Humans Forgo Reward to Instill Fairness into AI

Lauren S. Treiman, Chien-Ju Ho, Wouter Kool

Washington University in St. Louis
ltreiman@wustl.edu, chienju.ho@wustl.edu, wkool@wustl.edu

Abstract

In recent years, artificial intelligence (AI) has become an integral part of our daily lives, assisting us with decision making. During such interactions, AI algorithms often use human behavior as training input. Therefore, it is important to understand whether people change their behavior when they train AI and if they continue to do so when training does not benefit them. In this work, we conduct behavioral experiments in the context of the ultimatum game to answer these questions. In our version of this game, participants 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 that their choices would be used to train AI, while others did not receive this information. In the first experiment, we found that participants were willing to sacrifice personal earnings to train AI to be fair as they became less inclined to accept unfair offers. The second experiment replicated and expanded upon this finding, revealing that participants were motivated to train AI even if they would never encounter it in the future. These findings demonstrate that humans are willing to incur costs to change AI algorithms. Moreover, they suggest that human behavior during AI training does not necessarily align with baseline preferences. This observation poses a challenge for AI development, revealing that it is important for AI algorithms to account for their influence on behavior when recommending choices.

Introduction

Artificial intelligence (AI) plays an increasingly important role in the decision-making processes we encounter in everyday life. For example, AI assists viewers with deciding what YouTube movie to watch, doctors with determining patient care (Bayati et al. 2014; Giordano et al. 2021; Jiang et al. 2017; Koh et al. 2022), judges with granting bail (Angwin et al. 2016; Hayashi and Wakabayashi 2017; Završnik 2020), and public policymakers with allocating resources for the homeless (Azizi et al. 2018; Kube, Das, and Fowler 2019). One of the most promising prospects of these human-computer interactions is that AI holds the potential to help us make more optimal and less biased decisions (Bansal et al. 2019, 2020). Before this can be achieved, it is imperative to understand the issues that arise when humans and AI interact. These issues become especially apparent when AI systems are trained on human behavior.

When training data does not properly represent the population, generalizability problems arise (Lai et al. 2022). For example, when trained to classify gender, AI models trained predominantly on lighter-skinned people are best at classifying lighter-skinned individuals overall and particularly bad at classifying darker-skinned females (Buolamwini and Gebru 2018). These findings suggest that if individuals are biased when constructing training sets for AI, then they will cause it to perpetuate their biases when recommending and making decisions (Cazes et al. 2021; Soleimani, Intezari, and Pauleen 2021; Soleimani et al. 2021). Therefore, it is crucial for the individuals that develop AI-assisted choice systems are highly aware of their biases (Ntoutsi et al. 2020).

However, not all training sets are under the control of the developers. AI is often trained on data directly supplied by the people with which it interacts. This notion raises the question of how humans change their behavior when aware that AI will learn from their interactions. A notable example is Microsoft's AI chatbot Tay in 2016, which launched on Twitter to learn from conversations with other users on the platform. However, within just 16 hours, the account had to be abruptly shut down since Twitter users were intentionally training Tay to be racist, sexist, and anti-Semitic, which resulted in Tay generating offensive tweets (Mathur, Stavrakas, and Singh 2016; Wolf, Miller, and Grodzinsky 2017). This sequence of events suggests that when given the opportunity to train AI systems, people change their behavior based on their motives. However, research on this phenomenon remains sparse.

A related set of findings demonstrates that humans change their behavior when interacting with AI systems. For example, individuals are more likely to cheat on a coin-tossing task when they report to a machine compared to a human (Cohn, Gesche, and Maréchal 2022). Moreover, individuals tend to experience less guilt when making unfair offers to AI than to another human (de Melo, Marsella, and Gratch 2016). Furthermore, individuals generally prefer interacting with other humans over AI and may change their responses to avoid AI interaction altogether (Erlei et al. 2022). These findings suggest that if AI is trained using human interactions, it should account for existing human biases as well as behavioral changes driven by these interactions.

In this work, we report two experiments that directly test whether humans adopt a different decision-making strategy when they know their responses will train AI. To the best of our knowledge, we are the first to examine changes in human behavior during AI training. In our first experiment, we investigated whether people would train AI to exhibit fair behavior if they could profit from this training in a subsequent session. In our second experiment, we tested whether individuals would train AI to be fair in situations where only others could benefit. We report two experiments that directly test whether humans adopt a different decision-making strategy when they know their responses will train AI. We investigated whether people would train AI to exhibit fair behavior if they could profit from this training in a subsequent session. Next, we tested whether individuals would train AI to be fair in situations where only others could benefit. These experiments also allowed us to determine whether individuals make different choices, especially pertaining to perceived fairness, when interacting with AI compared to other humans (Sanfey et al. 2003; van 't Wout et al. 2006).

