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 June 24, 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. We report three studies that indicate this assumption is invalid, as people change their behavior to instill desired routines into AI. To show this behavioral shift, we recruited participants to play the ultimatum game, where they were asked to decide whether to accept or reject 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 three experiments, we found that people modify 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. Interestingly, 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 structures to our advantage (19–22). A more applied example comes from (23), 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.

Interestingly, this change in behavior may persist beyond AI training. Research from experimental psychology has shown that behavior initially implemented to pursue a goal will eventually become habitually engrained (24, 25). 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; (26, 27)). 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

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

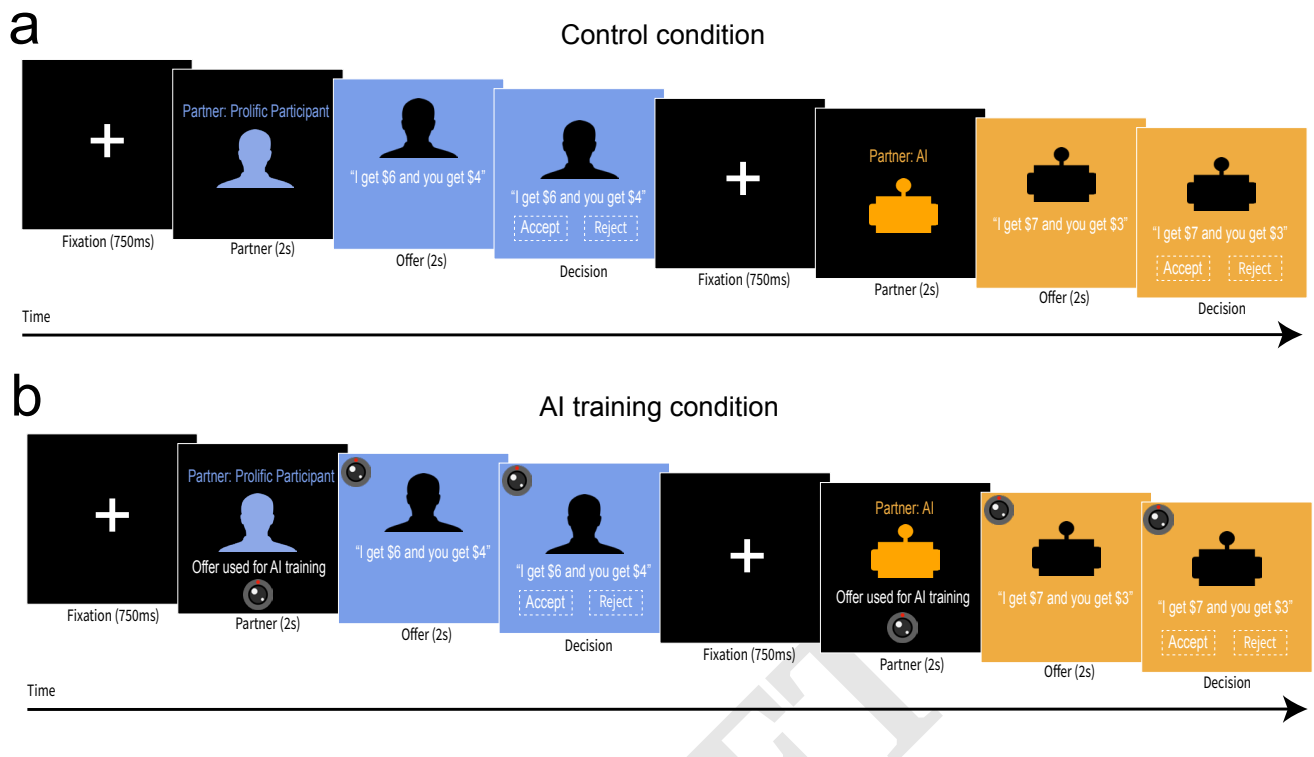

Fig. 1. Example trial sequences for the control (a) and AI training (b) conditions. For participants assigned to the AI training condition, a webcam was shown to remind the participants that their responses were training AI. Participants in the control condition did not see a webcam since their responses were not training AI. With the exception of the webcam, the trial format was the same for both conditions. Specifically, 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.

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 (28). 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 (29–31). In other words, the ultimatum game measures how subjective fairness affects decision making (32). 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 three preregistered experiments, we tested whether people modify their behavior on the ultimatum game (28) when informed their responses would be used to train an AI (Figure 1). Each round of this game involved a different partner, which was either a participant recruited from another experiment or an AI algorithm. On each round, participants were shown the partner's proposal on how they allocated $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.

