The Effects of AI Biases and Explanations on Human Decision Fairness: A Case Study of Bidding in Rental Housing Markets

Xinru Wang[1] , **Chen Liang**[2] , **Ming Yin**[1]
[1]Purdue University
[2]University of Connecticut
xinruw@purdue.edu, chenliang@uconn.edu, mingyin@purdue.edu

Abstract

The use of AI-based decision aids in diverse domains has inspired many empirical investigations into how AI models' decision recommendations impact humans' decision accuracy in AI-assisted decision making, while explorations on the impacts on humans' decision fairness are largely lacking despite their clear importance. In this paper, using a real-world business decision making scenario–bidding in rental housing markets—as our testbed, we present an experimental study on understanding how the bias level of the AI-based decision aid as well as the provision of AI explanations affect the fairness level of humans' decisions, both *during* and *after* their usage of the decision aid. Our results suggest that when people are assisted by an AI-based decision aid, both the higher level of racial biases the decision aid exhibits and surprisingly, the presence of AI explanations, result in more unfair human decisions across racial groups. Moreover, these impacts are partly made through triggering humans' "disparate interactions" with AI. However, regardless of the AI bias level and the presence of AI explanations, when people return to make independent decisions after their usage of the AI-based decision aid, their decisions no longer exhibit significant unfairness across racial groups.

1 Introduction

As Artificial Intelligence (AI) technology advances in the past decade, more AI-based decision aids have been developed to assist human decision making in various domains such as hiring [Peng *et al.*, 2022], smart pricing [Zhang *et al.*, 2021], and criminal justice [Angwin *et al.*, 2016]. Many empirical studies have been conducted to evaluate the effectiveness of the collaborations between humans and AI in AI-assisted decision making settings, especially in terms of the decision making accuracy of the human-AI team. It has been found that AI recommendations often help the human-AI team make more accurate decisions, surpassing the performance of human decision makers alone [Bansal *et al.*, 2021; Lai and Tan, 2019]. However, they seldom lead to a team that outperforms both humans and AI, possibly due to decision makers' inappropriate reliance on AI recommendations [Dietvorst *et al.*, 2015; Buçinca *et al.*, 2021].

On the other hand, while many studies find that AI models can inherit biases from the training data and show unequal treatment to individuals from different groups [Angwin *et al.*, 2016], much less attention has been paid on evaluating the *fairness* of human decisions in AI-assisted decision making, despite its clear importance. Only a few most recent research finds that humans have some capabilities in correcting the biases of AI models, thereby the racial and socioeconomic disparities exhibited in their AI-assisted decisions are smaller than those inherent in the AI models [Fogliato *et al.*, 2022], although the specific architecture of the AI model employed may also influence human decision fairness [Peng *et al.*, 2022]. In general, however, systematic understandings on which factors may influence human decision fairness in AI-assisted decision making and how are still largely lacking.

In this paper, we make an initial attempt to fill this research gap. Specifically, we conjecture that a key factor that may affect human decision fairness in AI-assisted decision making is *the level of bias* exhibited by the AI model, though the precise impact is unclear—while higher levels of AI biases may exacerbate the unfair human decisions, it is also possible that humans may consciously avoid being influenced by the AI model after recognizing it as highly biased. Another potential influencing factor is *the provision of AI explanations*—the AI explanations may explicitly expose the AI model's biases to human decision makers [Dodge *et al.*, 2019], but they may also lead to decision makers' unwarranted faith in the model and subsequently change their receptivity to AI recommendations [Ehsan *et al.*, 2021; Schaffer *et al.*, 2019]. Thus, our first research question is:

- **RQ1**: During AI-assisted decision making, how do the bias level of an AI model and the provision of AI explanations affect **(a)** the fairness level of humans' decision *outcome*, and **(b)** the fairness level of humans' decision *process*, in terms of the extent to which they are influenced by the AI recommendations in their interactions with AI?

Moreover, as an AI model is often depicted as being able to uncover hidden patterns from data, the AI-assisted decision making process may also present an opportunity for humans to "learn" the data-driven insights from the AI model. Thus,

we are prompted to investigate a second research question:

- **RQ2**: How do the bias level of an AI model and the provision of AI explanations affect the fairness level of humans' independent decision outcomes, *after* they have been assisted by the AI model?

To answer these questions, we conducted a randomized human-subject experiment. Our experiment involved a real-world business decision making scenario, i.e., bidding in a rental housing market. We recruited participants on Amazon Mechanical Turk (MTurk) and asked them to act as travel agents to help clients bid on rental property listings provided by hosts of different races, given the client's budgets. We created a total of 5 treatments in our experiment, including a control treatment in which participants had no access to AI models, and the other four treatments were arranged in a 2 by 2 factorial design varying on the bias level of the AI model (low vs. high) and the provision of AI explanations (with vs. without). Moreover, decision making tasks in our experiments were divided into two phases—participants made decisions with AI assistance in Phase 1 (when applicable), and independent decisions without AI assistance in Phase 2.

