The consequences of AI training on human decision making

Lauren S. Treiman[a,2], **Chien-Ju Ho**[a,b,1], **and Wouter Kool**[a,c,1]

This manuscript was compiled on July 21, 2024

Artificial intelligence (AI) is now an integral part of everyday decision making, assisting us in both routine and high-stakes choices. These AI models often learn from human behavior, assuming this training data is unbiased. However, we report five studies that show that people change their behavior to instill desired routines into AI, indicating this assumption is invalid. To show this behavioral shift, we recruited participants to play the ultimatum game, where they were asked to decide whether to accept proposals of monetary splits made by either other human participants or AI. Some participants were informed their choices would be used to train an AI proposer, while others did not receive this information. Across five experiments, we found that people modified their behavior to train AI to make fair proposals, regardless of whether they could directly benefit from the AI training. After completing this task once, participants were invited to complete this task again but were told their responses would not be used for AI training. People who had previously trained AI persisted with this behavioral shift, indicating that the new behavioral routine had become habitual. This work demonstrates that using human behavior as training data has more consequences than previously thought since it can engender AI to perpetuate human biases and cause people to form habits that deviate from how they would normally act. Therefore, this work underscores a problem for AI algorithms that aim to learn unbiased representations of human preferences.

Human-AI Interaction | AI Training | Ultimatum Game | Fairness | Decision Making

Significance Statement

In recent years, people have become more reliant on AI to help them make decisions. These models not only help us but also learn from our behavior. Therefore, it is important to understand how our interactions with AI models influence them. Current practice assumes that the human behavior used to train is unbiased. However, our work challenges this assumption. We show that people change their behavior when they are aware it is used to train AI. Moreover, this behavior persists days after training has ended. These findings highlight a problem with AI development: assumptions of unbiased training data can lead to unintentionally biased models. This AI may reinforce these habits, resulting in both humans and AI deviating from optimal behavior.

Artificial intelligence (AI) plays an increasingly important role in everyday decision making. It is used not only by social media and streaming services to provide recommendations but also in more crucial contexts including patient care (1–4), the judicial system (5–7), and policymaking (8, 9). Most of these models learn how to make decisions from human behavior. One important implicit assumption underlying such training is that the observed choice data is unbiased (10). However, when people are aware their behavior is used to train AI, they might deviate from how they would normally act (11, 12). For example, they may deliberately change their behavioral policy to instill desired behaviors in the algorithm. This behavioral shift would pose a fundamental problem for AI algorithms that aim to learn unbiased representations of human decision making. Therefore, we sought to investigate how humans modify their behavior when they are aware they are training AI.

It is well-established that humans act differently when interacting with AI systems (13–15), displaying less socially desirable traits (16) and becoming more prone to cheat (17). In fact, humans are willing to incur a cost to avoid interacting with AI altogether (18). These findings demonstrate that people are sensitive to the presence of AI systems, but they do not reveal whether people alter their behavior if they are aware that it is used for AI training.

Nevertheless, humans are characterized by their ability for goal-directed behavior. Psychological science is replete with demonstrations of how we exploit task structure to our advantage (19–22). Computer scientists study similar principles from a reverse perspective, focusing on designing AI systems that consider users' responses to AI deployment (23–25). For example, Goodhart's Law, originating from economic and financial policy making, predicts that when AI systems set a particular measure as a target, people change their behavior to "game the system" by focusing on that target (26). A notable example of Goodhart's law comes from Camacho and Conover (27), who showed that, as the rules of social welfare distribution became known, people in Colombia reported exaggerated financial needs so that they just reached the threshold to qualify for aid. Thus, if humans become aware AI learns from their behavior, then they may start using (intuitive) internal models of the algorithm's learning rules to strategically instill behavior that benefits them.

This change in behavior may persist beyond AI training. Research from experimental psychology has shown that behavior initially implemented to pursue

Author affiliations: [a]Division of Computational and Data Sciences, Washington University in St. Louis, St. Louis, MO 63130; [b]Division of Computer Science & Engineering, Washington University in St. Louis, St. Louis, MO 63130; [c]Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO 63130

[1]C.H. (Chien-Ju Ho) contributed equally to this work with W.K. (Wouter Kool)

[2]To whom correspondence should be addressed. E-mail: ltreiman@wustl.edu

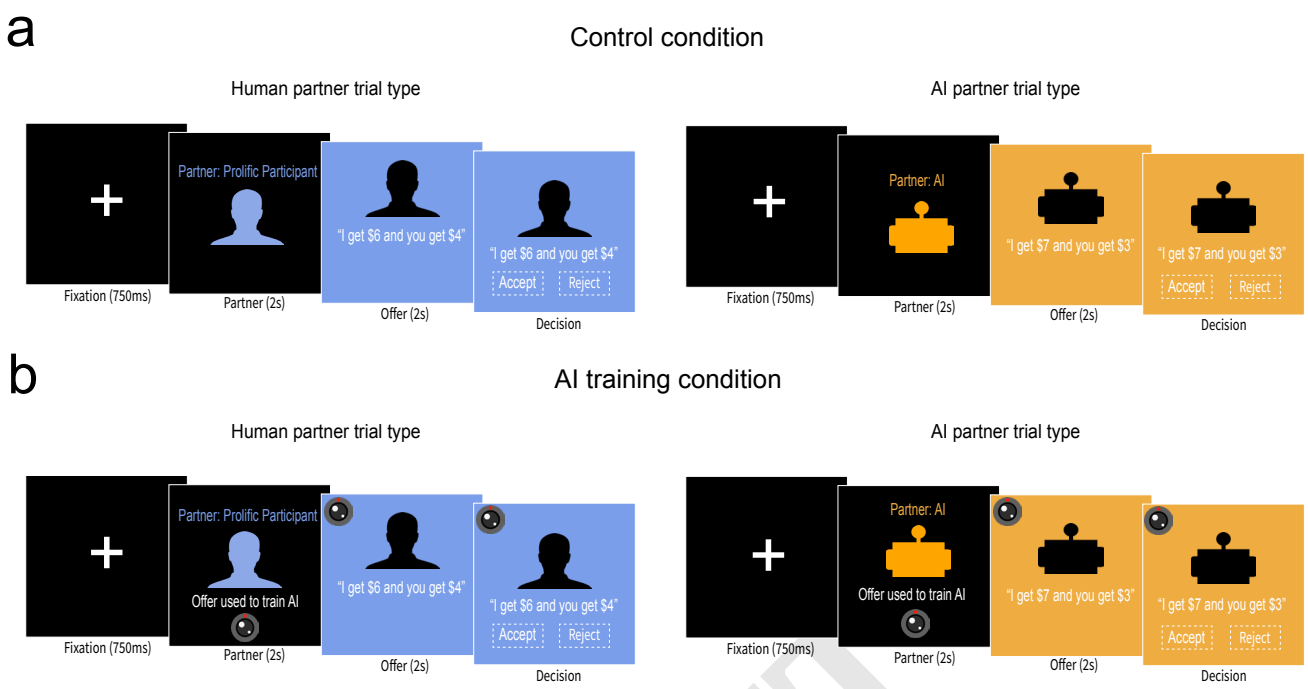

Fig. 1. Example trials for the control (a) and AI training (b) conditions for each partner type (left human participant and right AI). For participants assigned to the AI training condition, a webcam was shown to remind the participants that their responses were training AI. Participants in the control condition did not see a webcam since their responses were not training AI. With the exception of the webcam, the trial format was the same for both conditions. On each trial, participants first saw a fixation cross (750ms) to indicate the start of the trial. Next, they saw the partner type (human or AI) for 2 seconds. They then saw the offer amount for 2 seconds before they could make a choice. Participants had unlimited time to choose. Each participant made multiple choices with varying partner types and offer amounts. Only training condition was varied between participants.

a goal will eventually become habitually engrained (28, 29). A hallmark characteristic of such habits is that they persist even when the environment has changed in such a way that they become costly (i.e., reinforcer devaluation; (30, 31)). In our case, habits would reveal themselves when the behavior initially used to train AI is implemented in the absence of AI training. This would be problematic for machine learning algorithms that are designed to learn people's unbiased preferences.

Here, we aimed to determine whether humans modulate their behavior when they know it will be used to train AI and whether these changes persist after AI training. To do this, we used the ultimatum game (32). In this game, two players allocate a sum of money. One player, the proposer, divides the money, and the other player, the responder, decides to accept or reject it. Even though rational responders should accept any nonzero offer, behavior on this task shows that people are prone to reject 'unfair' offers (e.g., below 30% of the total), foregoing monetary rewards (33–35). In other words, the ultimatum game measures how subjective fairness affects decision making (36). We used this feature of the task to test our main hypothesis, predicting that people are less likely to accept unfair offers when their behavior trains an AI proposer.

Our studies* show that people are willing to incur costs to train an AI system to make fair offers, even when this does not result in increased personal gains. Importantly, this behavior persisted across several days and in the absence of AI training, suggesting that people form habits when training AI systems. Our work reveals an important blind spot for AI developers, who should account for these biases when designing algorithms that aid human choice (10). They also provide a novel, applied context in which goal-directed and habitual forms of control coordinate to guide decision making. Stimuli, data, and analysis scripts from all experiments can be found on the Open Science Framework (OSF)†.

