and over-reliance on the AI model. Error bars represent the standard errors of the mean. *, ** and * represent the statistical significance level of 0.05, 0.01 and 0.001, respectively.**

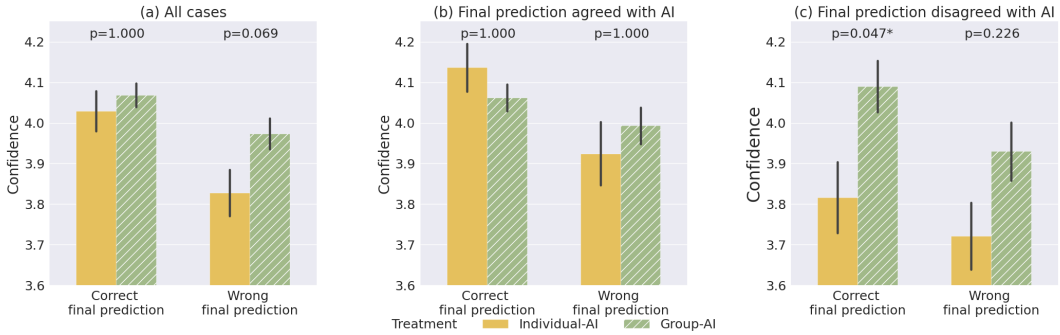


Figure 5: Comparing individuals' and groups' confidence in their correct final predictions and incorrect final predictions on (a) all cases, (b) only those cases when their final predictions agreed with the AI recommendation, (c) only those cases when their final predictions disagreed with the AI recommendation. Error bars represent the standard errors of the mean. * represents the statistical significance level of 0.05.

To examine whether individuals' and groups' reliance on the AI model was appropriate, we further plot the comparisons in subjects' under-reliance and over-reliance on the AI model in Figures 4b and 4c, respectively. We found that when the AI model was correct, groups were more likely to rely on the AI model compared to individuals, resulting in a lower level of under-reliance. The result of our Welch's t-test with Bonferroni correction confirms that the difference is statistically significant (Individual-AI: $M = 0.37, SD = 0.34$; Group-AI: $M = 0.24, SD = 0.30$; $t(184) = 2.86$, *adjusted-p* = 0.009). Meanwhile, groups were also more likely to rely on the AI model than individuals when the AI model was wrong, that is, groups exhibited a significantly higher level of over-reliance on the AI model (Individual-AI: $M = 0.53, SD = 0.36$; Group-AI: $M = 0.68, SD = 0.33$; $t(180) = -2.89$, *adjusted-p* = 0.010). Together, these results suggest that in human-AI collaborative recidivism risk assessment, groups rely on the AI model more than individuals, regardless of the correctness of the AI model's recommendation.

4.3 Comparison on Decision Confidence

Next, we look into individuals' and groups' decision confidence in AI-assisted decision making. Figure 5a compares the average decision confidence for subjects in the Individual-AI and Group-AI treatment, and the comparison is conducted for subjects' correct and incorrect final decisions separately. Visually, regardless of whether the final decisions were correct, groups appeared to be slightly more confident in them than individuals. According to the results of the Welch's t-test with Bonferroni correction, however, these differences in decision confidence are not significant at the level of $p = 0.05$, both for correct decisions and incorrect decisions.

We note that since the AI model's decision recommendations were presented to subjects in AI-assisted decision making, whether subjects' final decisions agreed with the AI model may affect subjects' confidence in them. Thus, to obtain a more in-depth understanding of how groups' decision confidence compares with that of individuals, we further compared groups' and individuals' decision confidence conditioned on both the correctness of the final decision, and the agreement between the final decision and the AI model's recommendation (i.e., if subjects decided to rely on the AI model). The results are shown in Figures 5b and 5c. We found that individuals always had much lower confidence in their final decisions if they decided *not* to rely on the AI model, but we did not observe a similar decrease in decision confidence among group decision makers. Applying Welch's t-tests with Bonferroni corrections on each of the 4 scenarios (i.e., rely on AI and correct, rely on AI and wrong, not rely on AI and correct, not rely on AI and wrong), we found a significant difference in decision confidence between the two treatments when subjects' final decisions disagreed with the AI model's incorrect recommendation and made a correct

Treatment	ΔPOS	ΔTwin	ΔACC	ΔFPR	ΔFNR	ΔREL-POS	ΔREL-NEG
Individual-AI	0.089	0.011	-0.140	0.188	0.075	-0.103	-0.097
Group-AI	0.083	0.000	-0.018	0.120	-0.065	-0.053	-0.016
(adjusted) p-value	0.903	0.890	0.028*	0.658	0.270	1.000	0.395

Table 1: Comparing the fairness of individuals and 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.

final decision (Individual-AI: $M = 3.81, SD = 0.88$; Group-AI: $M = 4.09, SD = 0.85$; *adjusted-p* $= 0.047$), but no significant differences in the other three scenarios. This means that compared to individuals, groups are more confident when they reject an AI model's incorrect recommendation. This can be desirable because it implies that groups' decision confidence can be more calibrated than individuals' when they decide not to rely on the AI model.

4.4 Comparison on Understanding of AI

For subjects in the Individual-AI treatment, the average Pearson correlation coefficient between the true feature importance and the subject's perceived feature importance was 0.05. Meanwhile, for subjects in the Group-AI treatment, the average Pearson correlation coefficient value was 0.10, and the result of a Welch's t-test suggests that it is not statistically different than that in the Individual-AI treatment ($p = 0.344$). That is, in AI-assisted decision making, whether people make decisions on their own or in groups does not seem to significantly affect their understanding of the AI model.

4.5 Comparison on Decision Making Fairness

Table 1 reports the comparisons between the fairness of individuals and groups in AI-assisted decision making, on both how fair their decisions were and how fair they interacted with the AI model.