Our experimental results demonstrate that when human decision makers are assisted by AI models in their bidding, the bias that the AI model exhibits across hosts of different races can propagate to humans' decisions, as humans' decision outcomes become less fair across racial groups when the AI model's bias level gets higher. Surprisingly, we also find that the provisions of AI explanations, which is often believed to help people identify fairness issues of the AI model, turns out to result in higher levels of unfairness in humans' decision outcomes. Moreover, both the higher bias level of the AI model and the presence of AI explanations are shown to result in more unfair human decisions across racial groups by increasing the human decision makers' "*disparate interactions*" [Green and Chen, 2019a; Green and Chen, 2019b] with the AI model (i.e., the extent to which humans' decisions are influenced by the AI model's recommendations is different across hosts of different races). For example, when the AI bias level becomes higher, decision makers are more strongly influenced by the AI recommendations to decrease their bid prices to Black hosts than to White hosts. Finally, we find that the AI biases and explanations have no impacts on the fairness of humans' independent decisions after they have been assisted by the AI model, which implies that the impacts of AI models on human decision fairness do not extend beyond the short-term usage of the model. Together, these results highlight the needs for gaining deeper understandings on the impacts of AI model properties and presentations on various aspects of human decision making.

2 Related Work

Biases and fairness concerns throughout the AI model development pipeline have been increasingly recognized by researchers and practitioners alike. The machine learning community has responded by proposing many algorithmic fairness definitions [Verma and Rubin, 2018; Mehrabi *et al.*, 2021] and developing many bias mitigation methods and tools [Dwork *et al.*, 2012; Bird *et al.*, 2020; Hu *et al.*, 2020; Duan *et al.*, 2020]. More recently, a growing number of human-centered studies have been carried out to understand people's fairness perceptions of AI models [Saxena *et al.*, 2019; Wang *et al.*, 2020; Gemalmaz and Yin, 2022]. For example, Srivastava *et al.* [2019] matched individual perceptions of fairness to mathematical definition of fairness, and found that "demographic parity" aligns the best with humans' ideas of fairness. Following many calls to present explanations to decisions to promote informational justice, researchers have also explored the effects of providing AI explanations on people's perceived algorithmic fairness. The results are mixed, however, as choices of explanation styles impact fairness perception in different ways [Binns *et al.*, 2018; Dodge *et al.*, 2019; Angerschmid *et al.*, 2022].

On the other hand, despite AI models have been increasingly used in supporting humans in their decision making, research on how the fairness level of humans' decisions is affected by their usage of AI-based decision aids is limited. Indeed, most empirical research on AI-assisted decision making has been carried out to understand how humans trust [Chiang and Yin, 2022; Zhang *et al.*, 2020] and understand the AI model [Wang and Yin, 2021; Poursabzi-Sangdeh *et al.*, 2021], and how their decision accuracy is affected by their usage of the AI-based decision aids [Lai and Tan, 2019; Bansal *et al.*, 2021]. Only most recently, some efforts have been spent on examining the fairness of AI-assisted decisions. For example, it is found that interacting with an AI model may have differing impacts on the fairness of humans' decisions depending on the model configurations [Peng *et al.*, 2022] and the balanced degree of data representation in the decision making tasks humans encounter [Peng *et al.*, 2019]. Also, humans show a degree of bias in their interactions with the AI model [Green and Chen, 2019a; Green and Chen, 2019b], thus they may be influenced by AI recommendations to different extents on decision making cases concerning different demographic groups. Our work complements prior work by systematically examining how the bias level and the provision of explanations of an AI model affect human decision fairness, both *during* and *after* humans' interactions with the AI model.

3 Study Design

To understand how the bias level of AI models and the provision of AI explanations affect human decision fairness during and after AI-assisted decision-making, we conducted a randomized human-subject experiment on MTurk.

3.1 Experimental Setup

Tasks. In our experiment, participants were asked to complete a set of bidding tasks. Specifically, each participant was told to act as a "travel agent" to help their "clients" secure rental places to stay for one night in New York City. In each task, participants were presented with a profile of a rental house listing with 21 features, including information on the host (e.g., race, superhost status), the house (e.g., number of beds), and reviews of the listing (e.g., overall rating and rating on sub-scales). Participants were also given the client's "*budget*" for this listing, which was the highest possible price

the client could pay for it. With these information, participants could then submit a bid price on behalf of the client to the host of the listing, which should not exceed the client's budget. Participants were told that the host also had an "*asking price*" in their mind, which was the lowest possible price that they can accept to rent their place. As a result, if the participant's bid price was no less than the host's asking price, the transaction would succeed—the client would pay the bid price to the host, and pay 50% of the difference between their budget and the bid price to the participant as the reward to them. However, if the participant's bid price was lower than the host's asking price, the transaction would not happen and the participant would earn zero reward[1]. Thus, to maximize their rewards, in each task, participants should try to submit a bid price that is as low as possible, while ensuring that it is greater than the host's asking price for the transaction to happen—in other words, participants need to make accurate predictions of the host's asking price. To assist participants in making these predictions, an AI model's predictions and explanations may be provided to participants in some experimental treatments (see Section 3.2 for details).