Our experiments are conducted in the context of the ultimatum game. In this game, two people are asked to split a sum of money. One person, the proposer, decides how to divide the money between the two of them. The second person, the responder, chooses to accept or reject this offer. If the responder accepts the offer, then both people receive payments according to the offer. If the responder rejects the offer, then neither person receives anything (Güth, Schmittberger, and Schwarze 1982). The game theoretical analysis suggests that *rational* responders should accept any nonzero offer because receiving a small amount of money is better than receiving nothing. However, empirical studies suggest that responders tend to reject 'unfair' offers (e.g., less than 30% of the total), giving up monetary rewards in the process (Camerer 2003, 2011; Oosterbeek, Sloof, and van de Kuilen 2004). According to one prominent theory (Pillutla and Murnighan 1996), this deviation from optimality occurs because unfair offers elicit anger (see Barclay and Stoller 2014; Harris et al. 2020 for alternative explanations), making humans more willing to punish the proposer. In short, this game has become one of the gold-standard tasks to investigate how individuals integrate fairness concerns into their decision making (van Dijk and Dreu 2021).

Of course, the ultimatum game is only one of many approaches for studying human behavior and fairness perceptions (Yang and Stoyanovich 2017; Avi-Itzhak and Levy 2004). We acknowledge that this limits our ability to generalize our findings to the full scope of issues arising in the context of human-AI interactions and fairness research. However, we believe the ultimatum game provides a straightforward, scientifically rigorous, way to answer our research questions.

In this work, we leverage the ultimatum game to ask whether individuals are willing to incur monetary costs to instill fairness into AI. In two experiments, participants responded to fair and unfair offers made by AI or another human participant. We tested whether participants' perceptions of fairness would change if told their choices would train an AI proposer. In this paradigm, individuals willing to train AI to promote fairness should become more likely to reject unfair offers, thereby sacrificing their rewards.

We found that individuals were willing to incur a cost to train AI to make fair offers, regardless of whether individuals would encounter the AI in the future. In Experiment 1, participants in the 'AI training condition' accepted fewer unfair offers than participants in the control condition. In Experiment 2, we observed the same behavior in a group of participants who trained an AI they would never encounter again. These findings suggest that people are intrinsically motivated to modify their behavior when aware that AI is using their responses to learn, even willing to sacrifice their winnings to train AI to exhibit fairness towards others. Moreover, they indicate this behavior stems from a desire for fairness and not just maximizing personal gain. Stimuli, data, and analysis scripts from all experiments can be found on the Open Science Framework (OSF) [1].

Related Work

Our experiments expand on prior work that uses the ultimatum game to investigate how and when people change their perception of fairness. For example, there are many studies that show that humans' valuation of fairness changes when interacting with AI compared to human counterparts (Acosta-Mitjans et al. 2019; Di Dio et al. 2019; de Melo and Gratch 2015; Nishio et al. 2012; Swiderska, Krumhuber, and Kappas 2019; Sandoval et al. 2015; Tulk and Wiese 2018). The majority of these studies indicate that humans are more likely to accept unfair offers made by AI compared to human participants (Chen, Zhao, and Lai 2018; Moretti and di Pellegrino 2010; Sanfey et al. 2003; van 't Wout et al. 2006). However, Torta et al. (2013) discovered an opposite effect, with individuals rejecting unfair offers made by the computer more frequently than those made by human participants or robots. According to Torta et al. (2013), this discrepancy may have been driven by a relatively low emphasis on the nature of the partners compared to other studies (Moretti and di Pellegrino 2010; Sanfey et al. 2003). Consistent with this, there is a modest literature demonstrating that the way partners are presented influences acceptance rates in the ultimatum game. For example, humans reject offers more from in-group members than from out-group members (Mendoza, Lane, and Amodio 2014), from partners who show negative affect (Mussel, Göritz, and Hewig 2013), and from partners who are described as selfish (Marchetti et al. 2011). Here, we build upon this research by investigating whether humans respond differently to AI than other humans when presenting each partner type using anonymous silhouettes.