First, we focus on the five metrics that reflect how fair the individuals' or groups' decisions were—ΔPOS, ΔTwin, ΔACC, ΔFPR, and ΔFNR. Recall that we selected the 6 formal tasks for each subject in a way such that the true reoffending likelihood was the same within the set of Black defendants and the set of White defendants that the subject saw. Results in Table 1 suggest that in AI-assisted recidivism risk assessment, whether subjects made decisions individually or in a group did not seem to significantly change their likelihood of predicting Black defendants to reoffend in relative to White defendants (i.e., no significant difference is observed on ΔPOS or ΔTwin between the two treatments). However, compared to subjects who made decisions in groups, individual subjects' predictions on Black defendants appear to have a lower accuracy, a higher false positive rate, and a higher false negative rate than White defendants. When using Welch's t-tests to examine whether there exist any significant differences between the fairness level in individuals' and groups' decisions, we found that the difference on ΔACC was significant ($p = 0.028$). Indeed, for subjects in the Individual-AI treatment, their average prediction accuracy on Black defendants was 48.4%, which was 14% lower than that on White defendants. For subjects in the Group-AI treatment, however, the groups' average prediction accuracy on Black and White defendants were 56.9% and 58.7% respectively, with a much smaller gap of 1.8%; this indicates that groups' decisions are fairer than individuals' decisions with respect to the accuracy equality definition.

Next, we proceed to the two metrics that reflect how fair individuals and groups were when interacting with the AI model—ΔREL-POS and ΔREL-NEG. The results shown in Table 1 suggest that when the AI model predicted a defendant to reoffend (i.e., made a positive prediction), both individuals and groups seemed to rely on the AI model's predictions less when the defendant was Black than the case when the defendant was White (i.e., ΔREL-POS<0). Similar observations can also be made when the AI model predicted that a defendant will not reoffend (i.e., ΔREL-NEG<0). Yet, our statistical test results suggest that the differences on these two metrics are not significant across the two treatments, meaning that how fairly groups interacted with the AI model was not reliably different than how fairly individuals interacted with the AI model.

4.6 Comparison on Accountability

Lastly, we look into how subjects attributed accountability to themselves and the AI model differently when they made decisions as individuals or as groups[8]. Figure 6a illustrates the comparison across the two treatments on the normalized accountability that subjects assigned to themselves, conditioned on the final decisions being correct or incorrect. Figure 6b shows a similar comparison for the normalized accountability subjects assigned to the AI model. According to the results of the Welch's t-tests with Bonferroni correction, we found that if the final decision was correct, compared to individual decision-makers, the group decision-makers assigned a significantly lower level of credit to themselves (Individual-AI: $M = 1.10, SD = 0.29$; Group-AI: $M = 1.01, SD = 0.19$; $t(1115) = 5.31$, *adjusted-p* < 0.001) and a significantly higher level of credit to the AI model (Individual-AI: $M = 0.89, SD = 0.30$; Group-AI: $M = 0.96, SD = 0.30$; $t(1115) = -3.30$, *adjusted-p* $= 0.002$). In contrast, when the final decision was wrong, the assigned accountability to themselves or the AI model were not significantly affected by whether the decisions were made by individuals or groups.

Intuitively, the accountability assignment between oneself and the AI model in an AI-assisted decision making task may depend on whether the AI recommendation on that task is correct. To obtain a more fine-grained understanding, we further look into the comparisons in accountability assignment between individuals and groups for the cases where the AI recommendations were correct (Figure 7) and the cases where the AI recommendations were wrong (Figure 8) separately. Here, we found that the differences in the assigned accountability between the Individual-AI treatment and the Group-AI treatment that we previously saw in Figure 6

[8]Here, we did not compare the assigned accountability to the teammates, since subjects in the Individual-AI treatment did not have any teammates.

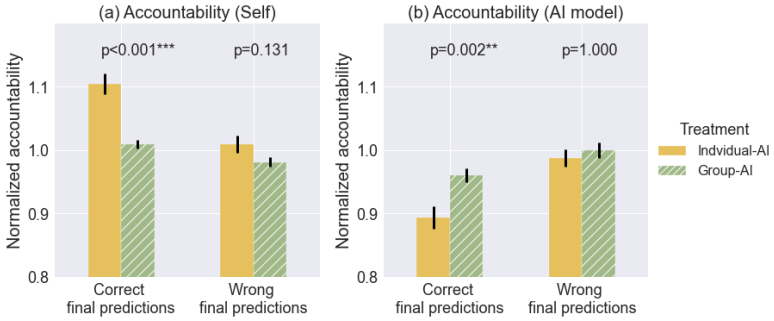


Figure 6: Comparing how subjects who made decisions as an individual or in a group assigned accountability to (a) themselves and (b) the AI model, conditioned on whether their final decision was correct. Error bars represent the standard errors of the mean. ** and * represent the statistical significance level of 0.01 and 0.001, respectively.**

Figure 7: Comparing when the AI recommendation was correct, how subjects who made decisions as an individual or in a group assigned accountability to (a) themselves and (b) the AI model, conditioned on whether their final decision was correct. Error bars represent the standard errors of the mean. ** and * represent the statistical significance level of 0.01 and 0.001, respectively.**

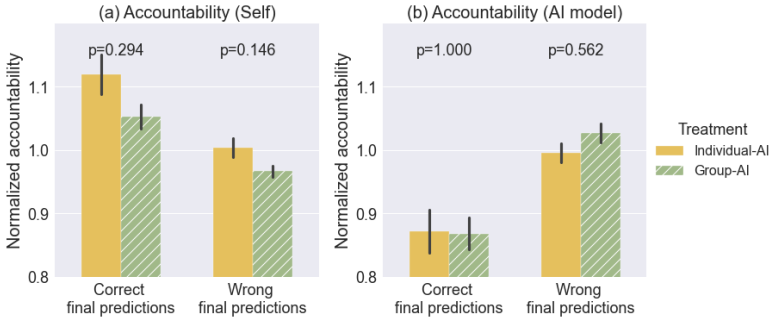


Figure 8: Comparing when the AI recommendation was wrong, how subjects who made decisions as an individual or in a group assigned accountability to (a) themselves and (b) the AI model, conditioned on whether their final decision was correct. Error bars represent the standard errors of the mean.

primarily came from those scenarios when the AI recommendations were correct. Specifically, we found that compared to individual decision makers, group decision makers significantly decreased the credit to themselves (*adjusted-p* < 0.001) while increasing the credit to the AI model (*adjusted-p* $= 0.001$) when they made correct final decisions with the help of correct AI recommendations. In contrast, when the AI recommendations were wrong, regardless of the correctness of subjects' final decisions, we found no significant differences in their accountability assignment between themselves and the AI model across the two treatments.