Dataset. Rental house profiles that participants saw in the experiment were taken from a dataset provided on the Inside Airbnb website[2]—we downloaded a version of the NYC Airbnb dataset that contained records about all 38,277 Airbnb listings in New York City scraped in December 2021, and we restricted our attention to short-term rental listings whose minimum nights to stay were fewer than 7 days. For each listing, in addition to information on the host, house, and reviews, we also had its daily price that was given by the host of the listing; this was used as the ground truth for the host's asking price in our experiment. To determine each host's race, following [Zhang *et al.*, 2021], we first used Deepface [Serengil and Ozpinar, 2021], a lightweight facial recognition and attribute analysis framework, to categorize the host's race based on their profile photo, and then recruited a group of in-house annotators to manually verify the correctness of the White/Black race labels given by Deepface. This procedure yielded a cleaned dataset of 3,884 Airbnb listings whose hosts were verified as either White or Black[3].

Within the cleaned dataset, we found a clear price gap between listings provided by Black and White hosts in NYC (top two rows in Table 1, Mann-Whitney U test: $p < 0.001$). To control for potential confounding influences on the listing's price caused by factors other than the race of the host, we conducted coarsened exact matching (CEM) [Iacus *et al.*, 2012] to group together listings with similar characteristics but different host race. We considered the number of guests the listing can accommodate, number of bedrooms, whether the listing is a private room or an entire unit, whether the host is a superhost, number of reviews, review scores, and the neighborhood[4], as covariates to be used in creating the subclasses, and we coarsened all numeric covariates into two bins. This yielded a subset of 1,853 matched listings belonging to 231 subclasses. As shown in Table 1 (bottom rows), the price gap between listings provided by Black and White hosts is still significant after matching ($p < 0.001$).

	Host's Race	# of Listings	Avg. Price
All	Black	876	\$136.5
	White	3,008	\$196.8
Matched	Black	580	\$134.6
	White	1,273	\$161.8

Table 1: Price gap between listings provided by Black and White hosts in the entire cleaned dataset and the matched subset.

3.2 Experimental Treatments

We adopted a between-subject design by randomly assigning participants into one of the 5 treatments—a control treatment in which participants had no access to AI models in all bidding tasks, and another four experimental treatments arranged in a 2×2 factorial design. In the later 4 treatments, participants were assisted by an AI model in predicting the host's asking price in the first phase of bidding tasks (see Section 3.3 for details). The AI model used in different treatments differed on *the level of racial bias* it exhibited when predicting prices for listings provided by White/Black hosts (low-bias vs. high-bias), and *the provision of explanations* for why the AI model made certain predictions (with vs. without).

Bias level of AI models. We randomly selected 50% of data samples in the cleaned dataset, both within and outside of the matched subset as the held-out test data, while the rest 50% of the data was used as the training dataset for the AI model. We followed the fair regression algorithm proposed in [Agarwal *et al.*, 2019] to train AI models with different levels of bias.

Specifically, given training examples in the form of (X, A, Y) triples, where X is a feature vector, $A \in \mathcal{A}$ is a protected attribute (e.g., host's race), and $Y \in \mathcal{Y} = [0, 1]$ is the label, the fair regression algorithm attempts to train a predictor f that satisfies *demographic parity* such that $f(X)$ is independent of A. Since $f(X) \in [0, 1]$, achieving the (approximate) demographic parity is equivalent to ensure that on the cumulative distribution function of $f(X)$, $|\mathbb{P}[f(X) \geq z|A = a] - \mathbb{P}[f(X) \geq z]| \leq \epsilon, \forall a \in \mathcal{A}, z \in [0, 1]$. Intuitively, the smaller the slack parameter $\epsilon (0 \leq \epsilon \leq 1)$, the less biased the predictor is across subgroups with different protected attribute values. Thus, based on our training dataset, we trained two linear regression models by setting the slack parameter value as $\epsilon_h = 1$ and $\epsilon_l = 0.005$, separately. These two models were then used as the AI models in the experiment for the high-bias and low-bias AI treatments, respectively (see

[1]To induce realism in tasks, we used incentive structures that align with actual bid agent behaviors by adapting the widely used incentive-aligned contingent valuation method [Becker *et al.*, 1964] to the travel agent scenario.

[2]http://insideairbnb.com/get-the-data.

[3]Note that participants in our experiment did not see the host's actual profile photo in a task. Instead, based on the host's race and gender, a photo was randomly selected from the public Chicago Face Database [Ma *et al.*, 2015] and presented to participants.

[4]Based on the geographical information of each listing, we obtained the zip code of the area that the listing locates at and mapped it to different neighborhoods [Des Jarlais *et al.*, 2018].

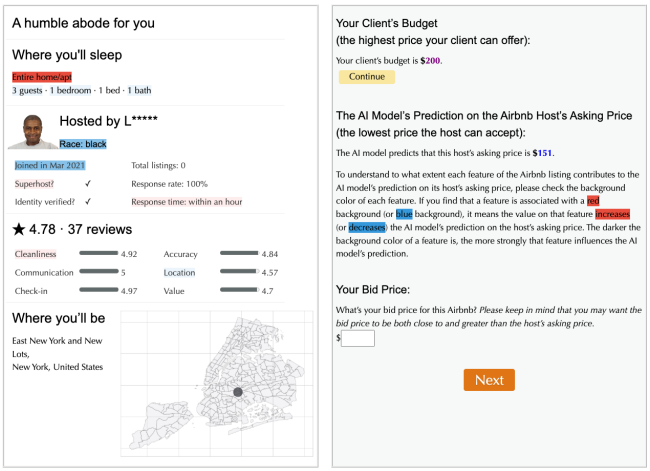


Figure 1: An example of the task interface (with AI explanation).

the supplemental materials for the evaluations of these two models on the test dataset).