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 of the ultimatum game started with a visual cue reminding participants behavior would be used to train AI. Two days later, all participants were invited to complete a second

*A preliminary version of this dataset was published in (33)

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

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

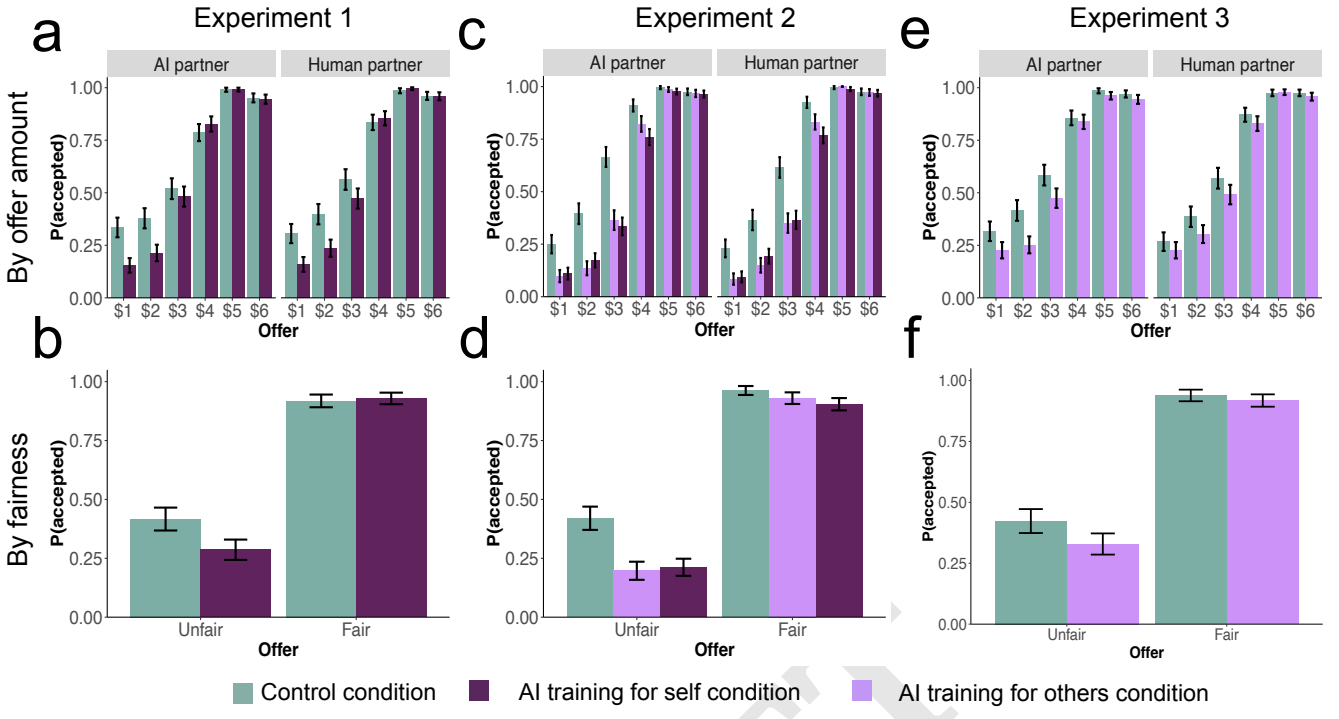

Fig. 2. Results for session 1 for Experiments $1-3$. All figures show the proportion of accepting an offer based on offer amount (top row) and by fairness (bottom row) conditioned on partner type and fairness. Error bars indicate standard error.

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 ($\beta = 1.80, p < 0.001$), replicating prior findings (34–36). However, they did not respond differently when partnered with a human compared to AI ($\beta = -0.11, p = 0.051$). Most importantly, even though we found no main effect of training condition ($\beta = -0.37, p = 0.064$), there was a significant interaction between training condition and offer amount ($\beta = 0.17, p = 0.002$). As can be seen in Figure 4a and Figure 4b, 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 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 ($n = 101$), 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