Results

Across five preregistered experiments, we tested whether people modify their behavior in the ultimatum game (32) when informed their responses would be used to train an AI (Figure 1). In all of these experiments, we manipulated this "AI training" condition between groups of participants. Some participants were told their behavior would train an AI while other participants were not given this information. For both groups of participants, each round of this game involved a different partner. On some rounds, this was a participant recruited from another experiment. On other rounds, the partner was an AI algorithm. On each round, participants were shown the type of partner and their proposal on allocating $10 between the two players (offers ranging from $1 – $6). Participants then chose whether to accept or reject the offer. At the end of each session, one randomly selected proposal was resolved. If the participant accepted it, the money would be shared according to the proposal. If they rejected it, neither player received a reward.

*A preliminary version of this dataset was published in (37)

†Preregistration link found here: https://osf.io/b7w5c

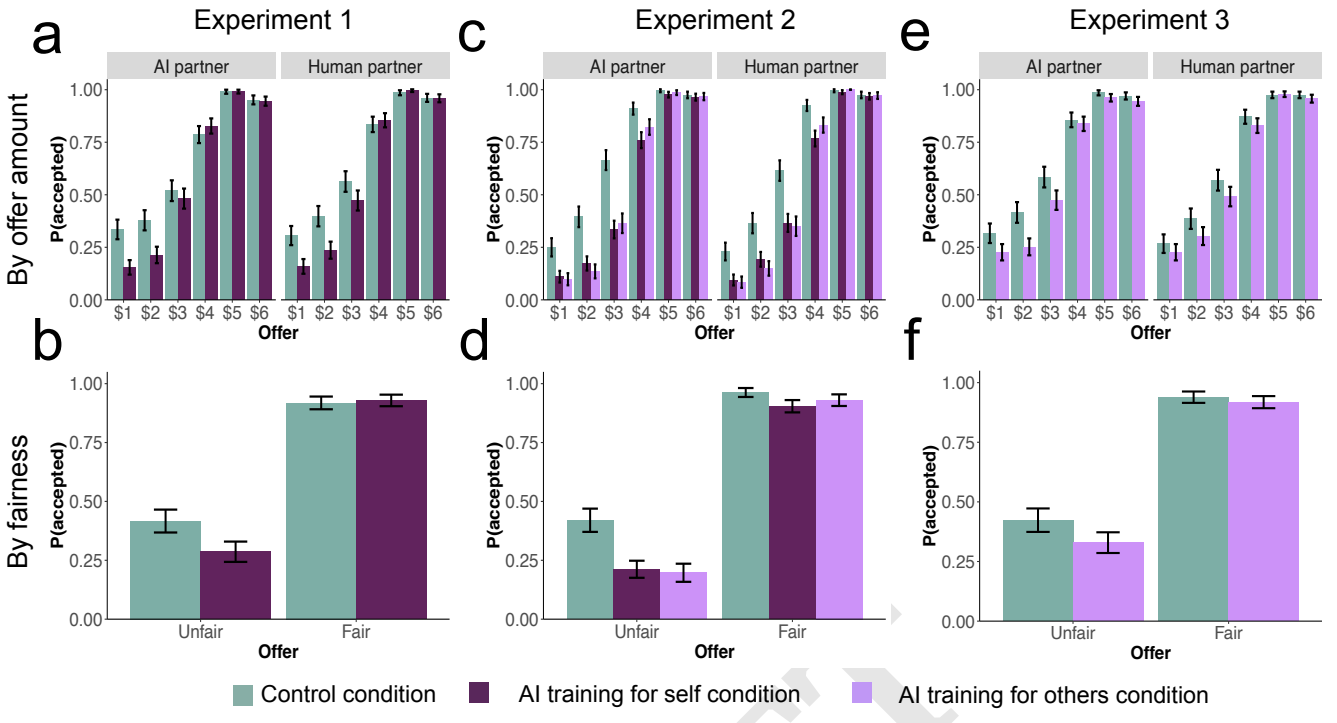

Fig. 2. Results for session 1 for Experiments $1-3$. All figures show the proportion of accepting an offer based on offer amount (top row) and by fairness (bottom row) conditioned on partner type and fairness. Error bars indicate standard error.

Humans forgo reward to train an AI to make fair offers. In Experiment 1[‡], some participants ($n = 110$) were informed that their responses would be used to train an AI they would encounter in a subsequent session ('AI training' condition) while others ($n = 103$) did not receive this information (control condition). In the AI training condition, each round started with a webcam icon and a line of text reminding participants their behavior would be used to train AI (see Figure 1b). Two days later, all participants were invited to complete a second session of the same task, where participants in the AI training condition were informed their responses would no longer train AI. To incentivize participation, the bonus rate was increased by 300% for this second session. The results from this second session are described at the end of the Results section.

We analyzed the data by modeling participants' probability of accepting an offer as a function of the dollar amount, the partner type, and the training condition using both mixed-effects models as well as ANOVAs. Full regression and ANOVA tables can be found in the Supplemental Materials.

A preregistered logistic mixed-effects model revealed that participants were more likely to accept offers with larger dollar amounts ($b = 1.87,\ \mathrm{SE} = 0.06,\ p < 0.001$), replicating prior findings (38–40). However, they did not respond differently when partnered with a human compared to AI ($b = -0.11,\ \mathrm{SE} = 0.05,\ p = 0.051$). Most importantly, even though we found no main effect of training condition ($b = -0.37,\ \mathrm{SE} = 0.20,\ p = 0.064$), there was a significant interaction between training condition and offer amount ($b = 0.17,\ \mathrm{SE} = 0.05,\ p = 0.002$). As can be seen in Figure 2a and Figure 2b, participants in the AI training condition were more sensitive to the offer amount than participants in the control condition, particularly for unfair offers ($\leq \$3$). No other interaction effect was significant ($ps \geq 0.22$).

A preregistered ANOVA provided evidence for this interpretation. Here, we found main effects of fairness ($F_{1,211} = 592,\ p < 0.001$) and training condition ($F_{1,211} = 4.23,\ p = 0.04$), but these effects were qualified by an interaction between these terms ($F_{1,211} = 8.97,\ p = 0.003$). Specifically, participants in the AI training condition were more likely to reject unfair offers compared to participants in the control condition ($t_{196} = 2.62,\ p = 0.01$), but no difference was found for fair offers ($t_{200} = -0.55,\ p = 0.58$). The ANOVA did not show a main effect of partner type ($F_{1,211} = 1.82,\ p = 0.18$) or any additional interactions ($ps \geq 0.52$).

Our first experiment showed that participants rejected more unfair offers in the AI training condition. This result indicates that people are willing to forgo monetary reward to train an AI system to make fairer proposals. However, it does not reveal the motivation behind this change in behavior. While participants may have been motivated by an intrinsic motivation to increase fairness, it's also possible that they rejected more unfair offers to increase their rewards in the second session (where they would encounter the AI they trained).

Humans incur costs to train a fair AI for other people. We designed Experiment 2[§] to distinguish between these explanations. This experiment followed the same procedure as Experiment 1 (Figure 1), but we introduced a third AI

[‡] Preregistration link for session 1 found here: https://osf.io/ajxk4

[§] Preregistration link for session 1 found here: https://osf.io/krhz9

training condition. In this new condition, participants were informed their responses would train an AI they wouldn't encounter but others would face in the second session ('AI training for others' condition; $n = 107$). They were not explicitly told they would face an AI trained by someone else in the second session. By comparing behavior in this new condition to a replication of the original AI training condition (now 'AI training for self' condition) ($n = 127$) and the control condition ($n = 100$), we could test whether people would still be willing to train an AI to be fair for only altruistic motivation.

The results of this new experiment were clear. In short, participants in the new 'AI training for others' condition showed a similar willingness to incur a monetary cost to train AI to be fair.

A preregistered logistic mixed-effects model revealed a main effect of offer amount ($b = 2.25$, SE $= 0.11$, $p < 0.001$). Additionally, participants in both the AI training for self condition($b = -1.92$, SE $= 0.43$, $p < 0.001$) and the AI training for others condition ($b = -1.76$, SE $= 0.45$, $p < 0.001$) were more likely to reject proposer offers than those in the control condition. However, two significant interaction effects between offer amount and both the AI training for self ($b = 0.30$, SE $= 0.14$, $p = 0.03$) and the AI training for others ($b = 0.55$, SE $= 0.16$, $p < 0.001$) showed that participants in the training conditions were particularly punitive for lower offers than those in the control condition. This pattern of behavior is shown in Figure 2c and Figure 2d. An additional mixed-effects model indicated that participants in the AI training conditions accepted offers similarly ($b = -0.16$, SE $= 0.41$, $p = 0.71$) and were comparably sensitive to the offer amounts ($b = -0.24$, SE $= 0.15$, $p = 0.11$). There was no main effect of partner type ($b = 0.07$, SE $= 0.10$, $p = 0.51$). No other interaction effects were significant ($ps \geq 0.22$).

The results from a preregistered ANOVA were consistent with this interpretation. We found a significant interaction between training condition and offer fairness ($F_{2,331} = 11.43$, $p < 0.001$). Specifically, compared to the control condition, unfair offers were less likely to be accepted by participants in both the AI training for self condition ($t_{193} = 4.67$, $p < 0.001$) and AI training for others condition ($t_{187} = 4.99$, $p < 0.001$). There was no statistical difference in acceptance rates between participants in the two AI training conditions ($t_{230} = -0.39$, $p = 0.69$).