4.7 Exploratory Analyses

Finally, we conduct a few exploratory analysis to gain deeper insights into the group dynamics in AI-assisted decision making, and how the composition of groups affects their behavior and performance in AI-assisted decision making.

4.7.1 How does the presence of AI model impact group dynamics? Compared to the typical group decision making settings, in AI-assisted group decision making, group members can not only

Impacts on group dynamics	Examples
AI recommendation as a starter	*"I agree with the machine prediction."* (Subject 435, Group 42) *"I checked "will not", like the machine learning model."* (Subject 62, Group 8) *"one prior, could go either way. machine says will."* (Subject 907, Group 80) (All of the above quotes are the first message in a group's chat history)
Justify/Refute AI recommendation	*"I agree with the machine model. No prior charges, female, and white."* (Subject 1374, Group 125) *"I disagree with model. Too many past incidents and this was a serious crime."* (Subject 121, Group 12) *"The model thinks yes and that's a bit of history so I think they will reoffend."* (Subject 1320, Group 80)
Back up one's opinion using AI	*"I agree that the person will not reoffend because of no prior charges and the machine learning model prediction."* (Subject 142, Group 56) *"I think she will not reoffend. She was squeaky clean for 42 years...Also the machine agrees."* (Subject 27, Group 59)
AI recommendation as tiebreaker	*"Well I guess since we can't agree, let's just go with the machine."* (Subject 27, Group 59) *"okay we can go with the machine model then if you would prefer."* (Subject 1541, Group 130) *"should we go with the machine?"* (Subject 1541, Group 130)
Remind the group of AI decision	*"Our machine learning model predicts that this defendant will not reoffend in 2 years."* (Subject 158, Group 11)
Analyze the trustworthiness of AI	*"I think he will reoffend - 4 crimes after 18. The machine was wrong on many cases in the learning part."* (Subject 718, Group 61, AI predicts "not reoffend") *"I trust the model bot a lot."* (Subject 1269, Group 124)

Table 2: Example chat logs for the impacts of AI on group dynamics in AI-assisted group decision making.

interact with each other but also get access to the AI model's decision recommendations. Thus, we are interested in obtaining more understandings of how the presence of the AI model impacts the group dynamics in decision making. We analyzed the chat logs for all groups in our experiment, and identified a few representative impacts of AI on group dynamics. Some example chat logs illustrating these impacts are given in Table 2.

AI recommendations can help initiate the group deliberation. Due to the presence of the AI recommendations, we found that many groups started their deliberation process with some discussions related to the AI recommendations. For instance, some group members may explicitly express whether they agree with the AI model's recommendation or not at the beginning of the chat. Another example is that for some groups, the deliberation within the group often starts with one member giving a "briefing" of the current decision making case, which includes both the key features on the defendant's profiles and the AI model's recommendation. This observation implies that the AI recommendation may provide a reference point for some groups to initiate their deliberation process from some concrete decisions.

AI recommendations play various roles in the formation of group decisions. As groups go through the deliberation process, we found that different groups may incorporate the AI recommendations into their final decisions in different ways. For some groups, especially the ones that initiated the deliberation with some discussions related to the AI recommendations, the entire deliberation process could be centered around analyzing whether the AI recommendations were reasonable. As a result, group members attempted to justify why the AI model's recommendation was sensible or argue for why the AI model's recommendation could be wrong (see the "Justify/Refute AI recommendation" row in Table 2). In contrast, some other groups treated the AI recommendation more as a second opinion—group members expressed their opinions on the decision making case by analyzing the defendant's profiles on their own, while they cited the agreeing AI recommendations as the supporting evidence of their opinions to convince others (see the "Back up one's opinion using AI" row in Table 2). Interestingly, we found that sometimes the AI recommendation could also serve as the tiebreaker—when group members held different opinions and could not convince one another, they may eventually decide to go with the AI model's recommendation to reach a consensus (see the "AI recommendation as tiebreaker" row in Table 2).

Some groups made the efforts to ensure that AI's "voice" is heard and the trustworthiness of AI is discussed. On the group chatroom interface, by default, the first message displayed was always from a "risk analysis bot," which stated the AI model's decision recommendation. Interestingly, we noticed that during the discussion, members in some groups would re-post the AI recommendation, which effectively reminded group members of the AI recommendation and made sure that the AI model's opinion would not be left out. In addition, some groups also explicitly discussed their perceptions of the reliability and trustworthiness of the AI model to decide how to best utilize the AI recommendations (see the "Analyze the trustworthiness of AI" row in Table 2).

Not all groups discussed about AI recommendations. Finally, we note that not all groups discussed AI recommendations during their deliberation processes. Indeed, we found that some groups appeared to arrive at their final decisions solely based on their analyses on the defendant's profiles. While AI recommendations may have influenced how each group member approached the task and formed their own opinion (even though they did not explicitly state that in their chat messages), it is possible that AI recommendations were simply ignored by some groups in their decision making.

4.7.2 How does a group's cognitive diversity influence their behavior and performance in AI-assisted decision making? While we have obtained some understandings on how individuals and groups differ in their behavior and performance in AI-assisted decision making now, previous research on groups suggests that not all groups are the same. For example, many previous studies have suggested that the ways that a group behaves and how well it can perform depends on the group composition, such as the cognitive diversity of the group members [55]. Thus, in this section, we report the results of our second exploratory analysis on understanding how

the cognitive diversity of a group may impact its behavior and performance in AI-assisted decision making.