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

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 ($\beta = 2.89, p < 0.001$). Additionally, participants in both the AI training for self condition ($\beta = -1.92, p < 0.001$) and the AI training for others condition ($\beta = -1.76, p < 0.001$) 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 ($\beta = 0.30, p = 0.03$) and the AI training for others ($\beta = 0.55, p < 0.001$) 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 4c and Figure 4d. An additional mixed-effects model indicated that participants in the AI training conditions accepted offers similarly ($\beta = -0.16, p = 0.71$) and were comparably sensitive to the offer amounts ($\beta = -0.24, p = 0.11$). There was no main effect of partner type ($\beta = 0.07, p = 0.51$). 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 4c, 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 (37–39): 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 ($\beta = 2.29, p < 0.001$) but no effect of partner type ($\beta = -0.02, p = 0.70$). More importantly, we found a main effect of training condition ($\beta = -0.59, p = 0.014$), 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 ($\beta = -0.19, p = 0.003$), replicating that participants in the AI training condition were more likely to reject unfair offers (Figure 4e and Figure 4f). 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 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).

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 and 2. In these sessions,

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

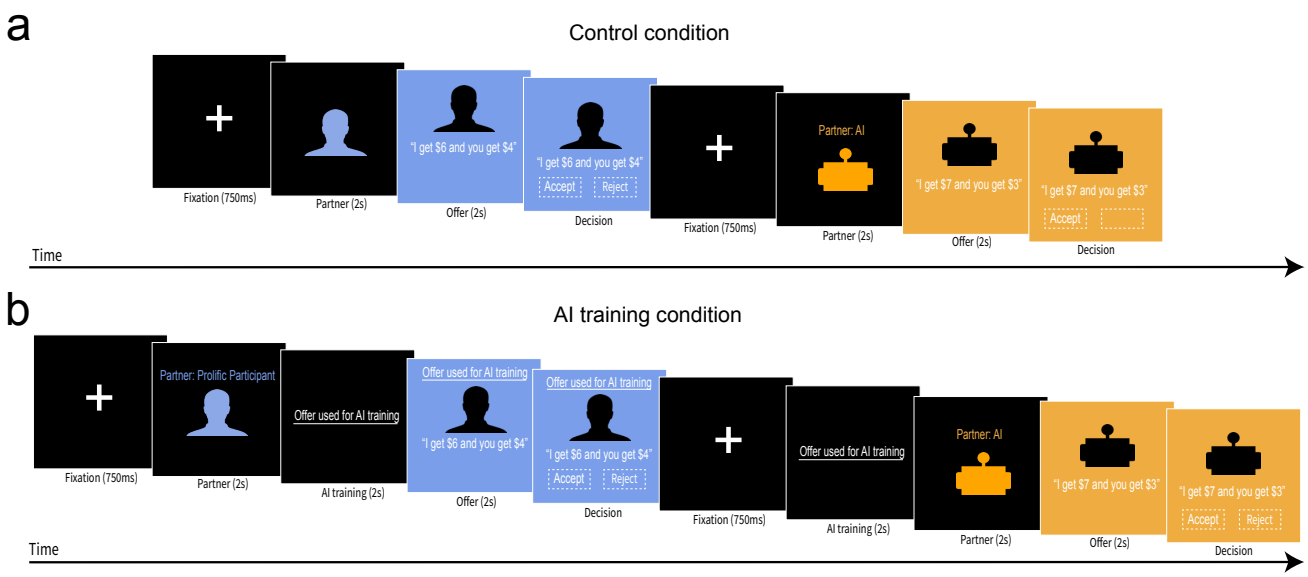

Fig. 3. Example trial sequences for the control (a) and AI training (b) conditions. For participants assigned to the AI training condition, additional text was shown to remind the participants that their responses were training AI. Participants in the control condition did not see this text since their 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. Specifically, 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.

participants in the AI training conditions (Experiment 1: $n = 95$; Experiment 2: AI training for others condition $n = 89$, AI training for self condition $n = 111$) were informed that their responses would no longer be used for AI training while those in the control condition (Experiment 1: $n = 87$; Experiment 2: $n = 93$) 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 ($\beta = 4.45, p < 0.001$). Although there was no main effect of training condition ($\beta = 0.48, p = 0.22$), there was an interaction with offer amount ($\beta = 0.75, p < 0.001$), 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 ($\beta = -0.11, p = 0.21$) 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 ($\beta = 0.04, p = 0.74$) or training condition for either the AI training for self ($\beta = -0.97, p = 0.15$) or AI training for others ($\beta = -1.29, p = 0.06$) conditions than the control condition. Participants were sensitive to the offer amount ($\beta = 5.08, p < 0.001$), but this sensitivity was increased for both the AI training for self condition ($\beta = 0.75, p < 0.001$) and the AI training for others condition ($\beta = 0.55, p = 0.001$). These groups of participants continued to reject unfair offers at a higher rate, even when they were informed their behavior would no longer train AI. There was neither a main effect of condition ($\beta = 0.32, p = 0.62$) nor an interaction effect ($\beta = 0.20, p = 0.28$) between condition and offer amount between AI training conditions. There were no additional significant interactions ($ps \geq 0.10$).