Interestingly, we found similar results for fair offers. Participants in both the AI training for self condition ($t_{201} = 3.55$, $p < 0.001$) and AI training for others condition ($t_{179} = 2.08$, $p = 0.04$), accepted fewer fair offers than the control condition, but there were no differences between training conditions ($t_{232} = 1.34$, $p = 0.18$). This finding stands in contrast to Experiment 1, where we found no difference in acceptance rates for fair offers. We believe this is driven by responses to \$4 offers. As shown in Figure 2c, participants in the control condition were more likely to accept \$4 offers compared to either AI training condition, but there was no difference for higher offers. Regardless of this effect, the observed interaction indicated that the effect of training condition was stronger for unfair offers compared to fair offers, which is consistent with our hypothesis. The ANOVA found no main effect of partner type ($F_{1,331} = 0.011$, $p = 0.91$) nor other significant interactions ($ps \geq 0.16$)

In addition to replicating the findings from Experiment 1, these results show that participants were willing to incur a personal cost to train an AI to make more fair offers even if they couldn't directly benefit. Strikingly, participants that trained an AI for others responded no differently than those who trained an AI for themselves. These results are consistent with the idea that people are motivated to train AI to promote fairness. However, it's also possible that they did so for reciprocal reasons (41–43): people may have only trained an AI to be fair because they assumed other participants were doing the same for them.

Humans are willing to train fair AI in the absence of personal benefits. We designed Experiment 3[¶] to test whether people would still be willing to train AI to be fair even if they could not personally benefit in future sessions. This experiment closely followed the design of Experiment 2 (Figure 1), except there was only the AI training for others condition (now referred to as 'AI training' condition) ($n = 117$) and control condition ($n = 101$). The key change, however, was that we removed the second session, eliminating the possibility of anyone benefiting from AI training in the future. By removing the second session, we could determine whether people are genuinely motivated to train AI to be fair.

A preregistered logistic mixed-effects model once again showed an increase in acceptance rates with higher offer amounts ($b = 1.99$, SE $= 0.07$, $p < 0.001$) but no effect of partner type ($b = -0.02$, SE $= 0.06$, $p = 0.70$). More importantly, we found a main effect of training condition ($b = -0.59$, SE $= 0.24$, $p = 0.014$), suggesting that participants accepted less offers in the AI training condition compared to the control condition. This main effect was qualified by an interaction with offer amount ($b = -0.19$, SE $= 0.06$, $p = 0.003$), replicating that participants in the AI training condition were more likely to reject unfair offers (Figure 2e and Figure 2f). No other interaction effects were significant ($ps \geq 0.17$).

A preregistered ANOVA provided results that were mostly consistent with this analysis. Even though there was no main effect of partner type ($F_{1,216} = 0.006$, $p = 0.94$), training condition ($F_{1,216} = 3.68$, $p = 0.056$), or their interactions with offer amount (partner type: $F_{1,216} = 0.626$, $p = 0.43$; training condition: $F_{1,216} = 2.04$, $p = 0.15$), we found a significant three-way interaction between these variables and partner type ($F_{1,216} = 4.40$, $p = 0.037$). Post-hoc two-way ANOVAs suggested that when playing against an AI, participants were more likely to reject offers in the AI training condition ($F_{1,216} = 5.43$, $p = 0.02$), even though this was not qualified by an interaction with fairness ($F_{1,216} = 3.60$, $p = 0.06$). When playing against a human, we found neither a main effect of training condition ($F_{1,216} = 2.02$, $p = 0.16$) nor an interaction effect of training condition and fairness ($F_{1,216} = 0.81$, $p = 0.37$).

These results suggest that participants were willing to incur personal costs to train an AI to make more fair offers, even if they were unable to personally benefit. This pattern of behavior reveals a genuine motivation to promote fairness in AI. We should note that there was some inconsistency in the results between our ANOVA and mixed-effects models. The mixed-effects model showed increased sensitivity to unfair

[¶] Preregistration link found here: https://osf.io/hp3b2

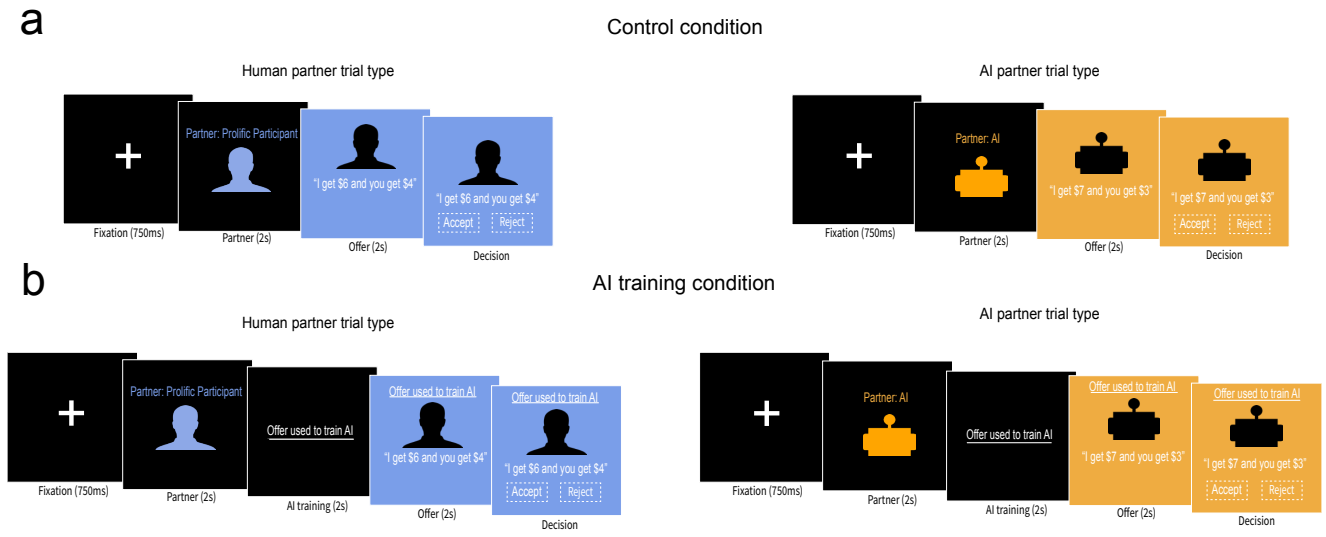

Fig. 3. Example trials for the control (a) and AI training (b) conditions for each partner type (left human participant and right AI). For participants assigned to the AI training condition, additional text was shown to remind the participants that their responses were training AI. Participants in the control condition did not see this text since their responses were not training AI. With the exception of the text appearing on an additional screen (2s) and when making a choice, the trial format was the same for both conditions. On each trial, participants first saw a fixation cross (750ms) to indicate the start of the trial. Next, they saw the partner type (human or AI) for 2 seconds. Participants in the AI training conditions then saw an additional screen reminding them of AI training for 2 seconds. They then saw the offer amount for 2 seconds before they could make a choice. Participants had unlimited time to choose. Each participant made multiple choices with varying partner types and offer amounts. Only training condition was varied between participants.

offers among participants in the AI training condition, but the ANOVA revealed this effect only for rounds that involved AI partners. While this latter pattern is intriguing, we place more confidence in the mixed-effects results. Not only are they consistent with our earlier findings, but this method of analysis predicts each individual response (unlike the ANOVA where a subject-wise averaging step occurs before conducting the analysis).

In these first three experiments, we found that people assigned to the AI training conditions rejected more unfair offers than those in the control conditions. We interpret these findings to suggest that people in the AI training conditions are more punitive due to their motivation to train AI for fairness. However, there is an alternative explanation for these results. It is possible that people in the AI training conditions rejected more unfair offers because they felt that they were observed when making choices (44). Specifically, participants in all AI training conditions saw an image of a webcam to remind them of AI training. Even though the webcam icon was only an image, it is possible that it made participants feel as if they were being watched. Consequently, participants may have been more self-aware in the AI training conditions (45), leading to more prosocial behavior and conforming to social norms (46–48). In other words, under this alternative explanation, the participants in our experiments were not deliberately trying to instill fairness into AI.

Humans train AI to be fair when no longer observed. To rule out the alternative explanation that the behavior change in the AI training conditions was only caused by the presence of the webcam image, we ran two new studies (Experiments 4 and 5). These studies were direct replications of Experiments 2 and 3[‖] except we did not use a webcam icon (Figure 3). Instead, to remind participants in the AI training condition (Experiment 4: AI training for others condition $n = 129$, AI training for self condition $n = 100$; Experiment 5: n = 105) that their responses would be used for AI training, they saw an additional screen with the text "Offer used to train AI" for two seconds. This text also appeared when participants were asked to make a choice. Participants in the control condition (Experiment 4: $n = 116$; Experiment 5: $n = 103$) completed the same task as before (Figure 3a). By removing the webcam icon, we could determine whether people change their behavior to train AI even if there are no visual cues suggesting they are being observed. We reasoned that if we replicated these results without the visual cue, we could more confidently claim that people changed their behavior to train AI.

In Experiment 4, we found that participants in both AI training conditions rejected more offers than those in the control condition (Figure 4a and Figure 4b), replicating the results of Experiment 2. A pre-registered logistic mixed-effects model showed that people in the AI training for self condition ($b = -1.17$, SE $= 0.45$, $p = 0.009$) and AI training for other condition ($b = -1.45$, SE $= 0.42$, $p < 0.001$) rejected more offers than those in the control condition. Although we found a main effect of offer amount ($b = 1.90$, SE $= 0.08$, $p < 0.001$), there were no significant interactions between offer amount and training condition when comparing the AI training for self condition ($b = 0.14$, SE $= 0.12$, $p = 0.24$) and AI training for others condition ($b = -0.055$, SE $= 0.10$, $p = 0.60$) to the control condition. However, from inspection of Figure 4a, we

[‖] We did not run a version of Experiment 1 without the webcam icon since Experiment 2 includes its conditions.