Provision of AI explanations. We further adopted the SHAP algorithm [Lundberg and Lee, 2017], a model-agnostic explanation method, to compute the contribution that each feature in the listing profile made to the AI model's prediction on the host's asking price for that listing. In each task across all treatments with AI explanations, we color-coded each feature of the rental place listing based on its SHAP value to help participants understand how different features influence the AI model's prediction—A feature was highlighted in a red (or blue) background if based on the SHAP algorithm, its value increased (or decreased) the AI model's price prediction; the darker the background color, the larger the influence was (see Figure 1 for an example).

3.3 Experimental Procedure

We posted our experiment as a human intelligence task (HIT) on MTurk. Upon arrival, participants were randomly assigned to one of the 5 treatments as described in Section 3.2. They first completed a questionnaire on their background, including their demographics, familiarity with Airbnb, technical literacy, and expertise in AI and machine learning. Then, we presented participants with an interactive tutorial to explain the bidding task to them as well as walking through the interface with them. To help participants get a sense of the Airbnb rental housing market in NYC, in the tutorial, we also presented the price distribution and summary statistics for a one-night Airbnb stay in NYC based on our training dataset. Upon completion of the tutorial, participants were asked to answer a few qualification questions to show they understood all the information presented. They could not proceed to the next part of the experiment unless they answered all qualification questions correctly.

After passing the qualification, participants started to work on a sequence of bidding tasks divided into two phases, each consisting of 10 tasks. Phase 1 was designed to evaluate the fairness of human decisions *when* they are assisted by an AI model. Thus, prior to the commencement of Phase 1, all participants, except for those in the control group, were explicitly informed that they were provided with a free trial of an AI-powered predictive tool to help them predict the host's asking price for each rental place listing. Then, in each task, we presented to participants the profile of the listing, along with the client's budget, which was set to $200 for one-bed listing or $250 for listings with more than one bed. Depending on the treatment a participant was assigned to, the AI model's prediction on the host's asking price with or without its explanations might also be provided. Participants then needed to submit their bid prices, although feedback on whether the transaction succeeded and how much reward they earned was not revealed to them until the end of the experiment.

Note that the task instances we presented to participants in Phase 1 were carefully selected from our test dataset. In particular, the 10 tasks in Phase 1 were consisted of 8 "regular" instances and a pair of "counterfactual" instances (presentation order was randomized). We conjectured that participants' bid price for a listing can be largely affected by the host's race (Black vs. White), the number of beds (1 vs. 1+), and whether the host is a superhost. We selected 8 regular instances to be balanced on these three features (i.e., one instance from each feature combination). Additionally, in accordance with the correspondence study design [Bertrand and Duflo, 2017], the pair of counterfactual instances was selected from the matched subset in the test dataset, so they were very similar to each other except for the host's race. In the bidding task sequences, we intentionally presented the two counterfactual tasks to participants one after another—after participants finished bidding on the first instance in the pair, we told them that the previous listing turned out to be not available, but there was a very similar listing nearby, except that it was owned by a host of a different race (the presentation order of race within the counterfactual instances was randomized). Then, in the next task, we showed participants the second instance in the pair and asked them to bid on it, and we included their bid price and if applicable, the AI model's prediction, on the previous instance on the interface for their reference. Importantly, for the 10 task instances of Phase 1, we manually verified that the explanations of the high-bias AI model consistently suggested a host's race being Black resulted in a relatively large decrease in the AI model's price prediction, while the explanations of the low-bias AI model indicated very little impact from the host's race.

Phase 2 was very similar to Phase 1, except for that it was designed to evaluate the fairness of human decisions *after* they have had the experience of being assisted by an AI model in their decision making. Prior to the commencement of Phase 2, we told participants who had access to an AI model in Phase 1 that the free trial of the AI-powered predictive tool was ended. As a result, participants of all treatments predicted the host's asking price in Phase 2 tasks on their own. We still included 8 balanced regular instances and a pair of counterfactual instances in Phase 2, although they were randomly sampled from a larger pool of task instances (i.e., 40 regular instances—5 for each feature combination, plus 5 pairs of counterfactual instances).

We included three attention check questions at different places throughout the experiment, in which participants were

instructed to select a pre-specified option. These attention check questions later helped us to exclude inattentive participants. Our experiment was only open to U.S. workers who had completed at least 1,000 HITs before and had an approval rate of at least 95%, and each worker could participate only once. The base payment of the experiment was $2.00. To incentivize participants to carefully deliberate on how to bid in each task, at the end of our experiment, we revealed to participants the outcome of their bids and the total rewards they earned as travel agents, and we converted the rewards to actual bonus payments using a conversion ratio of 200:1.

4 Data and Methods

We collected response data from 678 participants in total. The median time a participant spent on our HIT was 15.8 minutes, leading to a median hourly wage of $12.5. We considered a participant as inattentive if they failed on any attention check question, or their bid prices were less than $20 on any of the 20 task instances[5]. After excluding inattentive participants, we retained valid data from 459 participants (see supplemental materials for participant demographics), which we utilized to perform our analyses and address our research questions.