Our experiments also speak to a body of research that examines how behavior in the ultimatum game changes based on how AI partners are implemented. For example, de Melo, Gratch, and Carnevale (2014) manipulated the emotional response of AI responders. They found that participants were more inclined to make offers to AI that displayed emotion compared to AI that did not display emotion. This finding suggests that humans anthropomorphize AI, and change

[1] Link found here: https://osf.io/b7w5c

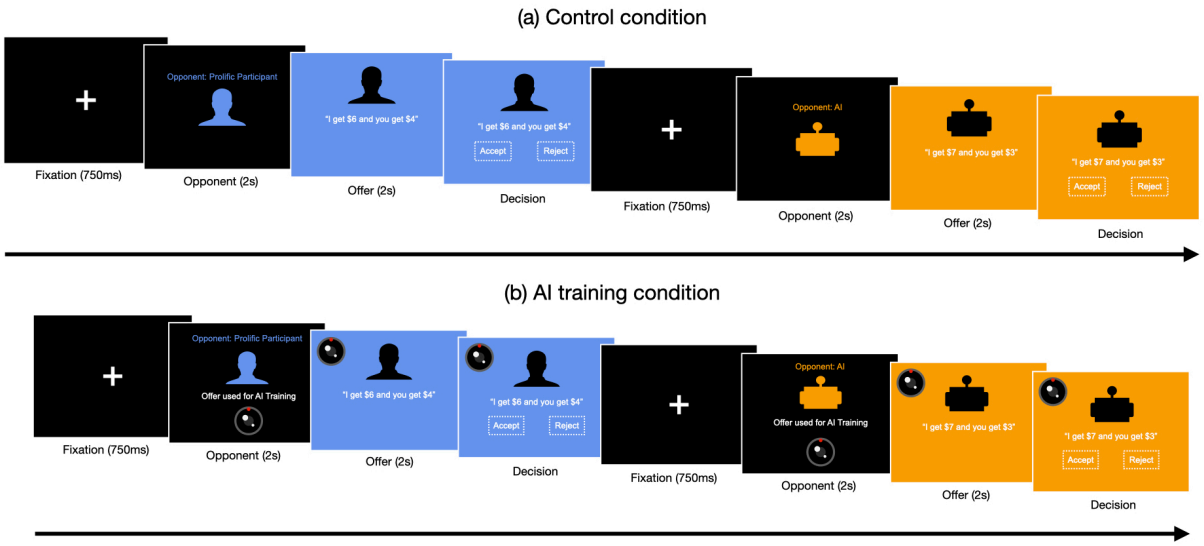


Figure 1: There were two possible conditions: control (a) and AI training (b). For participants assigned to the AI training condition, a webcam was shown to remind the participants that their responses were training AI. Participants in the control condition did not see a webcam since their responses 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. Once they made a decision, they started the next trial.

their own behavior based on how they think AI will respond (de Kleijn et al. 2019). Indeed, humans are more inclined to change their behavior when they know how AI is implemented. Russo, Duradoni, and Guazzini (2021) found that participants offer more money when partner with AI programmed to maximize their profits compared to randomly acting bots. Finally, humans are sensitive to the goal of AI in ultimatum games, exhibiting fairer behavior when interacting with computer agents that make decisions on behalf of human participants than when interacting with the participants themselves (de Melo, Marsella, and Gratch 2017).

In this study, we build on these findings by investigating how humans change behavior when aware their responses will be used for AI training. This research is crucial because even though previous research has already established AI algorithms that promote different levels of fairness (Añasco et al. 2023), it remains unknown how they deal with human biases, especially when AI is trained by human behavior.

Our work is also related to a recent literature in machine learning that describes the effects of strategic manipulation (Hardt et al. 2016; Perdomo et al. 2020; Chen, Liu, and Podimata 2020; Miller, Perdomo, and Zrnic 2021). For example, Hardt et al. (2016) study the design of AI algorithms when the training data is controlled by humans who can incur costs to manipulate their features. Related to this, Hu, Immorlica, and Vaughan (2019) and Milli et al. (2019) consider settings in which these costs for manipulation differ for different groups and explore the societal impacts of these differences. In this existing line of work, human behavior is always assumed to be rational. That is, people are expected to optimally respond to AI behavior. However, humans are known to consistently violate rationality assumptions. Here, we contribute to the understanding of how humans change their behavior when they are explicitly aware their behavior will be used to train AI. We believe this will help developers to design AI systems that are more robust, accounting for human behavior in the design of AI.