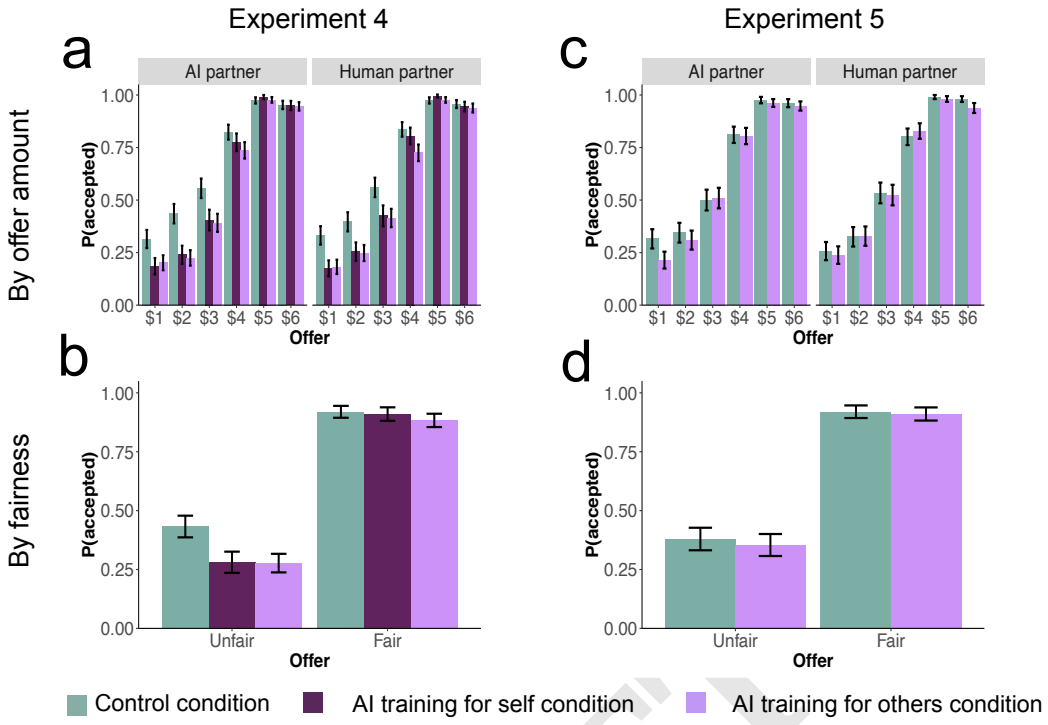

Fig. 4. Results for session 1 for Experiments 4 and 5. All figures show the proportion of accepting an offer based on offer amount (top row) and by fairness (bottom row) conditioned on partner type and fairness. Error bars indicate standard error.

hypothesize that the main effect of AI training was so strong for offers $\leq$ \$3 that the mixed effects model could not capture the interaction effects. Therefore, we ran an exploratory ANOVA and found both a main effect of training condition ($F_{2,342} = 6.15,\ p = 0.002$) and an interaction effect between fairness and training condition ($F_{2,342} = 4.87,\ p = 0.008$). Post hoc t-tests revealed that participants in the AI training for self condition ($t_{214} = 2.91,\ p = 0.004$) and AI training for others condition ($t_{230} = 3.18,\ p = 0.002$) rejected more unfair offers than those in the control condition. However, when considering only fair offers, there was no difference in acceptance rates when comparing the control condition to the AI training for self ($t_{213} = 0.45,\ p = 0.66$) and AI training for others ($t_{243} = 1.57,\ p = 0.12$), consistent with previous findings.

We ran an additional pre-registered mixed effects model to compare the acceptance rates between AI training conditions. The mixed effects model found no main effect ($b = 0.29,\ \text{SE} = 0.43,\ p = 0.51$) nor interaction effect between offer amount and training condition ($b = 0.19,\ \text{SE} = 0.11,\ p = 0.07$). Post hoc t-tests were consistent with this finding, as there was no difference in acceptance rates for unfair offers ($t_{211} = -0.08,\ p = 0.94$) and fair offers ($t_{227} = -1.24,\ p = 0.22$) between AI training conditions. We also found no effect of partner type ($b = -0.01,\ \text{SE} = 0.08,\ p = 0.89$) and no other interactions were significant ($ps \geq 0.54$).

We pre-registered the same analysis for Experiment 5. This study used the same procedure as Experiment 3, with participants in the AI training condition being informed that their behavior would train an AI for other participants but that they would not return for a follow-up session. As in Experiment 4, we did not use the webcam icon to remind people of the training condition, and used text ("Offer used to train AI") instead. The results were consistent with Experiment 3 (Figure 4c and 4d). Specifically, the mixed effects model showed an increase in acceptance rate as offer amount increases ($b = 1.86,\ \text{SE} = 0.06,\ p < 0.001$) but found no effect of partner type ($b = -0.07,\ \text{SE} = 0.05,\ p = 0.23$). More importantly, although there was no main effect of training condition ($b = -0.28,\ \text{SE} = 0.23,\ p = 0.22$), there was an interaction effect between training condition and offer amount ($b = -0.19,\ \text{SE} = 0.06,\ p = 0.001$), once again demonstrating that participants in the AI training condition were more sensitive to the offer amount than those in the control condition as the offer amount decreased. We also found a three-way interaction between offer amount, opponent, and training condition ($b = 0.10,\ \text{SE} = 0.04,\ p = 0.019$). To interpret this three-way interaction, we ran post-hoc mixed effects models conditioning on training condition. When conditioning on participants in the AI training condition, we found neither a main effect of partner type ($b = -0.08,\ \text{SE} = 0.07,\ p = 0.28$) nor an interaction effect of offer amount and partner type ($b = 0.02,\ \text{SE} = 0.05,\ p = 0.67$). Thus, participants in the AI training condition responded similarly to each partner type. When only considering participants in the control condition, we found no main effect of partner type ($b = -0.05,\ \text{SE} = 0.09,\ p = 0.54$). However, there was a significant interaction between partner type and offer amount ($b = -0.18, \text{SE} = 0.07,\ p = 0.0053$). Specifically, these participants accepted more unfair offers made by the AI than by the human participant. However, this difference in acceptance rates decreased as the offer amount became more fair. We found no other significant interactions ($ps \geq 0.08$).

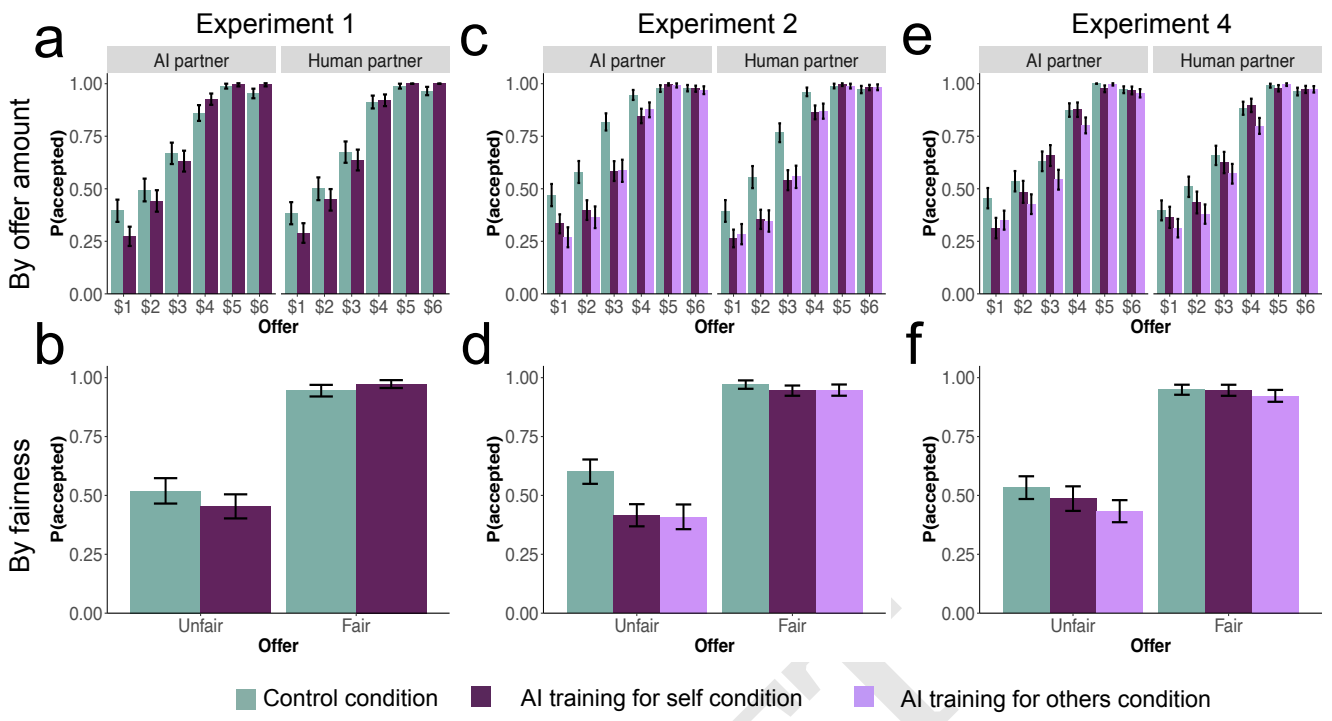

Fig. 5. Results for session 2 for Experiments 1, 2, and 4. All figures show the proportion of accepting an offer based on offer amount (top row) and by fairness (bottom row) conditioned on partner type and fairness. Error bars indicate standard error.