For this analysis, we first need to determine each subject's "cognitive style" in solving the recidivism risk assessment tasks. Cognitive style is the way that individuals perceive, process, and remember information; it reflects people's thought patterns and mental perspectives, and it is shown to impact how people make decisions [15, 58]. In this study, we operationalized the measurement of a subject's cognitive style in the context of the recidivism risk assessment task, and we quantified it as the subject's belief on how some selected features of the defendant's profile will affect the recidivism risk. Specifically, recall that in Phase 1, we asked each subject to work on a set of 8 tasks that were balanced on the defendant's race, charge degree, and the true recidivism status, with other features kept almost identical. Using a subject i's independent predictions on these 8 tasks, we computed the subject's belief in how the defendant's race affects recidivism risk (Belief_i^{race}) as the fraction of positive predictions that this subject made on all 4 Black defendants in Phase 1 minus the fraction of positive predictions that this subject made on all 4 White defendants in Phase 1. So, When $\text{Belief}_i^{race} > 0$ (or $\text{Belief}_i^{race} < 0$), it means that subject i associated the Black (or White) race with a higher chance of reoffending; the larger the value of $|\text{Belief}_i^{race}|$, the more subject i's recidivism risk assessment is influenced by the defendant's race. Similarly, we also computed subject i's belief in how the defendant's charge degree affects recidivism risk (Belief_i^{charge}) as the fraction of positive predictions subject i made on all 4 felony defendants in Phase 1 minus the fraction of positive predictions subject i made on all 4 misdemeanor defendants in Phase 1. Together, each subject i's cognitive style is then defined by their belief on both the defendant's race and charge degree and can be represented by the vector $\mathbf{s_i} = [\text{Belief}_i^{race}, \text{Belief}_i^{charge}]$.

Given the cognitive style of each subject in a group G, we then computed the similarity between two members i and j in the group as the cosine similarity score between their corresponding cognitive style vectors, i.e., $\text{sim}(i, j) = cos(\mathbf{s_i}, \mathbf{s_j})$. Thus, the similarity between two group members will be a value in the range of $[-1, 1]$—the larger the value of $\text{sim}(i, j)$ is, the more similar the cognitive styles of subjects i and j are; and $\text{sim}(i, j) < 0$ implies that the two subjects had *opposite* beliefs on the direction of either the impact of a defendant's race on recidivism risk (e.g., subject i believed Black race was associated with higher recidivism risk while subject j believed White race was associated with higher recidivism risk), or the impact of a defendant's charge degree on recidivism risk, or both. Finally, the group G's cognitive diversity was defined by the *largest* similarity score between any two members in the group—when $\max_{i,j \in G} \text{sim}(i, j) < 0$, the group is classified as a HIGH-DIVERSITY group (since even the most similar pair of members in the group is quite different from each other); otherwise, the group is classified as a LOW-DIVERSITY group.

With this classification, we got 13 HIGH-DIVERSITY groups and 79 LOW-DIVERSITY groups. We then compared how groups with different levels of cognitive diversity differ on their decision accuracy and confidence, reliance on AI, understanding of AI, fairness in decision making, and willingness to take accountability, following the same methodologies as we used for our earlier comparisons between individuals and groups. The complete set of results can be found in the supplemental materials. While for most of the metrics that we considered, we did not observe reliable differences between HIGH-DIVERSITY and LOW-DIVERSITY groups, in the following, we highlight a few significant differences that we detected.

First, we found that compared to LOW-DIVERSITY groups ($M = 0.75, SD = 0.28$), HIGH-DIVERSITY groups had a lower level of overall reliance on the AI model ($M = 0.60, SD = 0.27$; $t(91) = 2.16, p = 0.033$). We also found that HIGH-DIVERSITY groups and LOW-DIVERSITY groups exhibited some differences in the fairness level of their decision making. For example, groups with different levels of cognitive diversity appeared to differ significantly on how they make decisions for the twin cases ($p = 0.042$)—Given a pair of defendants who were identical on all features and their true recidivism status except for their race, while the LOW-DIVERSITY groups make roughly similar predictions for the Black defendant and the White defendant ($\Delta\text{Twin} = 0.038$), the HIGH-DIVERSITY groups tended to believe that the Black defendant in the pair has a much lower risk to reoffend ($\Delta\text{Twin} = -0.231$) and were therefore less fair. In addition, we also found that in general, when the AI model made a positive prediction, HIGH-DIVERSITY groups were more likely to follow it when the defendant was Black, which is significantly different from LOW-DIVERSITY groups who were more likely to follow it when the defendant was White (*adjusted-p* = 0.048).

Due to the significant unbalance in the number of HIGH-DIVERSITY and LOW-DIVERSITY groups and the exploratory nature of this analysis, we caution the readers to not over-interpret the results in this set of exploratory analysis. Nevertheless, the differences we observed between different groups in this analysis indicate that the cognitive diversity of groups *may* impact the behavior and performance of groups in AI-assisted decision making, and we hope these findings can inform more confirmatory studies in the future.

5 DISCUSSION

In this paper, using recidivism risk assessment as the decision domain, we present a comparative case study investigating the differences in the behavior and performance in AI-assisted decision making between groups and individuals. Our results suggest that groups and individuals show differences in both their decision making processes and outcomes in AI-assisted decision making. Specifically, whether decisions are made in a group or individually affects decision-maker's reliance on the AI model and confidence in the decisions, the fairness level of the decisions according to some fairness definition, as well as decision-maker's assignment of accountability in decision making, but does not significantly affect the accuracy of the decision or decision-maker's understanding of the AI model. In this section, we provide some potential explanations for the observed differences between groups and individuals in AI-assisted decision making, and discuss the limitations and implications of our findings.