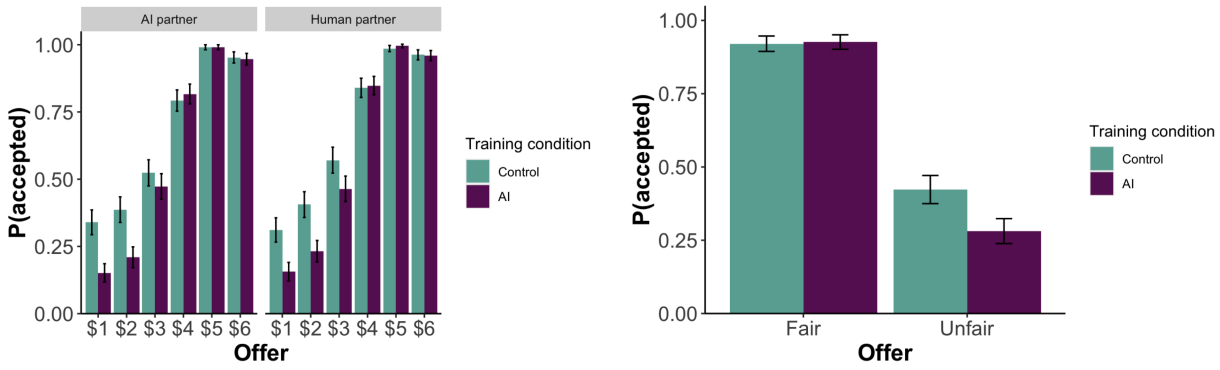
Experiment 1

In this study, we investigated whether individuals modify their behavior when they are aware that AI is observing their actions to learn how to make choices. To accomplish this, we used the ultimatum game, a well-known two-player economic bargaining game commonly employed to assess people's fairness attitudes (Güth, Schmittberger, and Schwarze 1982). In this task, participants played multiple rounds of the ultimatum game, partnered with either AI or another participant. They choose whether to reward or punish their partner based on the perceived fairness of the offers they received. One group of participants was informed that their responses would train an AI, which they would encounter later, while others were not provided with this information. The study was pre-registered on OSF [2].

Participants

A total of 217 participants (113 female, 3 non-binary, 1 missing; $M = 38.25$, $SD = 14.15$) were recruited from Prolific. Four participants were excluded from the analysis because they were exposed to both conditions since they refreshed the webpage and were assigned to a different condi-

[2]Link found here: https://osf.io/ajxk4



(a) The effect of offer amount.

(b) The effect of offer fairness.

Figure 2: Proportion accepted based on (a) offer amount (b) offer fairness conditioned on partner type and training condition for Experiment 1. Error bars indicate standard error.

tion than the original one. This experiment took 6 minutes to complete and the median pay rate for participants was approximately $10 per hour (all participants were paid $8.50 per hour before receiving a bonus). In this and the next experiment, all participants provided informed consent before completing each session, and the IRB of our institution approved this study.

Design

Participants were randomly assigned to either the 'AI training condition' ($n = 110$) or the 'control condition' ($n = 103$). All participants received instructions regarding the ultimatum game and were informed about the opportunity to participate in a follow-up session within the next few weeks. Participants in the AI training condition were specifically informed that their responses during this task would be used to train AI, which they would encounter during the follow-up session. However, they 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 were instructed to make decisions on how to divide a $10 sum with their partner. This partner was either AI or another participant recruited from a separate experiment. 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 condition 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 respectively. Participants were provided with unlimited time to make their decision.

Participants engaged in 24 rounds of the ultimatum game, playing 12 rounds with each partner type (AI and human participant). 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 (Moretti and di Pellegrino 2010).

To incentivize choice behavior, participants were informed that one trial would be randomly selected and resolved at the end of the experiment. They would receive a bonus of 5% of the amount they earned from the trial selected, and were informed that the bonus would increase to 15% in the follow-up session to encourage them to return[3].

Analysis For each trial, we measured whether participants accepted the offer depending on the offer amount, the fairness of the offer (categorized as fair: $4-$6, and unfair: $1-$3), partner type, and their training condition. We employed a logistic mixed-effects model to assess the factors that predict participants' acceptance of offers, including partner type, condition, offer amounts, and their interactions. Additionally, we conducted an ANOVA test to examine how the fairness of offers, partner type, condition, and their interactions influenced the likelihood of accepting offers Any significant interactions were interpreted using post-hoc t-tests.

Results

Effect of fairness of offer on accepting offers The mixed-effects model results indicated that participants were more likely to accept offers as the offer amount increased, ($\beta = 1.88,\ p < 0.001$). The results from the ANOVA were consistent with the model, finding that participants accepted more fair offers than unfair offers, ($F_{1,216} = 604,\ p < 0.001$). This general pattern replicates previous experiments

[3]The purpose of the follow-up session is mainly to provide stakes for participants to care about the AI trained on their data. Because the questions asked in this paper do not apply to this second session, we do not report the results here.

using the ultimatum game (Sanfey et al. 2003; Moretti and di Pellegrino 2010; van 't Wout et al. 2006).

Effect of AI training on accepting offers Next, we turned our attention to the effect of AI training. The mixed-effects model results revealed that participants in the AI training condition accepted less offers than participants in the control condition, ($\beta = -0.74,\ p < 0.001$). This main effect was classified by an interaction between training condition and offer amount, ($\beta = 0.18,\ p < 0.001$). The sign of this interaction effect indicates that participants in the AI training condition were more sensitive to the offer amount than participants in the control condition when deciding whether to reject the offer. A quick inspection of the results in Figure 2(a) suggests that this was mainly due to acceptance rates for unfair offers ($< \$4$) being lower in the AI training condition.