4.1 Measurements

Based on the experimental data we collected, we defined a few metrics to quantify the fairness of humans' decisions, in terms of both the decision *outcome* and the decision *process*.

Demographic disparity. We used demographic disparity to evaluate the fairness level of humans' decision outcomes. Specifically, the demographic disparity of a participant k's decisions on a set of N tasks can be defined as $DD_k = \sum_{Race_i=black} \hat{b_i^k}/N_{black} - \sum_{Race_i=white} \hat{b_i^k}/N_{white}$, where $\hat{b_i^k}$ is the normalized bid price participant k made on listing i (i.e., their original bid price b_i^k divided by the client's budget on that listing), while N_{black} (or N_{white}) is the number of listings in the N tasks for which the host's race is Black (or White). When a participant was not biased towards either Black or White hosts in deciding their bid prices, the value of DD_k would be close to zero. On the other hand, $DD_k < 0$ (or $DD_k > 0$) suggests that participant k tended to offer lower (or higher) bid prices on listings provided by Black hosts than on listings provided by White hosts.

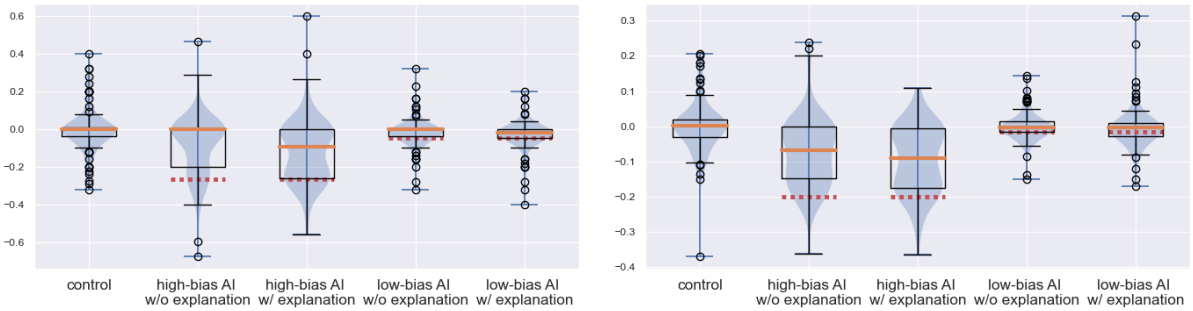
AI influence disparity. Next, to gain more insights into *why* humans' decision outcomes across racial groups were fair or unfair during their usage of the AI model, we further measured the fairness level of humans' decision processes. In particular, we focused on evaluating how fairly participants interacted with the AI model in Phase 1 across listings provided by hosts of different races. Following prior work [Green and Chen, 2019b], we quantified the extent to which the AI model influenced participants' decisions by comparing the bid prices made by participants who were given the AI model's predictions with the bid prices made by those who were not given the AI model's prediction on the same task instances. That is, we defined I_i^k as the "AI influence" on participant k for listing i, and $I_i^k = \frac{b_i^k - c_i}{a_i - c_i}$, where b_i^k is participant k's bid price on listing i after seeing the AI model's prediction a_i on the host's asking price, while c_i is the average bid price on listing i made by participants in the *control* treatment. This metric was similar to the "*weight of advice*" metric widely used in the advice-taking literature [Yaniv, 2004], and it measured how much participant altered their decisions when presented with the AI's advice—$I_i^k = 0.5$ means that participant k equally weighed their independent bid price and the AI's advice to make their final bid on listing i, while $I_i^k < 0.5$ (or $I_i^k > 0.5$) suggests participant k's final bid price on listing i was closer to their independent bid price (or the AI's advice).

Given the quantification of AI influence, we then used AI influence disparity to measure whether participants were influenced by the AI model to a similar degree between listings provided by Black/White hosts. In particular, we first separated the task instances in Phase 1 into two sets based on whether the AI model's predicted host's asking price a_i was greater or smaller than c_i. Within the set of tasks where $a_i > c_i$ (i.e., the AI model attempted to pull participants towards higher bids), the AI influence disparity for participant k can be defined as $AID_k^{\text{AI>control}} = \text{mean}\{I_i^k | \forall i, \textit{Race}_i = \textit{black}, a_i > c_i\} - \text{mean}\{I_i^k | \forall i, \textit{Race}_i = \textit{white}, a_i > c_i\}$. Thus, when $AID_k^{\text{AI>control}} < 0$ (or $AID_k^{\text{AI>control}} > 0$), it means that participant k was less (or more) strongly influenced by the AI model to increase their bid prices on listings provided by Black hosts than on listings provided by White hosts. Similarly, within the set of tasks where $a_i < c_i$ (i.e., the AI model attempted to pull participants towards lower bids), the AI influence disparity for participant k can be defined as $AID_k^{\text{AI<control}} = \text{mean}\{I_i^k | \forall i, \textit{Race}_i = \textit{black}, a_i < c_i\} - \text{mean}\{I_i^k | \forall i, \textit{Race}_i = \textit{white}, a_i < c_i\}$. An $AID_k^{\text{AI<control}}$ value that is below (or above) zero implies that participant k was less (or more) strongly influenced by the AI model to decrease their bid prices to Black hosts than to White hosts.