5.1 On the (Seemingly) Conflicting Results between Reliance and Confidence

In our study, we find that groups rely on the AI recommendations significantly more than individuals, regardless of the AI correctness. Given our exploratory analysis findings in Section 4.7.1 on how the

presence of AI model changes group dynamics, we provide several possible explanations for this observation. First, as many groups directly start their deliberation process with discussions about the AI recommendations, it is possible that decision makers in these groups treat the AI recommendations as a reference point of their decisions and experience the "*anchoring effect*" to some extent. We note, however, since the AI recommendations are provided upfront to both individuals and groups, the individual decision makers may have also experienced the anchoring effect in their decision making. It is possible for the anchoring effect to either get amplified in groups (if each member in the group is affected by the anchoring effect to some degree) or get mitigated in groups (if exposure to disagreeing opinions from other group members reduces the anchoring effect). Rigorously comparing the magnitude of anchoring effect that individuals and groups experience in AI-assisted decision making is an interesting future work. Second, we observe that in some groups, certain member of the group will take the responsibility to remind group members of the AI recommendations. Thus, it is not surprising that AI's "opinion" is reflected in groups' decisions more often than in individuals' decisions. In fact, subjects who kept reminding others of the AI recommendation might even be perceived as the "defenders" of AI authority by their groupmates, who may feel the pressure to agree in order to reach a consensus. Finally, groups' higher level of reliance on the AI model could also relate to some group members' preferences to use the AI recommendation as the tiebreaker to resolve disagreement. This tendency to rely on the AI recommendation when consensus can not be easily reached may be further facilitated by the "*Bandwagon effect*", that is, people tend to adopt what others are doing especially when the cognitive load of making decisions is high [96]. This means that people may easily agree to follow the AI recommendation when others suggest to do so, as illustrated in the chat log of Group 61 below:

"Felony 33 stole a car. I think he possibly will re-offend. Not sure though. What do you think?" (Subject 718)

"So you believe he will offend?" (Subject 27)

"I kind of do. It's a felony and grand theft is a serious crime not petty." (Subject 718)

"will not reoffend within two years." (Subject 795)

"Okay, that's true." (Subject 27)

"Well I guess since we can't agree, let's just go with the machine." (Subject 27)

"I'm not sure. I'll go with what you guys think." (Subject 718)

On the other hand, despite the higher level of reliance on the AI model, we note that it is also inaccurate to claim that groups tend to *blindly* trust or rely on the AI recommendation, either. In fact, our results show that compared to individuals, groups can more confidently overturn the AI model's incorrect recommendation. Here, a plausible explanation is that whether groups will reject an AI recommendation depends on whether some member of the group confidently disagrees with the AI recommendation, which usually means the member considers the task to be "obvious" or easy for them, as shown in the next chat log of Group 80:

"one prior, could go either way. machine says will." (Subject 907)

"I'm disagreeing with the model, only second offense and first was juvie" (Subject 1320)

"ok, I can go with will not reoffend" (Subject 907)

"Yeah..true" (Subject 1602)

In other words, we conjecture that groups are more likely to accept the AI recommendation when all members have weak but disagreeing opinions, but they are more likely to confidently reject the AI recommendation if some member has a strong opinion that disagrees with AI.

5.2 More on Fairness in Decision Making

Our study demonstrates that AI-assisted decisions made by groups can be fairer than AI-assisted decisions made by individuals according to some fairness definition (e.g., accuracy equality). Meanwhile, groups' average decision accuracy is also not lower than that of individuals. In other words, groups have the potential to make fairer decisions according to some fairness criterion without sacrificing the decision accuracy. These results are consistent with previous research in psychology, which suggests that given the opportunities for discussions within a group, compared to individuals, bias can be mitigated within groups [124]. We note that this improvement in decision fairness is particularly valuable given that the AI model we used in the experiment (i.e., the COMPAS algorithm) was known to be biased [74], while the groups' overall increased reliance on the AI model does not result in a lower level of fairness in their decisions compared to the individuals. This again indicates that groups appear to rely on the AI recommendation in a selected manner rather than blindly. Moreover, we also find some exploratory evidence suggesting that the cognitive diversity of a group *may* affect the fairness level of the group's decisions. Interestingly, according to our exploratory analysis results, groups with higher level of cognitive diversity appear to associate a much lower recidivism risk to Black defendants than White defendants when holding everything else equal, which shows the opposite behavior to that of the AI model's. It is an interesting future work to understand why HIGH-DIVERSITY groups exhibit this behavior. In addition, this may imply the opportunity to intentionally construct groups with high cognitive diversity, and then obtain fairer final decisions by aggregating the decisions of the groups and those of the AI models.

5.3 On the Choice of Recidivism Risk Assessment as the Decision Domain

In this case study, we choose recidivism risk assessment as the decision domain to study the comparisons between individuals and groups in AI-assisted decision making. The recidivism risk assessment task can be representative of many real-world decision making domains where people can make decisions—either individually or in groups—with AI assistance, while they may have their own "bias" towards certain demographic groups when making these decisions, and the decision stakes are relatively high. Some examples of other decision making tasks that share a similar flavor include determining loan application approvals [45] and judging if a job interviewee is qualified [31]. In this sense, we conjecture that the results of our study may better generalize to this type of tasks.

However, we acknowledge that the generalizability of results in this study may be limited by a few key characteristics of the recidivism risk assessment task as well as the specific way that this task is operationalized in our study. For example, the recidivism risk

assessment task is relatively difficult for human decision makers. According to our pilot study (see supplemental materials for details), when subjects did not have access to the AI recommendations, their decisions on the formal tasks in Phase 2 HIT was relatively inaccurate (i.e., the average accuracy was 47.5%) and not very confident (i.e., the average confidence was 3.61 on a 5-point Likert scale, which was between "3: neither confident nor unconfident" and "4: confident"), and subjects' decisions on a task may often disagree with each other (i.e., on 80% of the tasks, at least 20% of people disagree with the majority decision). All of these characteristics may make humans tend to rely on the AI model, and possibly even more so when they make decisions in groups. On the other hand, we note that the AI model's performance on this task is also not very high, which is not uncommon for real-world decision making tasks that have a high level of inherent uncertainty. For instance, in the practice tasks of the Phase 2 HIT, an average subject's accuracy was 63.1%, which was even slightly higher than the AI model (62.5%). The fact that the performance gap between humans and AI is small reflects an opportunity to achieve human-AI complementarity [8], which may imply that decision makers can be relatively critical when determining whether to rely on AI recommendations in this task. More future studies need to be carried out to understand how the behavior and performance of individuals and groups differ when the task is substantially easier or even more difficult for humans, when the performance gap between humans and AI becomes larger, and when decision stakes are significantly different. We also highlight a few operationalizations of the recidivism risk assessment task in this study which may limit the generalizability of the results, including the choice of revealing AI recommendations to humans without having them make independent decisions first, allowing humans to get a sense of the performance comparison between the AI model and themselves before working on the formal AI-assisted decision making tasks, and using the original COMPAS algorithm as the AI model which was biased itself. It is important for future work to better understand whether the results of this study generalize to those settings where humans do not see AI recommendations upfront in their decision making, have limited information on the performance comparison between the AI model and themselves, or if the AI model used is much fairer itself.