The ANOVA results provided evidence for this interpretation. Here, we found a significant interaction between training condition and offer fairness, ($F_{1,216} = 10.06,\ p = 0.002$). Specifically, participants in the AI training condition were more likely to reject unfair offers compared to participants in the control condition, ($t_{201} = 2.88,\ p = 0.004$), but no difference was found for fair offers, ($t_{207} = -0.31,\ p = 0.76$). These findings are shown in Figure 2(b).

Effect of partner on accepting offers The mixed-effects model results indicated that participants were less likely to reject offers when playing with an AI than when playing with other participants, ($\beta = -0.11,\ p = 0.04$). There was no significant interaction between partner type and offer amount, ($\beta = -0.05,\ p = 0.21$).

Interestingly, the ANOVA did not find a main effect of partner type, ($F_{1,216} = 0.04,\ p = 0.15$). This discrepancy may be due to the increased sensitivity of mixed-effects models or potentially a false positive in the former analysis. There was no interaction between partner type and fairness, ($F_{1,216} = 0.40,\ p = 0.53$).

Discussion

We found that participants in the AI training condition were willing to incur a personal cost, becoming more likely to reject unfair offers to train the AI to be fair. This effect may reflect an intrinsic motivation to make AI fairer when it learns by observation. However, participants in the AI training condition knew they would return for a follow-up session, facing the AI they trained with more rewards at stake. Therefore, the changes in behavior in the AI training condition may reflect a strategy to increase personal gains in this follow-up session rather than a genuine desire to foster fairness. We designed Experiment 2 to adjudicate between these hypotheses.

Experiment 2

Experiment 2 was designed to test whether participants would still be motivated to train AI, even if they didn't personally benefit from it. We replicated the design of Experiment 1 while introducing a third condition. In this new condition, participants were informed that their responses would train an AI they wouldn't encounter but that other participants would face in a follow-up session. By directly comparing this condition with an AI training condition that would face the AI they train, we could specifically test for altruistic motivation. The design and analysis of this experiment were preregistered on OSF [4].

Participants

A total of 339 participants (160 female, 10 non-binary, 1 missing; $M = 38.30$, $SD = 12.85$) were recruited from Prolific. Three participants were excluded from the analysis because they were exposed to both conditions since they refreshed the webpage and were assigned to a different condition than the original one. This experiment took 6 minutes to complete and the median pay rate for participants was approximately $\$10$ per hour all participants were paid $\$8.50$ per hour before receiving a bonus).

Design

The design was largely similar to Experiment 1. Importantly, we now randomly assigned participants to one of three conditions: 'AI training for self' ($n = 127$), 'AI training for others' ($n = 107$), and control ($n = 102$). The AI training for self condition was identical to the AI training condition of Experiment 1. Participants in this condition were informed that their responses would be used to train AI, which they would encounter again in a follow-up session. In the new AI training for others condition, participants were also told that their responses would be used to train AI. However, they were explicitly informed that (a) they would not personally encounter the AI trained on their data but rather (b) other participants would face it. Participants in the control condition were again not informed about any AI training.

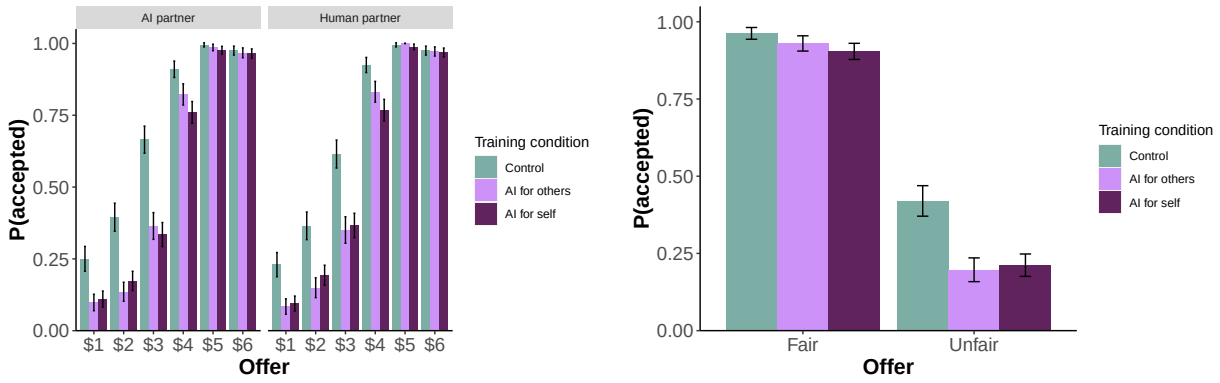
Participants assigned to the AI training for self and AI training for others conditions completed the same experiment as participants in the AI training condition of Experiment 1 (see Figure 1(a)). This involved seeing a webcam icon and the text "Offer used for AI Training" on each trial, reminding them that their responses would be used for AI training. In contrast, participants assigned to the control condition completed the same experiment as participants in the control condition of Experiment 1, with no webcam icon or mention of AI training (see Figure 1(b)).