We ran an exploratory ANOVA for Experiment 5 since we ran this analysis for all other experiments in this paper. Once again, we found a main effect for fairness ($F_{1,206} = 62.55$, $p < 0.001$). However, no other main nor interactions were statistically significant ($ps \geq 0.22$). Although the results from the ANOVA model are inconsistent with the results from the mixed-effects model, we place more confidence in the mixed-effects model results for similar reasons outlined in Experiment 3.

The results from Experiments 4 and 5 show that people training AI rejected more unfair offers than those in the control condition even without a visual cue suggesting they were being observed. This suggests that people in the AI training conditions are not changing their behavior because they are more self aware (48) but rather are genuinely motivated to instill fairness into AI.

Persistence of the effect of AI training. Finally, we investigated whether the effects of AI training persisted over time. To do this, we analyzed choice behavior from the second sessions of Experiments 1, 2 and 4. In these sessions, participants in the AI training conditions (Experiment 1: $n = 95$; Experiment 2: AI training for others condition $n = 88$, AI training for self condition $n = 110$; Experiment 4: AI training for others condition $n = 112$, AI training for self condition $n = 92$) were informed that their responses would no longer be used for AI training while those in the control condition (Experiment 1: $n = 84$; Experiment 2: $n = 90$; Experiment 4: $n = 107$) completed the same task as in the first session. Thus, rational responders should revert to their baseline preferences and no differences between groups should be observed.

However, in an exploratory analysis of the second session in Experiment 1, we found that the AI training group continued to reject unfair offers at a higher rate (Figure 5a and Figure 5b). A mixed-effects model revealed that people in the second session were more likely to accept higher offer amounts ($b = 2.68$, SE $= 0.12$, $p < 0.001$). Although there was no main effect of training condition ($b = 0.48$, SE $= 0.40$, $p = 0.22$), there was an interaction with offer amount ($b = 0.75$, SE $= 0.11$, $p < 0.001$), demonstrating that participants who were previously assigned to the AI training condition were more sensitive to the offer amount than those in the control condition. Specifically, participants who were previously training an AI continued to reject more unfair offers. There was no main effect for partner type ($b = -0.11$, SE $= 0.08$, $p = 0.21$) or additional interaction effects ($ps \geq 0.37$).

We preregistered this same analysis for the data from the second session of Experiment 2** and replicated these results (Figure 5c and Figure 5d). There were no main effects for either partner type ($b = 0.04$, SE $= 0.13$, $p = 0.78$) or training condition for either the AI training for self ($b = -0.99$, SE $= 0.66$, $p = 0.13$) or AI training for others ($b = -1.32$, SE $= 0.69$, $p = 0.06$) conditions than the control condition. Participants were sensitive to the offer amount ($b = 1.98$, SE $= 0.11$, $p < 0.001$), but this sensitivity was increased for both the AI training for self condition ($b = 0.77$, SE $= 0.17$, $p < 0.001$) and the AI training for others condition ($b = 0.55$, SE $= 0.17$, $p = 0.001$). These groups of participants continued to reject unfair offers at a higher rate, even when they were informed their behavior would no

** Preregistration link found here: http://osf.io/f8sp6

longer train AI. There was neither a main effect of condition ($b = 0.33,\ \mathrm{SE} = 0.65,\ p = 0.61$) nor an interaction effect ($b = 0.21, \mathrm{SE} = 0.18, p = 0.25$) between training condition and offer amount between AI training conditions. There were no additional significant interactions ($ps \geq 0.09$).

We preregistered the same analysis for the second session of Experiment 4[††] and found mixed results (Figure 5e and Figure 5f). Participants were once again sensitive to the offer amount ($b = 1.96,\ \mathrm{SE} = 0.10,\ p < 0.001$). However, the mixed effect model did not reveal a main effect of either partner type ($b = 0.10,\ \mathrm{SE} = 0.10,\ p = 0.29$) or training condition for either the AI training for others ($b = -1.08,\ \mathrm{SE} = 0.58,\ p = 0.065$) or AI training for self ($b = -0.77,\ \mathrm{SE} = 0.62,\ p = 0.21$) conditions compared to the control condition. Interestingly, there was no interaction effect between offer amount and training effect when comparing the AI training for others ($b = 0.06,\ \mathrm{SE} = 0.13,\ p = 0.64$) or AI training for self ($b = -0.07,\ \mathrm{SE} = 0.14,\ p = 0.63$) to the control condition, not replicating previous results. Additionally, there was no main effect of training condition ($b = 0.31,\ \mathrm{SE} = 0.60,\ p = 0.61$) or interaction effect between offer amount and training condition ($b = -0.13,\ \mathrm{SE} = 0.13,\ p = 0.33$) between training conditions. There were no additional significant interactions ($ps \geq 0.27$).

In Experiments 1 and 2, participants continue to reject more unfair offers after AI training, providing strong evidence that this behavior persists after AI training. Strikingly, this result did not replicate in Experiment 4. While it is possible that habit formation requires more salience than we used in this study (especially in the short experimental paradigms used here), visual inspection of Figure 5e, suggests that people previously in the AI training conditions were more punitive than those in the control condition for unfair offers. An exploratory follow-up analyses provided provisional evidence for this conjecture. Specifically, when conditioning our analyses only on *unfair* offers, we found a significant effect of training condition for the AI training for others condition ($b = -1.72,\ \mathrm{SE} = 0.81,\ p = 0.034$) but not for the AI training self condition ($b = -1.00,\ \mathrm{SE} = 0.85,\ p = 0.24$). However, there was a three-way interaction between the effect of the training for self, offer amount, and training type, ($b = 0.40,\ \mathrm{SE} = 0.20,\ p = 0.048$). As described in more detail in the Supplementary Materials, this effect suggests that people in the training for self-condition only showed persistence of training effects when interacting with AI ($b = 0.92,\ \mathrm{SE} = 0.44,\ p = 0.034$).

These effects indicate that the behavioral changes adopted to train AI can persist, even after AI training finishes. However, they also suggest that these effects may be amplified when people adopt changes in behavioral policy under more salient conditions (i.e., when a webcam icon emphasizes the idea that their behavior is directly observed).

Discussion. AI models help us make decisions, but we help AI models by letting them learn from our behavior. Therefore, AI models risk tailoring their recommendations around the human biases they observe. This paper presents evidence for this claim. In five experiments, we told participants that their behavior would be used to train an AI algorithm. Some of them were told their responses would train AI that they would encounter again (Experiments 1, 2 and 4), whereas others trained an AI that would play against other participants (Experiments $2-5$). Regardless of who the recipient was, or whether a visual cue enforced the idea that people were observed during decision making, people became less likely to accept unfair offers when their behavior was used to train AI compared to those in the control conditions. This effect appeared differently across experiments: as a main effect (Experiment 4), as an interaction effect with offer amount (Experiments 1 and 5), or as both (Experiments 2 and 3). Most importantly, the results from each experiment showed that people training AI were more punitive for lower offers. Therefore, we can conclude that people were willing to give up money to train AI to be fair, even if they wouldn't benefit from the training. These findings expose a problem for AI models that aim to learn user preferences: AI algorithms assume that human behavior provides an unbiased training set (10), but people shift their behavior away from baseline preferences when given control over training.

Such issues are most likely to occur when AI developers are unaware that their training data is biased. Specifically, when behavior exhibited during training differs from that in a natural setting, AI will learn preferences that do not reflect natural behavior but, rather, align with their biases (49–51). This is particularly problematic if people believe AI is making unbiased recommendations. Thus, AI developers need to consider the ways in which people can intentionally shift their behavior to shape their algorithms to their preferences. This information may help them to invent safeguards that can debias algorithms during training. It may also be useful to transparently relay this information to the organizations that plan to use the AI (52), allowing them to design environments in which training biases are less likely to arise.

The participants in our studies were motivated to train AI to be fair, even when they did not benefit from such training. This indicates that communal motivations can play a significant role in the training of AI (53). Specifically, if people are concerned about the needs and welfare of other people using the same AI system, this may prompt them to act in ways that will make the AI act more beneficial to those other users (54–57). This finding prompts an intriguing question: why do people change behavior when training AI? It has been argued that preferences for fairness may reflect a desire to adhere to societal norms instead of a genuine consideration of others' well-being (58, 59). However, participants in our studies never interacted, casting doubt on this reputational hypothesis. It remains possible, however, that people forgo rewards from unfair offers to maintain a positive image of the self (60). Using tools from social psychology, future research should distinguish between these competing hypotheses.

Regardless of the underlying mechanism, the motivation to train AI to make more fair offers seems to be positive. However, it is important to consider that interpretations of fairness vary greatly between people. For example, subjective estimates of fairness in the ultimatum game differ between geographic regions (35), with responders in Asian regions having higher rejection rates than those in the US. People may even disagree on definitions of fairness in AI contexts (5, 61–66). For example, some people prefer AI systems to strive for equality (i.e., equal opportunity) whereas others argue that AI systems should aim to implement equity (allocating resources

[††] Preregistration link found here: https://osf.io/fpqr9

needed to achieve equal outcomes). Thus, even with a shared motivation to integrate fairness into AI, diverse perspectives on what constitutes fairness may result in conflicting training objectives and AI systems that are not well-tailored for particular populations.