5.4 Other Limitations

We acknowledge a few additional limitations of our study. First, some metrics we adopted for comparing the behavior and performance of individuals and groups in AI-assisted decision making may not be perfect despite being widely adopted in previous literature. For example, we used the fraction of tasks where the individual or group's final decision agrees with the AI recommendation to quantify their reliance on the AI model. The value of this metric may partly reflect the anchoring effect of the AI recommendation, rather than people's true willingness to rely on AI. It is also possible that this metric, to some extent, reflects the "natural agreement" between humans and AI. Fortunately, the results of our pilot study show that the fraction of natural agreement between humans' independent decisions and the AI recommendation (i.e., 42.3%) is much lower than the actual level of human-AI agreement we observed in our experiment (e.g., 60% for subjects in the INDIVIDUAL-AI treatment as shown in Figure 4a), which means that our reliance metrics are not merely reflecting the natural agreement (see supplemental materials for more details). Our metric for measuring subjects' understanding of the AI model was also limited to their capability to sort out the importance of different features. Moreover, subjects' capability to understand the AI model may also be constrained by the limited interactions they had with the model (e.g., they only interacted with the model for 15 tasks in the Phase 2 HIT). Future studies should investigate into individuals' and groups' understanding of AI after a longer period of interactions and from more diverse angles (e.g., simulate model predictions, answer counterfactual questions).

In addition, while group decision making via online collaborations is prevalent nowadays, other interaction modes, such as in-person collaborations, also exist in the real-world group decision making settings. Therefore, findings in this study may not directly generalize to other modes of group collaborations. Even for group decision making via online collaborations, our study setup also has some unique characteristics—subjects in our study were anonymous to each other, the collaboration was one-shot, and subjects were all laypeople with limited domain expertise. Hence, we caution the readers not to over-generalize our results to other online AI-assisted group decision making settings where group members know each other, need to engage in long-term interactions, and some or all group members have substantial domain expertise.

5.5 Future Directions

There are many interesting future directions for further enhancing our understanding of how groups behave and perform in AI-assisted decision making, and how they compare with individuals. First, we note that the formation of a group is decided by many factors, such as the group hierarchy (e.g., whether a leader exists and how to decide the leader), group size, and decision aggregation methods (e.g., majority rule and consensus). These factors may all influence people's attitudes and behavior in a group. Exploring how different formations of a group affect its behavior and performance in AI-assisted decision making is a critical direction to explore in the future. In addition, while we have conducted some exploratory analyses on how the cognitive diversity of groups affects their behavior and performance in AI-assisted decision making, there are many other types of diversity, such as background diversity, information diversity, and value diversity. Cognitive diversity of a group could also be defined in different ways. Investigating how different types of diversity within a group affect how groups make use of AI assistance in decision making is also another future direction. Similarly, while we obtain some initial understandings of how the presence of the AI model may change the dynamics of a group in AI-assisted decision making, more studies are needed to rigorously examine how the group dynamics impact the outcomes of AI-assisted group decision making, and how to promote more positive group dynamics in AI-assisted group decision making. Overall, given our findings in this study that individuals and groups behave and perform differently in AI-assisted decision making, it is also necessary to examine to what extent the findings we have obtained when studying individuals' interactions with the AI model can be generalized to groups' interactions with AI. Finally, a key limitation

of groups in AI-assisted decision making is that they have higher levels of over-reliance on the AI model. Thus, developing effective interventions to decrease groups' over-reliance on the AI model can be essential. Compared to those interventions designed for individual decision makers, the interventions provided to groups should not only answer the question of "how to intervene," but also "intervene to whom" to potentially utilize the social structure and influences within the group to maximize the effectiveness of the interventions.

6 CONCLUSION

In this paper, we present a comparative case study to obtain a systematic understanding of the differences in how individuals and groups behave and perform in AI-assisted decision-making. Our results suggest that compared to individuals, groups are more likely to rely on the AI recommendations regardless of their correctness, may generate fairer decisions according to some fairness definition, are more confident in their correct decisions when overturning incorrect AI recommendations, and are willing to assign less credit to themselves while assigning more credit to the AI model when they make correct decisions under the help of AI. In contrast, we do not find sufficient evidence that individuals and groups exhibit different levels of decision accuracy or understanding of the AI models in AI-assisted decision making. Overall, our results highlight that groups may outperform individuals in some aspects of AI-assisted decision making. Our work provides important implications for human-AI partnership, and we hope our findings in this paper can inspire more explorations in the area of group-AI interaction.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers who provided many helpful comments. We thank the support of the National Science Foundation under grant IIS-1850335 on this work. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

REFERENCES

[1] Abdullah Almaatouq, Mohammed Alsobay, Ming Yin, and Duncan J Watts. 2021. Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences* 118, 36 (2021), e2101062118.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. ProPublica (2016). *URL: https://www. propublica. org/article/machine-bias-risk-asses sments-in-criminal-sentencing* (2016).

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[4] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. 2021. AI-Assisted Human Labeling: Batching for Efficiency without Overreliance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[5] Bahador Bahrami, Karsten Olsen, Peter E Latham, Andreas Roepstorff, Geraint Rees, and Chris D Frith. 2010. Optimally interacting minds. *Science* 329, 5995 (2010), 1081–1085.

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

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

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

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *Nips tutorial* 1 (2017), 2.

[10] Bernard Bass. 1982. Individual capability, team performance, and team productivity. *Human Performance and Productivity. Vols* 1, 2 (1982), 179–222.

[11] Suzanne T Bell. 2007. Deep-level composition variables as predictors of team performance: a meta-analysis. *Journal of applied psychology* 92, 3 (2007), 595.

[12] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[13] Yochanan E Bigman and Kurt Gray. 2018. People are averse to machines making moral decisions. *Cognition* 181 (2018), 21–34.