Results

Analysis We used the same approach to answer our research questions, except the analysis now included three between-subject training conditions instead of two.

Effect of fairness of offer on accepting offers A mixed-effects model indicated that participants accepted more offers as the offer amount increased, ($\beta = 2.03,\ p < 0.001$). Analogously, an ANOVA showed that participants accepted more fair offers than unfair offers, ($F_{1,331} = 1562,\ p < 0.001$).

[4] Link found here: https://osf.io/krhz9



(a) The effect of offer amount.

(b) The effect of offer fairness.

Figure 3: Proportion accepted based on (a) offer amount (b) offer fairness conditioned on partner type and training condition for Experiment 2. Error bars indicate standard error.

Effect of AI training on accepting offers Next, we again examined the effect of AI training on offer acceptance. Results from the mixed-effects model showed that participants in both the AI training for self condition, ($\beta = -1.69,\ p < 0.001$), and AI training for others condition, ($\beta = -1.55,\ p < 0.001$), accepted less offers than participants in the control condition. Since the mixed effects model only showed how participants' responses differed from the control condition, we ran a second mixed effects model, changing the reference level to the AI training others condition. From the second mixed effects model, we found that participants in the AI training others condition accepted offers similarly to the AI training self condition, ($\beta = -0.14,\ p = 0.70$).

Additionally, we found that the interactions between the offer amount and training condition were significant when comparing the control and AI training for self conditions, ($\beta = 0.25,\ p = 0.045$), and control and AI training for others conditions ($\beta = 0.46,\ p < 0.001$). From the second mixed effects model, we found that the interaction was not significant when comparing the AI training conditions, ($\beta = -0.20,\ p = 0.13$). The sign of these first two regressor coefficients indicate that participants in the AI training conditions were more sensitive to the offer amount than participants in the control condition. Similar to Experiment 1, visual inspection of these results in Figure 3(a) suggests that this effect was driven by a reduced willingness to accept unfair offers in the AI training conditions.

The results from the ANOVA were consistent with this interpretation, ($F_{2,331} = 17.32,\ p < 0.001$) (see Figure 3(b)). We found a significant interaction between 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 other condition, ($t_{187} = 4.99,\ p < 0.001$). However, there was no statistical difference in acceptance rates between participants in the AI training conditions, ($t_{230} = -0.39,\ p = 0.69$).

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 less fair offers than the control condition. Note that in Experiment 1, we found no difference in acceptance rates for fair offers between these conditions. This difference is likely driven by participants' responses to $4 offers. As shown in Figure 3, participants in the control condition accepted $4 more than those in either AI training conditions, but there was no difference for higher offers. In Experiment 1, $4 offers were equally likely to be accepted between the training conditions. Additionally, there was no statistical difference in acceptance rates between the AI training conditions, ($t_{232} = 1.34,\ p = 0.18$).

Effect of partner on accepting offers Participants acceptance rates when playing with AI compared to other participants was not statistically different, ($\beta = 0.06,\ p = 0.52$). Additionally, there was no significant interaction between partner type and offer amount, ($\beta = -0.05,\ p = 0.45$) nor between partner type and fairness, ($F_{1,331} = 2.01,\ p = 0.16$).

Discussion

The results of this experiment provide strong evidence that people are willing to invest resources to train AI, even if they don't personally benefit. Participants that trained an AI for other participants incurred personal costs by rejecting unfair offers more than participants in the control condition. Strikingly, our results suggest that their acceptance rates did not differ from those of the participants that trained an AI for themselves. Behavior in this task was not just motivated by self-interest, but also by a desire to promote fairness and assist others.

General Discussion

In this section, we first recap the main findings of our paper. We then discuss the generalizability of the novel finding that individuals are willing to train AI systems to be fair and are

even willing to incur a cost to do so. Next, we discuss the limitations of our study. We then connect our findings related to human-AI interaction to previous research. Finally, we outline potential future work.