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 three 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 and 2), whereas others trained an AI that would play against other participants (Experiments 2 and 3). Regardless of who the recipient was, people became less likely to accept unfair offers when their behavior was used to train AI compared to those in the control condition. In other words, 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.

Our findings suggest that such issues 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 (40–42). This is particularly problematic if people believe AI is making unbiased recommendations. Thus, AI

[‖] Preregistration link found here: http://osf.io/f8sp6

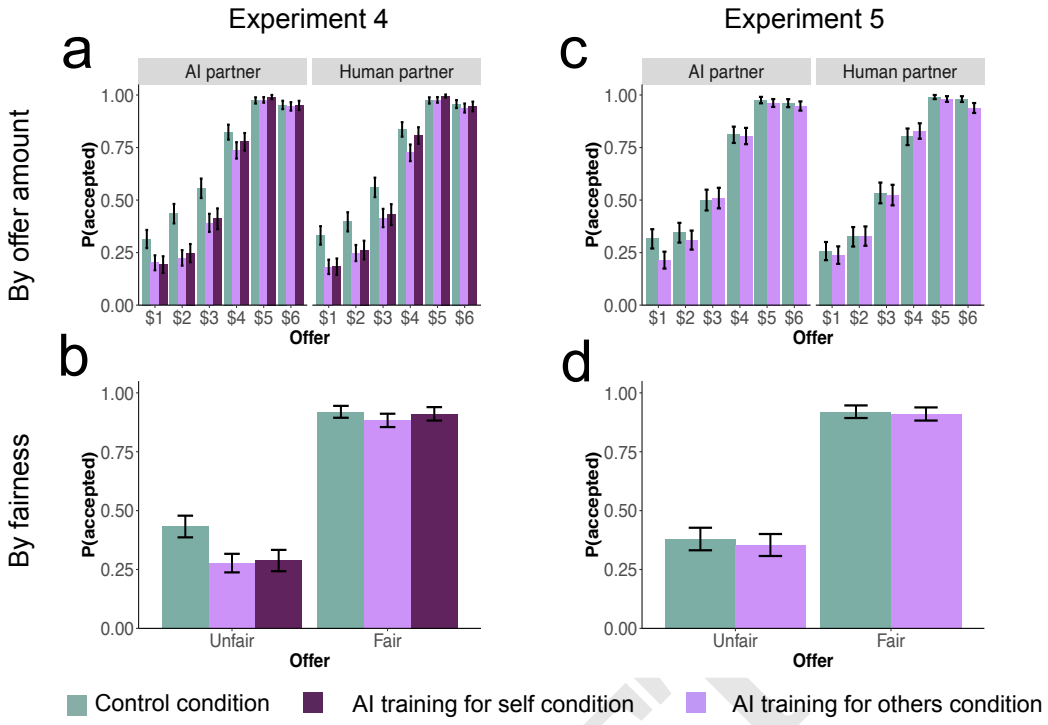

Fig. 4. Results for session 1 for Experiments 4 and 5. All figures show the proportion of accepting an offer based on offer amount (top row) and by fairness (bottom row) conditioned on partner type and fairness. Error bars indicate standard error.

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 (43), 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 (44). 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 (45–48). 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 (49, 50). 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 (51). 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 people have different interpretations of what fairness is. For example, subjective estimates of fairness in the ultimatum game differ between geographic regions (31), 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, 52–57). 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 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 (26, 58). That is, after initially engaging in more goal-directed deliberation, choosing between actions after reasoning through their consequences (59, 60), 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