[14] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[15] Bruce K Blaylock and Loren P Rees. 1984. Cognitive style and the usefulness of information. *Decision Sciences* 15, 1 (1984), 74–91.

[16] Jonathan David Bobaljik and Höskuldur Thráinsson. 1998. Two heads aren't always better than one. *Syntax* 1, 1 (1998), 37–71.

[17] Marcus T Boccaccini, Darrel B Turner, Daniel C Murrie, Craig E Henderson, and Caroline Chevalier. 2013. Do scores from risk measures matter to jurors? *Psychology, Public Policy, and Law* 19, 2 (2013), 259.

[18] Tim Brennan, William Dieterich, and Beate Ehret. 2009. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and behavior* 36, 1 (2009), 21–40.

[19] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[20] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. 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 (2021), 1–21.

[21] Madalina Busuioc. 2021. Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review* 81, 5 (2021), 825–836.

[22] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[23] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[24] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.

[25] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[26] Chun-Wei Chiang and Ming Yin. 2021. You'd better stop! Understanding human reliance on machine learning models under covariate shift. In *13th ACM Web Science Conference 2021*. 120–129.

[27] Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *27th International Conference on Intelligent User Interfaces*. 148–161.

[28] Leah Chong, Ayush Raina, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. The Evolution and Impact of Human Confidence in Artificial Intelligence and in Themselves on AI-Assisted Decision-Making in Design. *Journal of Mechanical Design* (2022), 1–37.

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

[30] Nancy J Cooke, Mustafa Demir, and Nathan McNeese. 2016. *Synthetic teammates as team players: Coordination of human and synthetic teammates*. Technical Report. Cognitive Engineering Research Institute Mesa United States.

[31] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[32] Bart A De Jong, Kurt T Dirks, and Nicole Gillespie. 2016. Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of applied psychology* 101, 8 (2016), 1134.

[33] Marie Delacre, Daniël Lakens, and Christophe Leys. 2017. Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology* 30, 1 (2017).

[34] Mustafa Demir, Nathan J McNeese, and Nancy J Cooke. 2016. Team communication behaviors of the human-automation teaming. In *2016 IEEE international multi-disciplinary conference on cognitive methods in situation awareness and decision support (CogSIMA)*. IEEE, 28–34.

[35] Dennis J Devine and Jennifer L Philips. 2001. Do smarter teams do better: A meta-analysis of cognitive ability and team performance. *Small group research* 32, 5 (2001), 507–532.

[36] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[37] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.

[38] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[39] Xiaoni Duan, Chien-Ju Ho, and Ming Yin. 2022. The influences of task design on crowdsourced judgement: A case study of recidivism risk evaluation. In *Proceedings of the ACM Web Conference 2022*. 1685–1696.

[40] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[41] Ferda Erdem, Janset Ozen, and Nuray Atsan. 2003. The relationship between trust and team performance. *Work study* (2003).

[42] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.

[43] Northpointe Institute for Public Management. 1996. COMPAS [Computer software].

[44] Donelson R Forsyth. 2018. *Group dynamics*. Cengage Learning.

[45] Jorge Galindo and Pablo Tamayo. 2000. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational economics* 15, 1 (2000), 107–143.

[46] Yashesh Gaur, Walter S Lasecki, Florian Metze, and Jeffrey P Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference*. 1–8.

[47] Meric Altug Gemalmaz and Ming Yin. 2022. Understanding Decision Subjects' Fairness Perceptions and Retention in Repeated Interactions with AI-Based Decision Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 295–306.

[48] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[49] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.

[50] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[51] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[52] Randy Y Hirokawa and Marshall Scott Poole. 1996. *Communication and group decision making*. Sage Publications.

[53] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[54] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

[55] Sujin K Horwitz and Irwin B Horwitz. 2007. The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of management* 33, 6 (2007), 987–1015.

[56] Yoyo Tsung-Yu Hou and Malte F Jung. 2021. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[57] George P Huber and Kyle Lewis. 2010. Cross-understanding: Implications for group cognition and performance. *Academy of Management review* 35, 1 (2010), 6–26.

[58] Raymond G Hunt, Frank J Krzystofiak, James R Meindl, and Abdalla M Yousry. 1989. Cognitive style and decision making. *Organizational behavior and human decision processes* 44, 3 (1989), 436–453.

[59] Niranjan S Janardhanan, Kyle Lewis, Rhonda K Reger, and Cynthia K Stevens. 2020. Getting to know you: motivating cross-understanding for improved team and individual performance. *Organization Science* 31, 1 (2020), 103–118.

[60] Irving Lester Janis. 1983. *Groupthink*. Houghton Mifflin Boston.

[61] Karen A Jehn, Gregory B Northcraft, and Margaret A Neale. 1999. Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative science quarterly* 44, 4 (1999), 741–763.

[62] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.. In *IJCAI*. 4070–4073.

[63] Steven J Karau and Kipling D Williams. 1993. Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology* 65, 4 (1993), 681.

[64] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[65] Norbert L Kerr, R Scott Tindale, et al. 2004. Group performance and decision making. *Annual review of psychology* 55, 1 (2004), 623–655.

[66] Antino Kim, Mochen Yang, and Jingjng Zhang. 2020. When Algorithms Err: Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms. *Late Errors on Users' Reliance on Algorithms (July 2020)* (2020).

[67] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W Malone. 2017. What makes a strong team? Using collective intelligence to predict team performance in League of Legends. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2316–2329.

[68] Keith Kirkpatrick. 2017. It's not the algorithm, it's the data. *Commun. ACM* 60, 2 (2017), 21–23.

[69] Asher Koriat. 2012. When are two heads better than one and why? *Science* 336, 6079 (2012), 360–362.

[70] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[71] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[72] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[73] Molly K Land and Jay D Aronson. 2020. Human rights and technology: new challenges for justice and accountability. *Annual Review of Law and Social Science (Forthcoming)* (2020).

[74] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9, 1 (2016), 3–3.

[75] James R Larson Jr. 2013. *In search of synergy in small group performance*. Psychology Press.