The participants in our studies didn't just change their behavior while it was used for AI training but also endured this behavioral change beyond the training session. In short, we invited some of our participants to take part in a follow-up session (at least two days later) and informed those who were previously assigned to the AI training condition that their responses would no longer be used for AI training. Even though they were no longer training AI, they continued to reject unfair offers at a higher rate than in the control condition. That is, they persisted with the behavioral policy they had adopted in the first session, even when the change in context rendered this strategy suboptimal. In the animal learning literature, such insensitivity to changes in the structure of the environment is a hallmark feature of the formation of habit (30, 67). That is, after initially engaging in more goal-directed deliberation, choosing between actions after reasoning through their consequences (68, 69), repeated sequences of behavior are encoded as habits. These habitual sequences of behavior are then triggered, irrespective of their current value, by the stimuli that initially elicited the response. Our results indicate that deliberately training an AI algorithm can lead to a similar formation of habits. Moreover, because habits are triggered by the specific learning contexts in which they are encoded (70), repeated encounters with the same AI systems will perpetuate and reinforce habitual behavior. In short, if humans have control over AI training, then this does not only cause issues for the AI systems during training but also in subsequent sessions. In order to better understand its implications for AI developers, future research should systematically explore how and when habits form during AI training as well as how they can be prevented.

However, the strength of this newly acquired behavior following the AI training phase was diminished when participants were not given the impression that they were being observed. Specifically, in Experiments 1 and 2, participants were reminded of AI training by seeing a webcam cue on every trial, whereas those in Experiment 4 were only reminded through text prompts. The presence of the webcam icon may have induced self-awareness in participants (45), prompting them to engage in more goal-directed deliberation and consequently reinforcing the adoption of the new behavioral policy (68, 69, 71). Nevertheless, exploratory analyses of our data suggested that AI training still contributes to these transfer effects, given that motivation plays a crucial role in habit formation (72, 73). Future research can elucidate the influences of AI training and self-awareness created by the webcam on these transfer effects.

In our first three experiments, a webcam icon reminded participants that they were in the AI training condition. It is possible that this image may have caused people to become more self-aware (45), and therefore to change their behavior because they were concerned about their social image (74). To rule out this alternative explanation of our results, we removed the webcam icon in Experiments 4 and 5. The results from these experiments demonstrated that people remained motivated to train AI, even without reinforcing the idea of being observed. Even though we were able to control for this variable, it should be noted that AI developers are typically unable to remove the feeling that users are being watched. Specifically, AI users may still infer their behavior is being observed even when they are not explicitly told so, especially when subsequent recommendations reflect prior behavior. Therefore, it is likely that our first three experiments are a better model to understand the consequences of AI training on human behavior.

Nevertheless, it is likely that humans are sensitive to how AI algorithms signal that they are learning from human behavior. To study this question, researchers could adopt the current framework and vary how AI training is cued. We should note that while this paper reports two sets of experiments that differ in the way in which AI training was cued, we are unable to draw conclusions from a direct comparison between them because these studies were conducted months apart. This makes it impossible to ensure that the populations only differed in their experimental treatment (75). Thus, the current work opens an avenue for future research systematically exploring the ways in which AI training can affect human decision making.

Participants in our experiments responded to offers made by both humans and AI. This manipulation was mainly included to lend credence to the idea that AI systems take part in our behavioral studies, and therefore that it would be useful to train them for future experiments. However, it also allowed us to test whether people changed their responses when interacting with AI compared to other humans. Indeed, several studies (38–40, 76, 77) report that people are more likely to accept unfair offers from AI proposers. In contrast, we found no difference in acceptance behavior between partner types. There are several explanations for this discrepancy. For example, participants may have felt less interpersonal connection with the human proposers, because they were displayed as abstract silhouettes. Another possibility is that our study was conducted entirely online, whereas the other studies were conducted in-person. Finally, our study framed the AI systems as capable of learning, which may have prompted our participants to treat them more like their human counterparts.

Of course, several limitations remain. Because we only used the ultimatum game, it remains unknown whether AI training effects can be observed in other decision-making contexts. For example, the ultimatum game involves the notion of morality (27, 32, 33, 35, 78, 79), but AI recommendations in real life often occur in non-moral contexts. It remains unknown how willing people are to train AI in non-moral situations. Moreover, the ultimatum game has a clear definition of fairness in our study population (27, 32, 33, 35, 78, 79), but many moral situations in which people rely on AI recommendations do not have an agreed-upon definition of fairness (61, 63). For example, humans rely on AI to make serious decisions such as allocating kidneys to patients (80) and resources to the homeless (81). Our results do not speak to how people approach the training of AI in such situations.

We should also note that the stakes in our study were relatively low (participants earned 5% of the amount they earned in one of their negotiations). Therefore, it is unclear how participants would train AI with higher stakes. Research

has shown that people are more likely to accept lower offers when the stakes are higher (82–84). Thus, it is possible that training effects can only be observed when the perceived benefits of AI training outweigh personal gains. This observation brings a host of intriguing questions to the fore, revolving around whether people are sensitive to this tradeoff and which cognitive mechanisms they use to resolve it.

However, in many practical applications of AI training, the stakes for each individual are low, both in terms of value and the immediacy of the outcome. On social networking sites, a poor recommendation from an AI only results in a waste of a few seconds. These real-life consequences are equal, or perhaps even lower, to what participants encountered in our task. At the same time, many low-stakes decisions can collectively lead to high-stakes outcomes. For example, not accounting for potential human behavior during AI training can lead to significant social impacts, such as the echo chamber effect (85). This highlights the importance of understanding human decision making in AI training, even when the individual stakes are low.

In short, because this is the first demonstration that AI training impacts human behavior, many questions remain. We hope that they will inspire a program of research that aims to achieve such an understanding. Researchers may study the effects of AI training in novel decision-making contexts to study issues of generalizability. They may also vary the payoffs associated with choices in these tasks to study the tradeoff between AI training and personal gains. We believe that the general AI-training methodology introduced here will provide a valuable experimental tool for this line of research.

To conclude, we found that people change their behavior when they are aware it is used to train AI. In the context of a social decision-making game, we found that people prioritize fairness when training AI, not just to increase their own reward but also because of a consideration for the well-being of others. This behavior change persisted even in subsequent sessions where no AI training took place. Together, our results suggest that, when presented with the opportunity, people instill their preference into AI algorithms. Our work poses a challenge for the development of AI systems that collaborate with humans since it is assumed that humans produce unbiased training data (10). Therefore, developers should consider how humans can exploit their algorithms and consider ways in which such bias can be minimized.

Materials and Methods

Participants. Participants were recruited from Prolific for all five experiments. In Experiment 1, a total of 217 participants (113 female, 3 non-binary, 1 missing; $M = 38.25, \mathrm{SD} = 14.15$) were initially recruited, with 181 participants returning for the second session (91 female, 3 non-binary, 1 missing; $M = 38.78, \mathrm{SD} = 14.07$). Four participants were excluded from the analysis because they were exposed to both conditions by refreshing the webpage and were assigned to a different condition than the original one. For Experiment 2, 337 participants (159 female, 10 non-binary; $M = 38.20, \mathrm{SD} = 12.81$) were recruited, and 291 participants returned for the second session (132 female, 8 non-binary; $M = 39.21, \mathrm{SD} = 13.16$). Three participants were excluded for refreshing the webpage and being exposed to more than one condition. In Experiment 3, a total of 221 participants (89 female, 1 non-binary; $M = 41.01, \mathrm{SD} = 13.51$) were recruited, with three participants excluded for the same reasons. For experiment 4, 346 participants (175 female, 4 non-binary, 1 missing; $M = 43.17, \mathrm{SD} = 25.66$) were recruited, with 312 of them returning for the second session (155 female, 4 non-binary, 1 missing; $M = 43.46, \mathrm{SD} = 26.49$). One participant was excluded for exposure to both conditions. In experiment 5, 208 participants (87 female, 2 non-binary; $M = 43.24, \mathrm{SD} = 13.53$) were recruited, with 0 participants excluded from the analysis.

Each session took approximately 6 minutes, and participants received a median pay rate of approximately $10 per hour for session 1 and $14 per four for session 2 (all participants were paid $8.50 per hour before receiving a bonus). All participants provided informed consent, and the study received approval from the Washington University in St. Louis IRB.

Experimental Design. At the beginning of the first session of each experiment, participants were randomly assigned to a condition. For Experiment 1, this was either the 'AI training condition' ($n = 110$) or the 'control condition' ($n = 103$). For Experiment 2, this was either the 'AI training for self' ($n = 127$), 'AI training for others' ($n = 107$) or 'control condition' ($n = 100$). For Experiment 3, this was either the 'AI training condition' ($n = 117$) or 'control condition' ($n = 101$). For Experiment 4, this was either the 'AI training for self' ($n = 100$), 'AI training for others' ($n = 129$) or 'control condition' ($n = 116$). For Experiment 5, this was either the 'AI training condition' ($n = 105$) or 'control condition' ($n = 103$).

Next, participants were extensively instructed about the rules of the ultimatum game and completed two practice trials. Participants in Experiments 1, 2, and 4 were told they would get an opportunity to participate in a follow-up session within the next few weeks. Next, participants in the AI training conditions were either informed that their responses would be used to train a separate AI they would encounter during the follow-up session (Experiment 1 and 'AI training for self' condition in Experiments 2 and 4) or told that they would train an AI for other participants (Experiment 3, Experiment 5 and 'AI training for others' condition in Experiments 2 and 4). Participants were not told what this training would encompass.

Next, participants played multiple rounds of the ultimatum game (Figure 1). On each round, participants played as the responder and decided whether to accept or reject a proposer's offer of how to allocate a $10 sum between both partners. We manipulated partner type within-subject: each participant played against both AI and human participants. To help distinguish between them, each partner type was associated with a color, either blue or orange, which were randomly assigned for each participant.