Recap and interpretation

Across two experiments, we found that people are willing to incur costs to train AI to be fair. In the first experiment, some participants were told their responses would train an AI they would encounter again. This group of participants rejected more unfair offers than participants in a control condition, who were not informed of any training. In other words, participants willingly sacrificed their bonuses to train AI to be fair. However, this result leaves unanswered whether they did this to make AI fair or to increase future rewards. Therefore, we ran a second experiment in which participants were told their responses would train an AI they would not encounter again but instead would face other people. Even in this condition, participants continued to reject more unfair offers than participants in the control condition. Moreover, they rejected unfair offers at the same rate as participants who trained AI for themselves. These findings suggest that people are intrinsically motivated to make AI fair, even if this comes at a personal cost and without any direct benefits.

Additionally, we found that participants did not accept more unfair offers when the proposer was an AI compared to when it was another human. This result differs from prior work (Moretti and di Pellegrino 2010; Sanfey et al. 2003; Torta et al. 2013; van 't Wout et al. 2006). We provide possible reasons for this discrepancy in our discussion later.

Implications for fairness in AI training

In this study, we found that individuals are motivated to train AI to be fair, even when they do not directly benefit from such training. This finding suggests a general inclination toward promoting fairness in AI training. While this motivation appears positive, it is imperative to consider that people have different definitions of fairness (Narayanan 2018). In the ultimatum game, there is a clear fairness notion that we can measure, such as the amount of money a responder is willing to forgo (Güth, Schmittberger, and Schwarze 1982; Camerer 2003, 2011; Oosterbeek, Sloof, and van de Kuilen 2004). However, in many real-world contexts, there is no agreed-upon definition of fairness when training AI (Kleinberg, Mullainathan, and Raghavan 2016; Narayanan 2018). For instance, there have been debates on which factors to prioritize to ensure that the process of allocating kidneys is fair. These factors include how long the person in need has been waiting, their age, and how urgent they need the kidney (Lee, Kanellis, and Mulley 2019).

This uncertainty invites debates on how AI should be trained to promote fairness. For instance, there have been debates on which factors AI algorithms should prioritize when predicting recidivism to achieve the highest levels of fairness (Angwin et al. 2016; Dieterich, Mendoza, and Brennan 2016). Even if humans are motivated to instill fairness into AI in more general contexts, different people might prioritize other notions of fairness. Therefore, it becomes crucial to moderate how individuals train AI to be fair, especially when their training objectives may diverge from the intended outcomes of the AI system.

This issue arises more generally. When given control over AI training, people can adapt their behavior based on how they want AI to act. As a result, individuals may behave differently during training than in a more natural setting. This bidirectionality leads to a complex interaction between human behavior and AI training that needs to be considered during the development and implementation of AI algorithms.

Limitations

Our study is one of the first to demonstrate that individuals are motivated to train AI to promote fairness. However, there are limitations worth mentioning. We can only generalize these findings to scenarios that resemble this game since we only conducted this study using the ultimatum game. As a result, it is unclear whether individuals are still motivated to train AI to be fair in other economic bargaining games, such as the dictator game (Engel 2011; Kahneman, Knetsch, and Thaler 1986) and the public goods game (Dawes 1980; Marwell and Ames 1979). By investigating this same question using other paradigms, we can determine whether our finding that individuals are motivated to train AI to promote fairness can be generalized to other contexts. For example, we can consider the role of AI training in real-world applications such as allocating kidneys to patients (Freedman et al. 2018) and resources to the homeless (Jo et al. 2022).

Connection to previous work

The main objective of our study was to determine whether individuals would train AI to be fair since this question has not been asked in the literature to our knowledge. To answer this question, we modified the ultimatum game, a well-estimated paradigm for measuring perceived fairness. Versions of this paradigm have repeatedly shown that people change their responses when interacting with AI compared to another individual (Moretti and di Pellegrino 2010; Sanfey et al. 2003; van 't Wout et al. 2006). Our design allowed us to ask the same question in addition to our main research question. Interestingly, participants in our task did not adapt their behavior when playing with AI compared to other human participants. This finding is inconsistent with several results from prior work (Chen, Zhao, and Lai 2018; Moretti and di Pellegrino 2010; Sanfey et al. 2003; Torta et al. 2013; van 't Wout et al. 2006) As mentioned before, some of these studies found that individuals are more inclined to reject unfair offers when interacting with another individual (Chen, Zhao, and Lai 2018; Moretti and di Pellegrino 2010; Sanfey et al. 2003; van 't Wout et al. 2006), while others found a reversed effect, that people accept more unfair offers when playing with another person (Torta et al. 2013). We discuss two hypotheses that explain the lack of effect of partner type on acceptance rates.