4.2 Statistical Methods

We first performed normality tests to all of the measurements and found that none of them were normally distributed. Thus, we conducted Aligned Rank Transform (ART) ANOVA [Wobbrock *et al.*, 2011], a non-parametric approach to factorial ANOVA, to analyze the data. We start by comparing the fairness level of humans' decision outcomes (i.e., the demographic disparity) across different treatments. As discussed earlier, our experiment had a 2×2 + 1 (control) design. Following recommendations on analyzing experimental data when the control treatment does not fit into the factorial design [Himmelfarb, 1975], we first conducted one-way ART ANOVA to examine if any significant differences in demographic disparity exist across all treatments. Then, we conducted two-way ART ANOVA on the data obtained from all but the control treatment to understand how the bias level of the AI model and the existence of AI explanations affect demographic disparity. These analyses were con-

[5]We chose $20 as the threshold to determine outlier bids since bid prices less than $20 consisted of the lowest 5% of all bid prices we obtained. We didn't consider a highest 5% outlier threshold since participants' bid prices were bounded by the client's budgets.



(a) Demographic Disparity, Phase 1 (counterfactual tasks)

(b) Demographic Disparity, Phase 1 (all tasks)

Figure 2: Violin plots with boxplots for demographic disparity comparisons in Phase 1. Orange lines represent the median values. Red dotted lines represent the demographic disparity values computed for the AI model used in the corresponding treatments.

ducted for Phases 1 and 2 separately, each time both within the pair of counterfactual instances only and on all task instances in the phase. We then compared the fairness level of humans' interactions with the AI model (i.e., the AI influence disparity) using two-way ART ANOVA tests on the data obtained from all but the control treatment. This analysis was conducted only for Phase 1 (since participants only had access to the AI model in Phase 1), and within subsets of tasks where the AI model's price prediction was higher or lower than the baseline bid price separately.

5 Results

5.1 Effects on Outcome Fairness in Phase 1

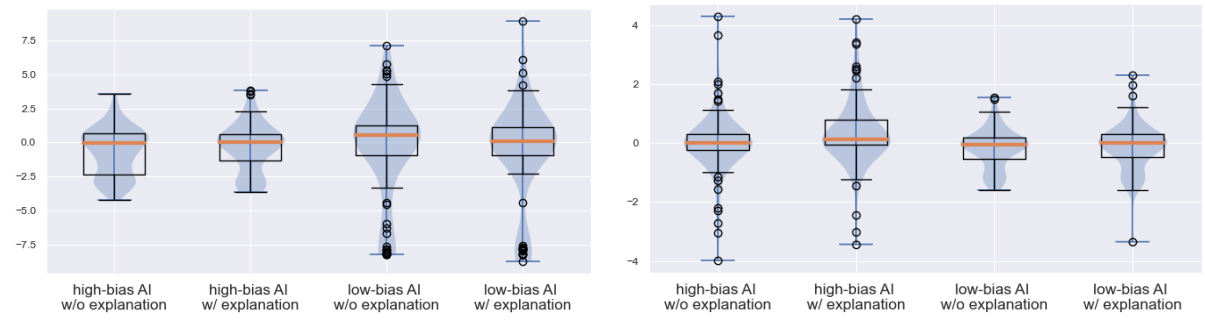
We start by examining **RQ1a**, how the AI biases and explanations affect the fairness in humans' decision outcomes *during* their usage of the AI model, i.e., in Phase 1. We first compare the demographic disparity in participants' bid prices on Phase 1 counterfactual instances—a pair of listings with very similar characteristics but are provided by Black/White hosts respectively—and we visualize the differences across treatments using violin plots with boxplots in Figure 2(a). Visually, it is clear that in our tasks, the bids that participants who had no access to AI models (i.e., in the control treatment) or received assistance from a low-bias AI model made were relatively unbiased, while participants assisted by a high-bias AI model tended to bid lower to Black hosts than to White hosts and thus had more negative demographic disparity values[6]. Results of the one-way ART ANOVA suggest a significant difference in demographic disparity across all treatments ($p < 0.001$). A post-hoc pairwise comparison with Bonferroni adjustment indicates the significant differences are between the control treatment and the two high-bias AI treatments (vs. high-bias w/o explanation: $p = 0.006$; vs. high-bias w/ explanation: $p < 0.001$), and between the high-bias AI w/ explanation treatment and the two low-bias AI treatments (vs. low-bias w/o explanation: $p < 0.001$; vs. low-bias w/ explanation: $p = 0.011$). Figure 2(b) shows similar results on the comparisons of demographic disparity across all tasks in Phase 1. Again, the one-way ART ANOVA suggests a significant difference across treatments ($p < 0.001$), with the pairwise comparisons suggesting the differences between the control and the two high-bias AI treatments, and between the two high-bias AI and the two low-bias AI treatments are all significant at the level of $p < 0.001$.