[76] James R Larson Jr, Pennie G Foster-Fishman, and Timothy M Franz. 1998. Leadership style and the discussion of shared and unshared information in decision-making groups. *Personality and Social Psychology Bulletin* 24, 5 (1998), 482–495.

[77] Patrick R Laughlin and John Adamopoulos. 1980. Social combination processes and individual learning for six-person cooperative groups on an intellective task. *Journal of Personality and Social Psychology* 38, 6 (1980), 941.

[78] Patrick R Laughlin, Bryan L Bonner, and Andrew G Miner. 2002. Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes* 88, 2 (2002), 605–620.

[79] Daniel Levi and David A Askay. 2020. *Group dynamics for teams*. Sage Publications.

[80] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.

[81] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[82] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *arXiv preprint arXiv:2101.05303* (2021).

[83] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[84] Zhuoran Lu and Ming Yin. 2021. 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*.

[85] Pamela J Ludford, Dan Cosley, Dan Frankowski, and Loren Terveen. 2004. Think different: increasing online community participation using uniqueness and

group dissimilarity. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 631–638.
[86] Merce Mach, Simon Dolan, and Shay Tzafrir. 2010. The differential effect of team members' trust on team performance: The mediation role of team cohesion. *Journal of Occupational and Organizational Psychology* 83, 3 (2010), 771–794.
[87] Andrew Mao, Winter Mason, Siddharth Suri, and Duncan J Watts. 2016. An experimental study of team size and performance on a complex task. *PloS one* 11, 4 (2016), e0153048.
[88] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.
[89] Winter Mason and Duncan J Watts. 2012. Collaborative learning in networks. *Proceedings of the National Academy of Sciences* 109, 3 (2012), 764–769.
[90] Nathan J McNeese, Beau G Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. 2021. Who/What Is My Teammate? Team Composition Considerations in Human–AI Teaming. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 288–299.
[91] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
[92] Larry K Michaelsen, Warren E Watson, and Robert H Black. 1989. A realistic test of individual versus group consensus decision making. *Journal of applied psychology* 74, 5 (1989), 834.
[93] Dena F Mujtaba and Nihar R Mahapatra. 2019. Ethical considerations in AI-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 1–7.
[94] Geoff Musick, Thomas A O'Neill, Beau G Schelble, Nathan J McNeese, and Jonn B Henke. 2021. What Happens When Humans Believe Their Teammate is an AI? An Investigation into Humans Teaming with Autonomy. *Computers in Human Behavior* 122 (2021), 106852.
[95] David G Myers and Helmut Lamm. 1976. The group polarization phenomenon. *Psychological bulletin* 83, 4 (1976), 602.
[96] Richard Nadeau, Edouard Cloutier, and J-H Guay. 1993. New evidence about the existence of a bandwagon effect in the opinion formation process. *International Political Science Review* 14, 2 (1993), 203–213.
[97] Lisa Hope Pelled, Kathleen M Eisenhardt, and Katherine R Xin. 1999. Exploring the black box: An analysis of work group diversity, conflict and performance. *Administrative science quarterly* 44, 1 (1999), 1–28.
[98] Kaśka Porayska-Pomsta and Gnanathusharan Rajendran. 2019. Accountability in human and artificial intelligence decision-making as the basis for diversity and educational inclusion. In *Artificial intelligence and inclusive education*. Springer, 39–59.
[99] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.
[100] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
[101] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. 1–14.
[102] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[103] Niloufar Salehi, Andrew McCabe, Melissa Valentine, and Michael Bernstein. 2017. Huddler: Convening stable and familiar crowd teams despite unpredictable availability. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1700–1713.
[104] Julian Sanchez, Wendy A Rogers, Arthur D Fisk, and Ericka Rovira. 2014. Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical issues in ergonomics science* 15, 2 (2014), 134–160.
[105] Donghee Shin. 2020. User perceptions of algorithmic decisions in the personalized AI system: perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media* 64, 4 (2020), 541–565.
[106] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.
[107] Tony Simons, Lisa Hope Pelled, and Ken A Smith. 1999. Making use of difference: Diversity, debate, and decision comprehensiveness in top management teams. *Academy of management journal* 42, 6 (1999), 662–673.
[108] Joachim Stempfle and Petra Badke-Schaub. 2002. Thinking in design teams-an analysis of team communication. *Design studies* 23, 5 (2002), 473–496.
[109] Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *12th ACM Conference on Web Science*. 315–324.
[110] James Surowiecki. 2005. *The wisdom of crowds*. Anchor.
[111] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In *CHI Conference on Human Factors in Computing Systems*. 1–17.
[112] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2021)*.
[113] Ray Tsaih, Yenshan Hsu, and Charles C Lai. 1998. Forecasting S&P 500 stock index futures with a hybrid AI system. *Decision support systems* 23, 2 (1998), 161–174.
[114] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.
[115] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.
[116] Victor H Vroom. 2000. Leadership and the decision-making process. *Organizational dynamics* 28, 4 (2000), 82–94.
[117] Daisuke Wakabayashi. 2018. Self-driving Uber car kills pedestrian in Arizona, where robots roam. *The New York Times* 19, 03 (2018).
[118] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[119] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. In *Proceedings of the ACM Web Conference 2022*. 1697–1708.
[120] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.
[121] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Transactions on Interactive Intelligent Systems (TiiS)* (2022).
[122] Jenny S Wesche and Andreas Sonderegger. 2019. When computers take the lead: The automation of leadership. *Computers in Human Behavior* 101 (2019), 197–209.
[123] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.
[124] Edward F Wright and Gary L Wells. 1985. Does group discussion attenuate the dispositional bias? *Journal of Applied Social Psychology* 15, 6 (1985), 531–546.
[125] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.
[126] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. Making sense of recommendations. *Journal of Behavioral Decision Making* 32, 4 (2019), 403–414.
[127] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
[128] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate? An Investigation from Perception to Decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277
[129] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.
[130] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An Ideal Human" Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.
[131] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. 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*. 295–305.
[132] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
[133] Sharon Zhou, Melissa Valentine, and Michael S Bernstein. 2018. In search of the dream team: Temporally constrained multi-armed bandits for identifying effective team structures. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–13.