In Experiments $1-3$, each round started with the display of a fixation cross ($750ms$). Next, a two-second presentation of an icon representing the partner type (human participant or AI) was displayed. Participants in the AI training conditions also saw an image of a webcam accompanied by the text "Offer used to train AI" on this screen. This served as a reminder that an AI would learn from their responses. Then, participants again saw the opponent icon, but now accompanied by the offer, which was displayed as a line of text indicating the proposed split (e.g., "I get $6 and you get $4"). In the AI training condition, a webcam icon was displayed in the top left corner of the screen as well. After two seconds, the words "accept" and "reject" appeared on the left and right sides of the screen, respectively, signaling that participants could make their choice using the 'F' and 'J' key on the keyboard. Participants were provided with unlimited time to make their decision.

For Experiments 4 and 5, participants completed the same task except with one critical change (Figure 3). Those in the AI training conditions did not see a webcam to remind them of AI training. Instead, after participants were informed of their partner type (with now no reminder of AI training), they saw an additional screen that said "Offer used to train AI." They saw this screen for $2s$. Participants also saw this text when asked to make a decision.

Participants completed 24 rounds of the ultimatum game, playing 12 rounds with each partner type. The offer amounts ranged from $1 to $6 and were presented in a random order. They were balanced across partner types for each participant, ensuring that all participants saw each offer two times for each partner type. For the AI partner trials, these offer amounts were programmed

to ensure that the offers were the same between conditions. For human partner trials, we recruited enough participants from various studies to ensure that we could balance offers between training conditions using the same amounts. Offer amounts $\$1 - \$3$ were considered to be unfair, while offers $\$4 - \$6$ were considered to be fair, consistent with previous literature (38).

For all sessions, each offer amount occurred with equal frequencies, except for in the second sessions of Experiments 2 and 4. Here, for trials where the AI was the partner, we used an update rule on this probability distribution with a learning rate of 0.5 to incorporate the responses of those assigned to an AI training condition. Specifically, participants in the control group played against an AI that was trained by those in the AI training others condition, while participants in both AI training conditions played against an AI that participants in the AI training self condition trained.

To incentivize choice behavior, participants were informed that one trial would be randomly selected and resolved at the end of each experiment. Participants received a bonus of 5% of the amount they earned from the trial selected in each first session. This bonus was increased to 15% for all second sessions to encourage them to return.

We used the participants' responses to pay the proposers who were recruited from various studies. Specifically, the proposers were informed that either a human participant or AI would respond to one of their proposals but were not informed which offer. If a human participant needed to respond to their offer, we randomly selected one response that corresponded to their offer amount and paid the proposer accordingly. If the AI needed to respond to their offer, we calculated the acceptance rates for each offer amount and used these acceptance rates to determine the participant's bonus.

For Experiments 1, 2 and 4, after completing session 1, participants were invited a few days later to complete session 2 (Experiment 1: AI training condition $n = 95$, control $n = 84$; Experiment 2: AI training for others condition $n = 88$, AI training for self condition $n = 110$, control $n = 90$; Experiment 4: AI training for others condition $n = 112$, AI training for self condition $n = 92$, control $n = 107$), and they completed the same task as before with a few modifications. Participants in the AI training conditions were informed their responses would no longer train AI. Therefore, they did not see a webcam (Experiments 1 and 2) or any text to remind them of AI training on each trial and completed the same task as those in the control group (see Figure 1b). Additionally, we changed the colors assigned to partner type to yellow and purple to avoid confusion across sessions. To encourage retention rates, we allotted 5 days to complete the experiment.

After completing the experiment, participants were asked to describe any strategies they developed and whether they knew what the ultimatum game was. If participants indicated they knew the ultimatum game, they were asked to describe the optimal strategy. These questions were not used in the analysis.

Analysis. The goal of our analyses was to determine whether participants' probability of accepting each offer was dependent on the offer amount, partner type, and (most crucially) the training condition. We analyzed data for each session and experiment separately.

For each session, we employed a logistic mixed-effects model to assess the factors that predict participants' acceptance of offers, including offer amounts, partner type, training condition, and their interactions. These models were estimated in R using lmerTest package [‡‡], and the following Generalized Linear Model equation:

$$\text{accept} \sim \text{partner} * \text{offer} * \text{training condition} + (1|\text{participant})$$

Here, our dependent variable 'accept' is binary (1 for acceptance, 0 for rejection). Independent variables include 'partner' (human or AI), 'offer' (integers $1 - 6$ centered around 0), and 'training condition' (control or AI training). We employed 'participant' as a random intercept to account for individual variability. We report the unstandardized estimates (b) along with their standard errors (SE).

[‡‡] We used the 'nlbinwrap' optimizer because it lets all models converge (except for the exploratory mixed effects models in experiment 4 session 2 since the 'nlbinwrap' optimizer did not converge, so we used the 'bobyqa' optimizer), but the 'bobyqa' optimizer produces qualitatively equal results for all experiments (but does not converge for Experiment 3).

Additionally, for session 1 of five three Experiments, we used R and JASP (86) to conduct a three-way within-between ANOVA to examine how the fairness of offers (categorized as fair: $\$4 - \$6$, and unfair: $\$1 - \$3$), partner type, training condition, and their interactions influenced the likelihood of accepting offers. Here, in contrast to the mixed-effects model specified above, we first computed subject-wise averages for each of the four combinations of fairness and partner type. These models allowed us to more precisely examine how fairness influenced offer acceptance. Any significant interactions were interpreted using post hoc t-tests and ANOVAs.

ACKNOWLEDGMENTS. We would like to thank members of the Control and Decision Making Lab and the Ho Lab for their advice and assistance. This work was supported in part by a seed grant from the Transdisciplinary Institute in Applied Data Sciences (TRIADS) at Washington University in St. Louis.

1. M Bayati, et al., Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS ONE* **9**, e109264 (2014).
2. C Giordano, et al., Accessing artificial intelligence for clinical decision-making. *Front. Digit. Heal.* **3** (2021).
3. F Jiang, et al., Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2**, 230–243 (2017).
4. DM Koh, et al., Artificial intelligence and machine learning in cancer imaging. *Commun. Medicine* **2** (2022).
5. J Angwin, J Larson, S Mattu, L Kirchner, Machine bias in *Ethics of data and analytics*. (Auerbach Publications), pp. 254–264 (2022).
6. Y Hayashi, K Wakabayashi, Can ai become reliable source to support human decision making in a court scene? in *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. (ACM), (2017).
7. A Završnik, Criminal justice, artificial intelligence systems, and human rights. *ERA Forum* **20**, 567–583 (2020).
8. MJ Azizi, P Vayanos, B Wilder, Rice, Designing fair, efficient, and interpretable policies for prioritizing homeless youth for housing resources in *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*. (Springer International Publishing), pp. 35–51 (2018).
9. A Kube, S Das, PJ Fowler, Allocating interventions based on predicted outcomes: A case study on homelessness services. *Proc. AAAI Conf. on Artif. Intell.* **33**, 622–629 (2019).
10. CK Morewedge, et al., Human bias in algorithm design. *Nat. Hum. Behav.* **7**, 1822–1824 (2023).
11. V Mathur, Y Stavrakas, S Singh, Intelligence analysis of tay twitter bot in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*. (IEEE), pp. 231–236 (2016).
12. MJ Wolf, K Miller, FS Grodzinsky, Why we should have seen that coming: Comments on microsoft's tay" experiment," and wider implications. *Acm Sigcas Comput. Soc.* **47**, 54–64 (2017).
13. A Cohn, T Gesche, MA Maréchal, Honesty in the digital age. *Manag. Sci.* **68**, 827–845 (2022).
14. CM de Melo, S Marsella, J Gratch, Social decisions and fairness change when people's interests are represented by autonomous agents. *Auton. Agents Multi-Agent Syst.* **32**, 163–187 (2017).
15. N Shechtman, LM Horowitz, Media inequality in conversation: How people behave differently when interacting with computers and people in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. (ACM), pp. 281–288 (2003).
16. Y Mou, K Xu, The media inequality: Comparing the initial human-human and human-ai social interactions. *Comput. Hum. Behav.* **72**, 432–440 (2017).
17. CM de Melo, S Marsella, J Gratch, People do not feel guilty about exploiting machines. *ACM Trans. Comput. Interact.* **23** (2016).
18. A Erlei, R Das, L Meub, Anand, For what it's worth: Humans overwrite their economic self-interest to avoid bargaining with ai systems in *CHI Conference on Human Factors in Computing Systems*. (ACM), (2022).
19. AB Karagoz, ZM Reagh, W Kool, The construction and use of cognitive maps in model-based control. *J. Exp. Psychol. Gen.* (2023).
20. W Kool, SJ Gershman, FA Cushman, Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychol. Sci.* **28**, 1321–1333 (2017).
21. GA Miller, E Galanter, KH Pribram, *Plans and the Structure of Behavior*. (Henry Holt and Co), (1960).
22. T Pouncy, P Tsividis, SJ Gershman, What is the model in model-based planning? *Cogn. Sci.* **45**, e12928 (2021).
23. M Hardt, N Megiddo, C Papadimitriou, M Wootters, Strategic classification in *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. pp. 111–122 (2016).
24. J Perdomo, T Zrnic, C Mendler-Dünner, M Hardt, Performative prediction in *International Conference on Machine Learning*. (PMLR), pp. 7599–7609 (2020).