When examining how the AI bias level and AI explanations affect the demographic disparity of participants' bids to Black/White hosts, our two-way ART ANOVA results on both the counterfactual instances and all task instances show that the higher level of bias in the AI model's predictions results in higher demographic disparity (counterfactuals: low-bias Mdn.$= -0.008$ vs. high-bias Mdn.$= -0.032$, $p < 0.001$; all: low-bias Mdn.$= -0.004$ vs. high-bias Mdn.$= -0.082$, $p < 0.001$). Surprisingly, we also detect a significant main effect of the AI explanation on demographic disparity (counterfactuals: w/o explanation Mdn.$= 0$ vs. w/ explanation Mdn.$= -0.036$, $p = 0.032$; all: w/o explanation Mdn.$= -0.011$ vs. w/ explanation Mdn.$= -0.023$, $p = 0.025$). This means that the provision of AI explanations actually made participants even *more* biased against Black hosts when making the bidding decisions, potentially as they made participants "justify" the bias of the AI model as capturing hidden patterns in the data more than raising participants' awareness of the AI model's biased predictions.

5.2 Effects on Interaction Fairness in Phase 1

We then move on to answer **RQ1b**, whether the AI biases and explanations affect the fairness in humans' interactions with the AI model in Phase 1. Figure 3(a) compares the AI influence disparity $AID_k^{\text{AI}>\text{control}}$ across the four treatments with AI models, within the subset of tasks where the AI model's predicted price is higher than the average bid price made by participants in the control treatment (i.e., $a_i > c_i$). A two-way ART ANOVA test suggests a significant main effect of the AI bias level (low-bias Mdn.$= 0.166$ vs. high-bias Mdn.$= 0$, $p = 0.035$). This suggests that when the AI bias

[6]Note that when comparing the demographic disparity of participants' bidding decisions to that of the AI model's predictions (shown as the red dotted lines in Figure 2), we find that participants' bid prices tended to be *less* biased towards the Black hosts compared to the AI model. This is consistent with findings in Fogliato *et al.* [2022] and implies that humans have some ability in correcting the AI model's bias in their AI-assisted decisions.



(a) AI Influence Disparity, AI > Control

(b) AI Influence Disparity, AI < Control

Figure 3: Violin plots with boxplots for AI influence disparity comparisons in Phase 1. Orange lines represent the median values.

level is low, participants are more strongly influenced by the AI model to increase their bid prices to Black hosts than to White hosts, but this is no longer the case when the AI model becomes more biased against Black hosts in general.

Interestingly, the inverse pattern is found for tasks where the AI model's predicted price is lower than the average bid price made by participants in the control treatment (i.e., $a_i < c_i$, see Figure 3(b)). Here, a significant main effect of AI bias (low-bias Mdn.$= -0.004$ vs. high-bias Mdn.$= 0.092$, $p < 0.001$) and a marginally significant main effect of AI explanation (w/o explanation Mdn. $= 0$ vs. w/ explanation Mdn. $= 0.045$, $p = 0.061$) are found from the two-way ART ANOVA test on $AID_k^{\text{AI<control}}$. This means that when the AI bias level increases, especially when provided with the AI explanations, participants are more strongly influenced by the AI model to decrease their bid prices when making bids to Black hosts than when making bids to White hosts.

Together, these results provide evidence that higher level of AI bias and the provision of AI explanations may result in participants' unfair bid decisions across racial groups partly because they lead participants to engage in "*disparate interactions*" with the AI model, i.e., participants respond to the AI model's predictions in a biased way that disproportionately result in lower bids to Black hosts than to White hosts.

5.3 Effects in Phase 2

Finally, we examine whether AI biases and explanations affect the fairness of humans' decisions, *after* they have had the experience of being assisted by the AI model and return to make independent decisions (**RQ2**). We again use ART ANOVA to compare the demographic disparity of participants' bid prices on both the counterfactual instances and all task instances in Phase 2. This time, we do not detect any significant difference across all five treatments (see supplemental materials for figures). This means that in our experiment, the AI model's impacts on human decision fairness are only limited to the duration of their interaction with the AI model. Humans do not seem to apply their observed patterns in AI decision making to their own decision making.

We provide two possible explanations on this null effect. Firstly, we did not provide participants with immediate performance feedback, potentially making participants unsure about the "quality" of the hidden patterns uncovered by the AI model. Secondly, participants only completed 10 tasks in Phase 1 of our experiment, and this limited exposure to the AI model may not be sufficient for learning to occur.

6 Conclusions and Discussions

In this paper, we ask the question of how AI biases and explanations impact human decision fairness during and after the usage of AI-based decision aids. Via an experimental study on a real-world business decision making setting, we find that both higher level of AI bias and the presence of AI explanation can result in more unfair AI-assisted human decisions across racial groups, partly as humans are influenced by the AI recommendations across racial groups in a biased way. However, the fairness level of humans' independent decisions after their usage of the AI model is not impacted by their previous interactions with AI.

Our results highlight the importance of thoroughly examining the fairness properties of AI-based decision aids, and perhaps explicitly mitigating their bias level, before providing them to human users to promote human decision fairness. More research should also be conducted to understand how AI explanations can be better designed to more effectively expose the internal biases of AI models , and support humans to interact with the AI models fairly and make fairer decisions.

