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

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

1 Evaluations of the AI Models Used in the Experiment

In our experiment, we followed the fair regression algorithm to train two AI models with different levels of bias to be used in our experiment. To adapt the label to a $[0, 1]$ range and thus utilize the fair regression algorithm, we first took the log of the original price of each Airbnb listing and used the min-max normalization. The resulting values were then discretized into 100 levels: $\widetilde{\mathcal{Y}} = \{0.01, 0.02, ..., 0.99, 1\}$. Table 1 reports the performance and bias level of the two AI models we obtained, when evaluating on the test dataset. Note that SP disparity is defined as $\max_{a,z} |\mathbb{P}[f(X) \geq z|A = a] - \mathbb{P}[f(X) \geq z]|$; the larger the SP disparity value, the more biased the AI model.

Metric	**low-bias AI** ($\epsilon = 0.005$)	**high-bias AI** ($\epsilon = 1$)
Mean abs. percentage error	0.345	0.313
SP disparity	0.030	0.130
Avg. price (Black, all)	\$154.2	\$124.5
Avg. price (White, all)	\$157.9	\$169.5
Avg. price (Black, matched)	\$156.1	\$126.4
Avg. price (White, matched)	\$142.3	\$151.0

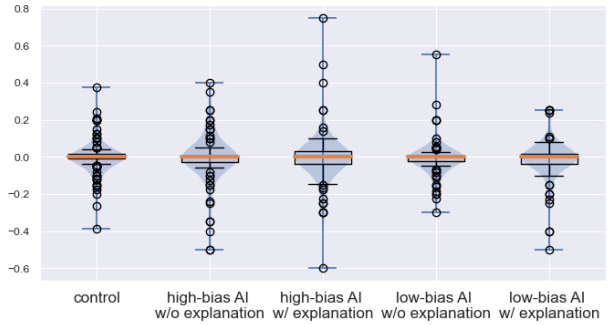
Table 1: Evaluation results of the performance and bias level of the two AI models on the test dataset.

2 Participant Demographics

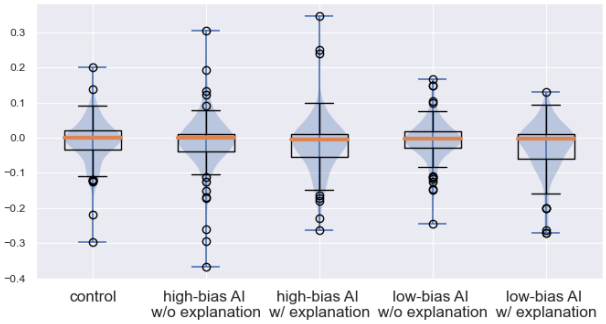
58.4% of our participants were self-reported as male, and 41.2% self-reported as female. The average age of our participants was 35.9. In addition, our participant sample had a predominantly white population, with 86% of participants self-identified as White and 7% as Black, although this reflects the general workforce on Amazon Mechanical Turk and the U.S. workforce[1]. A balance check reveals *no* significant difference in demographic composition across treatments.

3 Effects in Phase 2

Figure 1 compares the demographic disparity values for participants' bid decisions in Phase 2 on Black/White hosts across all five treatments, after they have had experience of being assisted by the AI model and return to make independent decisions. We do not detect any significant difference across all five treatments or any main effects of the AI bias level or the provision of AI explanations on the fairness level of participants' decisions. This means that the AI model's impacts on humans' decision fairness are not extended beyond the period of time that humans interact with the AI model. In other words, humans do not appear to apply their observed patterns in AI decision making to their own decision making.



(a) Demographic Disparity, Phase 2 (Counterfactual Tasks)



(b) Demographic Disparity, Phase 2 (All Tasks)

Figure 1: Violin plots with boxplots for demographic disparity comparisons in Phase 2. Orange lines represent the median values.

[1]https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/