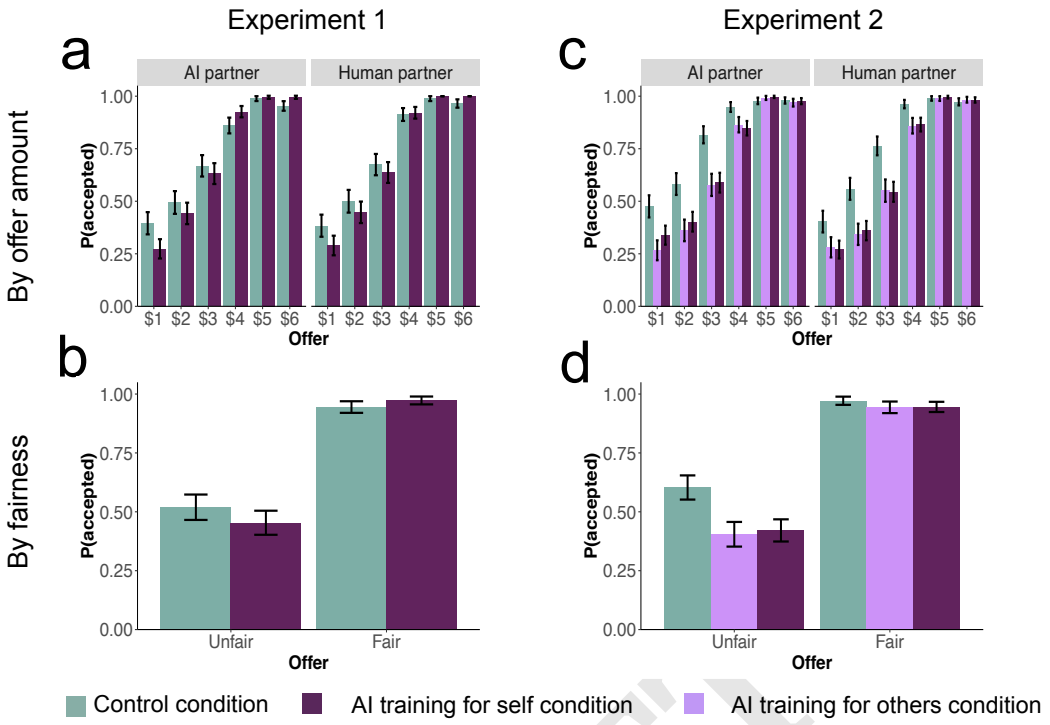

Fig. 5. Results for session 2 for Experiments 1 and 2. All figures show the proportion of accepting an offer based on offer amount (top row) and by fairness (bottom row) conditioned on partner type and fairness. Error bars indicate standard error.

habits are triggered by the specific learning contexts in which they are encoded (61), 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.

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 (34–36, 62, 63) 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 solely used the ultimatum game, it remains unknown whether AI training effects can be observed in other decision-making contexts. In real life, people rely on AI to make more serious decisions such as allocating kidneys to patients (64) and resources to the homeless (65). It is unclear how people approach the training in AI in such situations. Indeed, people's behavior in the ultimatum game changes when there is more reward at stake (66–68). It is possible that training effects can only be observed when the incentives are not compelling enough for individuals to prioritize personal gains. Of course, these issues of generalizability can be studied with the general AI-training methodology introduced in this paper.

Our paper reports evidence that people change behavior when they are aware their behavior 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 three 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 182 returning for the second session (91 female, 2 non-binary, 1 missing; $M = 38.68, \mathrm{SD} = 14.10$). 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, 339 participants (160 female, 10 non-binary; $M = 38.30, \mathrm{SD} = 12.85$) were recruited, and 291 returned for the second session (132 female, 8 non-binary; $M = 39.20, \mathrm{SD} = 13.14$). Four participants were excluded, three for exposure to both conditions and one for completing the task twice. 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.

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' ($n = 102$). For Experiment 3, this was either the 'AI training condition' ($n = 117$) or 'control condition' ($n = 101$).

Next, participants were extensively instructed about the rules of the ultimatum game and completed two practice trials. Participants in Experiments 1 and 2 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 Experiment 2) or told that they would train an AI for other participants (Experiment 3 and 'AI training for others' condition in Experiment 2). 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. This partner was either AI or another participant recruited from a separate experiment. Each partner type was associated with a color, either blue or orange, and was randomly assigned for each participant.

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.

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 randomized and balanced across partner types for each participant. Offer amounts $\$1 - \$3$ were considered to be unfair, while offers $\$4 - \$6$ were considered to be fair, consistent with previous literature (34).

For all sessions, each offer amount occurred with equal frequencies, except for in the second session of Experiment 2. 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 both second sessions to encourage them to return.