25. JP Miller, JC Perdomo, T Zrnic, Outside the echo chamber: Optimizing the performative risk in *International Conference on Machine Learning*. (PMLR), pp. 7710–7720 (2021).
26. CA Goodhart, C Goodhart, *Problems of monetary management: the UK experience*. (Springer), (1984).
27. A Camacho, E Conover, Manipulation of social program eligibility. *Am. Econ. Journal: Econ. Policy* **3**, 41–65 (2011).
28. H Aarts, B Verplanken, A Knippenberg, Predicting behavior from actions in the past: Repeated decision making or a matter of habit? *J. Appl. Soc. Psychol.* **28**, 1355–1374 (1998).
29. KJ Miller, A Shenhav, EA Ludvig, Habits without values. *Psychol. Rev.* **126**, 292–311 (2019).
30. A Dickinson, Actions and habits: The development of behavioural autonomy. *Philos. Transactions Royal Soc. London. B, Biol. Sci.* **308**, 67–78 (1985).
31. P Watson, C O'Callaghan, I Perkes, L Bradfield, K Turner, Making habits measurable beyond what they are not: A focus on associative dual-process models. *Neurosci. & Biobehav. Rev.* **142**, 104869 (2022).
32. W Güth, R Schmittberger, B Schwarze, An experimental analysis of ultimatum bargaining. *J. economic behavior & organization* **3**, 367–388 (1982).
33. CF Camerer, Strategizing in the brain. *Science* **300**, 1673–1675 (2003).
34. CF Camerer, *Behavioral game theory: Experiments in strategic interaction*. (Princeton university press), (2011).
35. H Oosterbeek, R Sloof, G Van De Kuilen, Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Exp. economics* **7**, 171–188 (2004).
36. E van Dijk, CK De Dreu, Experimental games and social decision making. *Annu. Rev. Psychol.* **72**, 415–438 (2021).
37. LS Treiman, CJ Ho, W Kool, Humans forgo reward to instill fairness into ai in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 11, pp. 152–162 (2023).
38. L Moretti, G Di Pellegrino, Disgust selectively modulates reciprocal fairness in economic interactions. *Emotion* **10**, 169 (2010).
39. AG Sanfey, JK Rilling, JA Aronson, LE Nystrom, JD Cohen, The neural basis of economic decision-making in the ultimatum game. *Sci. (New York, N.Y.)* **300**, 1755–1758 (2003).
40. M van 't Wout, RS Kahn, AG Sanfey, A Aleman, Affective state and decision-making in the ultimatum game. *Exp. Brain Res.* **169**, 564–568 (2006).
41. PM Blau, *Exchange and Power in Social Life*. (Routledge), 2 edition, (1986).
42. JA Colquitt, JA LePine, RF Piccolo, CP Zapata, BL Rich, Explaining the justice–performance relationship: Trust as exchange deepener or trust as uncertainty reducer? *J. Appl. Psychol.* **97**, 1–15 (2012).
43. JA Colquitt, et al., Justice at the millennium, a decade later: A meta-analytic test of social exchange and affect-based perspectives. *J. Appl. Psychol.* **98**, 199–236 (2013).
44. J Peterburs, et al., Processing of fair and unfair offers in the ultimatum game under social observation. *Sci. reports* **7**, 44062 (2017).
45. J Greenberg, Overcoming egocentric bias in perceived fairness through self-awareness. *Soc. Psychol. Q.* **46**, 152–156 (1983).
46. X Wang, F Li, B Cao, Both rewards and moral praise can increase the prosocial decisions: Revealed in a modified ultimatum game task. *Front. Psychol.* **9**, 344947 (2018).
47. Z Wei, Z Zhao, Y Zheng, Neural mechanisms underlying social conformity in an ultimatum game. *Front. Hum. Neurosci.* **7**, 896 (2013).
48. RA Wicklund, The influence of self-awareness on human behavior: The person who becomes self-aware is more likely to act consistently, be faithful to societal norms, and give accurate reports about himself. *Am. Sci.* **67**, 187–193 (1979).
49. M Cazes, N Franiatte, A Delmas, family=André, Rodier, Evaluation of the sensitivity of cognitive biases in the design of artificial intelligence in *Rencontres Des Jeunes Chercheurs En Intelligence Artificielle (RJCIA'21) Plate-Forme Intelligence Artificielle (PFIA'21)*. (2021).
50. M Soleimani, A Intezari, DJ Pauleen, Mitigating cognitive biases in developing ai-assisted recruitment systems. *Int. J. Knowl. Manag.* **18**, 1–18 (2021).
51. M Soleimani, A Intezari, N Taskin, D Pauleen, Cognitive biases in developing biased artificial intelligence recruitment system. *Hawaii Int. Conf. on Syst. Sci.* **54**, 5091–5099 (2021).
52. S Tolan, Fair and unbiased algorithmic decision making: Current state and future challenges. *arXiv preprint arXiv:1901.04730* (2019).
53. BM Le, EA Impett, EP Lemay, Muise, Communal motivation and well-being in interpersonal relationships: An integrative review and meta-analysis. *Psychol. Bull.* **144**, 1–25 (2018).
54. MS Clark, J Mills, Interpersonal attraction in exchange and communal relationships. *J. Pers. Soc. Psychol.* **37**, 12–24 (1979).
55. MS Clark, J Mills, The difference between communal and exchange relationships: What it is and is not. *Pers. Soc. Psychol. Bull.* **19**, 684–691 (1993).
56. MS Clark, JR Mills, A theory of communal (and exchange) relationships. *Handb. theories social psychology* **2**, 232–250 (2012).
57. VLP A.M, ET Higgins, AW Kruglanski, eds., *Social Psychology: Handbook of Basic Principles*. (The Guilford Press), Third edition edition, (2021) Includes bibliographical references and index.
58. R Golman, Acceptable discourse: Social norms of beliefs and opinions. *Eur. Econ. Rev.* **160**, 104588 (2023).
59. ET Higgins, Achieving 'shared reality' in the communication game: A social action that create; meaning. *J. Lang. Soc. Psychol.* **11**, 107–131 (1992).
60. R Bénabou, J Tirole, Incentives and prosocial behavior. *Am. economic review* **96**, 1652–1678 (2006).
61. S Barocas, M Hardt, A Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. (MIT Press), (2023).
62. W Dieterich, C Mendoza, T Brennan, Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc* **7**, 1 (2016).
63. J Kleinberg, S Mullainathan, M Raghavan, Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
64. D Lee, J Kanellis, WR Mulley, Allocation of deceased donor kidneys: A review of international practices. *Nephrology* **24**, 591–598 (2019).
65. N Grgic-Hlaca, EM Redmiles, KP Gummadi, A Weller, Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction in *Proceedings of the 2018 world wide web conference*. pp. 903–912 (2018).
66. M Srivastava, H Heidari, A 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*. pp. 2459–2468 (2019).
67. PC Holland, Relations between pavlovian-instrumental transfer and reinforcer devaluation. *J. Exp. Psychol. Animal Behav. Process.* **30**, 104 (2004).
68. A Dickinson, B Balleine, Motivational control of goal-directed action. *Animal learning & behavior* **22**, 1–18 (1994).
69. RJ Dolan, P Dayan, Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
70. W Wood, DT Neal, A new look at habits and the habit-goal interface. *Psychol. review* **114**, 843 (2007).
71. W Wood, D Rünger, Psychology of habit. *Annu. review psychology* **67**, 289–314 (2016).
72. A Dickinson, G Dawson, Pavlovian processes in the motivational control of instrumental performance. *The Q. J. Exp. Psychol.* **39**, 201–213 (1987).
73. A Dickinson, OD Pérez, Actions and habits: Psychological issues in dual-system theory in *Goal-directed decision making*. (Elsevier), pp. 1–25 (2018).
74. G Grimalda, A Pondorfer, DP Tracer, Social image concerns promote cooperation more than altruistic punishment. *Nat. communications* **7**, 12288 (2016).
75. S Edwards, M Clarke, S Wordsworth, J Borrill, Indirect comparisons of treatments based on systematic reviews of randomised controlled trials. *Int. journal clinical practice* **63**, 841–854 (2009).
76. M Chen, Z Zhao, H Lai, The time course of neural responses to social versus non-social unfairness in the ultimatum game. *Soc. Neurosci.* **14**, 409–419 (2018).
77. E Torta, E van Dijk, PA Ruijten, RH Cuijpers, The ultimatum game as measurement tool for anthropomorphism in human–robot interaction in *Social Robotics: 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013, Proceedings 5*. (Springer), pp. 209–217 (2013).
78. E Fehr, S Gächter, Fairness and retaliation: The economics of reciprocity. *J. economic perspectives* **14**, 159–182 (2000).
79. W Liu, et al., Morality is supreme: the roles of morality, fairness and group identity in the ultimatum paradigm. *Psychol. Res. Behav. Manag.* pp. 2049–2065 (2022).
80. R Freedman, JS Borg, Sinnott-Armstrong, V Conitzer, Adapting a kidney exchange algorithm to align with human values. *Proc. AAAI Conf. on Artif. Intell.* **32** (2018).
81. N Jo, et al., Fairness in contextual resource allocation systems: Metrics and incompatibility results in *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37, pp. 11837–11846 (2023).
82. S Andersen, S Ertaç, U Gneezy, Hoffman, Stakes matter in ultimatum games. *Am. Econ. Rev.* **101**, 3427–3439 (2011).
83. R Slonim, AE Roth, Learning in high stakes ultimatum games: An experiment in the slovak republic. *Econometrica* **66**, 569 (1998).
84. J Novakova, J Flegr, How much is our fairness worth? the effect of raising stakes on offers by proposers and minimum acceptable offers in dictator and ultimatum games. *PLoS ONE* **8**, e60966 (2013).
85. M Cinelli, G De Francisci Morales, A Galeazzi, W Quattrociocchi, M Starnini, The echo chamber effect on social media. *Proc. Natl. Acad. Sci.* **118**, e2023301118 (2021).
86. JASP Team, JASP (Version 0.18.3)[Computer software] (2024).