There are a few limitations in our current study. For example, since our study is conducted in a specific decision making context (i.e., bidding in rental housing markets) with relatively low stakes and humans are relatively unbiased in their independent judgements, we caution readers to not overgeneralize the results to other significantly different contexts. Also, we had a limited number of Black participants in our study. We recommend future studies to target marginalized subgroups for further investigation by aiming for a balanced sample of White and Black participants. Our results from a short online experiment may not generalize to scenarios involving persistent and prolonged biased AI exposure in the field. We hope this work can inspire more future studies to systematically examine how human decision makers' own biases, AI models' biases and presentations, and the biased interactions between humans and AI, together, impact decision fairness in the real-world AI-assisted decision making.

Ethical Statement

This study was reviewed and approved by Purdue University Institutional Review Board.

Acknowledgments

We thank the support of the National Science Foundation under grant IIS-1850335 on this work. We thank Chun-Wei Chiang, Xiaoni Duan, Altug Gemalmaz, Zhuoyan Li, Zhuoran Lu, Hasan Mahmood, and Amy Rechkemmer for their help in data annotation, and thank all reviewers for their feedback.

References

[Agarwal *et al.*, 2019] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR, 2019.

[Angerschmid *et al.*, 2022] Alessa Angerschmid, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. Fairness and explanation in ai-informed decision making. *Machine Learning and Knowledge Extraction*, 4(2):556–579, 2022.

[Angwin *et al.*, 2016] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. propublica, may 23, 2016.

[Bansal *et al.*, 2021] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[Becker *et al.*, 1964] Gordon M Becker, Morris H DeGroot, and Jacob Marschak. Measuring utility by a single-response sequential method. *Behavioral science*, 9(3):226–232, 1964.

[Bertrand and Duflo, 2017] Marianne Bertrand and Esther Duflo. Field experiments on discrimination. *Handbook of economic field experiments*, 1:309–393, 2017.

[Binns *et al.*, 2018] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. '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*, pages 1–14, 2018.

[Bird *et al.*, 2020] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

[Buçinca *et al.*, 2021] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[Chiang and Yin, 2022] Chun-Wei Chiang and Ming Yin. Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In *27th International Conference on Intelligent User Interfaces*, pages 148–161, 2022.

[Des Jarlais *et al.*, 2018] DC Des Jarlais, HLF Cooper, Kamyar Arasteh, J Feelemyer, Courtney McKnight, and Z Ross. Potential geographic" hotspots" for drug-injection related transmission of hiv and hcv and for initiation into injecting drug use in new york city, 2011-2015, with implications for the current opioid epidemic in the us. *PloS one*, 13(3):e0194799, 2018.

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

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

[Duan *et al.*, 2020] Xiaoni Duan, Chien-Ju Ho, and Ming Yin. Does exposure to diverse perspectives mitigate biases in crowdwork? an explorative study. In *Proceedings of the aaai conference on human computation and crowdsourcing*, volume 8, pages 155–158, 2020.

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

[Ehsan *et al.*, 2021] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509*, 2021.

[Fogliato *et al.*, 2022] Riccardo Fogliato, Maria De-Arteaga, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of misestimated algorithmic scores. *Available at SSRN 4050125*, 2022.

[Gemalmaz and Yin, 2022] Meric Altug Gemalmaz and Ming Yin. 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*, pages 295–306, 2022.

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

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

[Himmelfarb, 1975] Samuel Himmelfarb. What do you do when the control group doesn't fit into the factorial design? *Psychological Bulletin*, 82(3):363, 1975.

[Hu *et al.*, 2020] Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K Thiruvathukal, and Ming Yin. Crowdsourcing detection of sampling biases in image datasets. In *Proceedings of The Web Conference 2020*, pages 2955–2961, 2020.

[Iacus *et al.*, 2012] Stefano M Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24, 2012.

[Lai and Tan, 2019] Vivian Lai and Chenhao Tan. 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*, pages 29–38, 2019.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[Ma *et al.*, 2015] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[Peng *et al.*, 2019] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. What you see is what you get? the impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 125–134, 2019.

[Peng *et al.*, 2022] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. Investigations of performance and bias in human-ai teamwork in hiring. *arXiv preprint arXiv:2202.11812*, 2022.

[Poursabzi-Sangdeh *et al.*, 2021] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

[Saxena *et al.*, 2019] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, 2019.

[Schaffer *et al.*, 2019] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your ai: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 240–251, 2019.

[Serengil and Ozpinar, 2021] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021.

[Srivastava *et al.*, 2019] Megha Srivastava, Hoda Heidari, and Andreas Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2459–2468, 2019.

[Verma and Rubin, 2018] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[Wang and Yin, 2021] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

[Wang *et al.*, 2020] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 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*, pages 1–14, 2020.

[Wobbrock *et al.*, 2011] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 143–146, 2011.

[Yaniv, 2004] Ilan Yaniv. Receiving other people's advice: Influence and benefit. *Organizational behavior and human decision processes*, 93(1):1–13, 2004.

[Zhang *et al.*, 2020] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

[Zhang *et al.*, 2021] Shunyuan Zhang, Nitin Mehta, Param Vir Singh, and Kannan Srinivasan. Frontiers: Can an artificial intelligence algorithm mitigate racial economic inequality? an analysis in the context of airbnb. *Marketing Science*, 40(5):813–820, 2021.