For Experiments 1 and 2, after completing session 1, participants were invited a few days later to complete session 2 (Experiment 1: AI training condition $n = 95$, control $n = 87$; Experiment 2: AI training for others condition $n = 89$, AI training for self condition $n = 111$, control $n = 93$), 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 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 one week to complete the experiment.

After completing the experiment, participants were asked to describe any strategies they developed. Additionally, they knew what the ultimatum game was. If they answered yes, 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 MATLAB and R using the fitglme function in MATLAB and lmerTest package in R, 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.

Additionally, for session 1 of all three Experiments, we used R and JASP (69) 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, 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. A Camacho, E Conover, Manipulation of social program eligibility. *Am. Econ. Journal: Econ. Policy* **3**, 41–65 (2011).
24. 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).
25. KJ Miller, A Shenhav, EA Ludvig, Habits without values. *Psychol. Rev.* **126**, 292–311 (2019).
26. A Dickinson, Actions and habits: The development of behavioural autonomy. *Philos. Transactions Royal Soc. London. B, Biol. Sci.* **308**, 67–78 (1985).
27. 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).
28. W Güth, R Schmittberger, B Schwarze, An experimental analysis of ultimatum bargaining. *J. economic behavior & organization* **3**, 367–388 (1982).
29. CF Camerer, Strategizing in the brain. *Science* **300**, 1673–1675 (2003).
30. CF Camerer, *Behavioral game theory: Experiments in strategic interaction*. (Princeton university press), (2011).
31. 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).
32. E van Dijk, CK De Dreu, Experimental games and social decision making. *Annu. Rev. Psychol.* **72**, 415–438 (2021).
33. 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).
34. L Moretti, G Di Pellegrino, Disgust selectively modulates reciprocal fairness in economic interactions. *Emotion* **10**, 169 (2010).
35. 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).
36. 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).
37. PM Blau, *Exchange and Power in Social Life*. (Routledge), 2 edition, (1986).
38. 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).
39. 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).
40. 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).
41. M Soleimani, A Intezari, DJ Pauleen, Mitigating cognitive biases in developing ai-assisted recruitment systems. *Int. J. Knowl. Manag.* **18**, 1–18 (2021).
42. 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).
43. S Tolan, Fair and unbiased algorithmic decision making: Current state and future challenges. *arXiv preprint arXiv:1901.04730* (2019).
44. 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).
45. MS Clark, J Mills, Interpersonal attraction in exchange and communal relationships. *J. Pers. Soc. Psychol.* **37**, 12–24 (1979).
46. 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).
47. MS Clark, JR Mills, A theory of communal (and exchange) relationships. *Handb. theories social psychology* **2**, 232–250 (2012).
48. 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.
49. R Golman, Acceptable discourse: Social norms of beliefs and opinions. *Eur. Econ. Rev.* **160**, 104588 (2023).
50. ET Higgins, Achieving 'shared reality' in the communication game: A social action that create; meaning. *J. Lang. Soc. Psychol.* **11**, 107–131 (1992).
51. R Bénabou, J Tirole, Incentives and prosocial behavior. *Am. economic review* **96**, 1652–1678 (2006).
52. S Barocas, M Hardt, A Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. (MIT Press), (2023).
53. W Dieterich, C Mendoza, T Brennan, Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc* **7**, 1 (2016).
54. J Kleinberg, S Mullainathan, M Raghavan, Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
55. D Lee, J Kanellis, WR Mulley, Allocation of deceased donor kidneys: A review of international practices. *Nephrology* **24**, 591–598 (2019).
56. 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).
57. 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).
58. PC Holland, Relations between pavlovian-instrumental transfer and reinforcer devaluation. *J. Exp. Psychol. Animal Behav. Process.* **30**, 104 (2004).
59. A Dickinson, B Balleine, Motivational control of goal-directed action. *Animal learning & behavior* **22**, 1–18 (1994).
60. RJ Dolan, P Dayan, Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
61. W Wood, DT Neal, A new look at habits and the habit-goal interface. *Psychol. review* **114**, 843 (2007).
62. 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).
63. 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).
64. 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).
65. 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).
66. S Andersen, S Ertaç, U Gneezy, Hoffman, Stakes matter in ultimatum games. *Am. Econ. Rev.* **101**, 3427–3439 (2011).
67. R Slonim, AE Roth, Learning in high stakes ultimatum games: An experiment in the slovak republic. *Econometrica* **66**, 569 (1998).
68. 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).
69. JASP Team, JASP (Version 0.18.3)[Computer software] (2024).