First, unlike previous studies where the proposer was personified in various ways (e.g., an actor sitting across from the participant (Moretti and di Pellegrino 2010)), our study only presented participants with an outline of a human profile. This design may not have evoked the same emotional

engagement participants have felt in prior work. People may need to see a photo of their partner or physically meet them before they attach meaning to their responses. The more anonymous nature of our study may not have triggered a negative affective reaction (Pillutla and Murnighan 1996) in response to low offers.

Second, our study was conducted entirely online, unlike the laboratory-based settings of other studies. Since it's plausible that individuals could not be as motivated when completing a study online than in-person (Paolacci, Chandler, and Ipeirotis 2010), participants may have responded differently when playing with AI than humans if they completed this task in person. However, we should note participants still were sensitive to both the offer amount and the AI training conditions. Therefore, this lack of motivation should specifically target their sensitivity to the nature of the partner. It's also possible that the online nature of this study reduced demand characteristics, so participants might not have felt obligated to differ their responses when playing with AI than with another participant (Coles and Frank 2023).

Future Work

Our work shows that people are motivated to train AI to be fair. This finding sheds light on a new, intriguing question: Why are they willing to do so? As discussed above, one possible answer is that people have a general inclination toward promoting fairness in the AI training process. However, several alternative explanations come to mind. Here, we discuss them and suggest future work to distinguish between them. First, individuals may have been willing to train AI to help others for reciprocal reasons. Therefore, individuals may have trained AI to be fair, assuming that others would train AI for them to return the favor. This way, everyone would benefit from this training. This approach calls to mind the idiom "I scratch your back, you scratch mine." If this were the case, individuals would not train AI to be fair if there was no subsequent session to recap the reciprocated benefits. To test this hypothesis, one could only have individuals complete our experiment once, removing the possibility of benefiting from others' actions. If participants no longer train AI to be fair in this scenario, then this would support the notion that individuals train AI fairly for reciprocal reasons. Conversely, if they continued to promote fairness, then one could reasonably infer that people are genuinely motivated to train AI so that it will treat others fairly.

Another conjecture is that the reward itself could influence people's inclination to train AI for fairness. Specifically, individuals may train AI to exhibit fair behavior since the rewards were relatively small. Indeed, people accept more unfair offers when more reward is at stake (Andersen et al. 2011; Slonim and Roth 1998; Novakova and Flegr 2013). Thus, the small rewards in our study (maximum of $15¢$) may not have been enticing enough for individuals to prioritize personal gains. It would be interesting to determine whether individuals would still be willing to forgo their rewards in high stake scenarios. Systematically manipulating the reward at stake might reveal a threshold at which personal rewards outweigh the act of training AI to be fair.

In addition to understanding what motivates people to train AI to be fair, future work should test whether people will train AI to make choices that align with their preferences. Of course, in the current task, individuals could train AI to be fair because, presumably, they believe that AI should be equitable. To extend this hypothesis, future research could test whether people are motivated to train AI to provide recommendations that align with their own beliefs, even if those are not rationally optimal. This would result in AI systems that agree with individual preferences, reinforcing confirmation bias (Wason and Johnson-Laird 1972; Peters 2020). In real-life situations, this would carry risks in various domains. For example, in the legal field, AI trained in this way may exacerbate biases in judges' decision-making (Angwin et al. 2016). In medicine, where doctors may train AI to validate their diagnoses, this would potentially limit their openness to considering alternative evidence that points to different conclusions (Bornstein and Emler 2001). These observations reinforce the idea that it is critical to understand the motives and goals of individuals when they engage in AI training. After all, people may be inclined to train AI to align with their preferences, and those preferences may not be optimal. Pertinent to the current experiments, people may be inclined to train AI based on what they consider to be fair, but these fairness preferences may not be optimal or shared with other people.

It is important to note that even though we told some participants that their responses would be used for AI training, we did not tell them how the AI would use their choices to learn. A wide variety of learning algorithms would have been possible. For example, AI may have used the participants' choice to learn how to maximize its own reward, minimize the participants' reward, or make fair offers. Each of those requires different optimization procedures. Because we left this open to interpretation, our experiment critically relied on how participants think AI will use their data. What would happen when participants know how AI will learn from their data? We hypothesize they will adapt their behavior to the stated strategy (Russo, Duradoni, and Guazzini 2021). For example, if participants are told that AI aims to maximize its own rewards, then they may become more punitive compared to when AI tries to match perceived fairness. Future research could explore how informing individuals of the goals of AI influences their choices, and this affects how AI systems that learn from these choices.

Conclusion

This study used the ultimatum game to examine whether individuals are inclined to train AI to make equitable offers. The results indicate that individuals are motivated to train AI to prioritize fairness, not only for their advantage but also for the benefit of others. This finding suggests that, when given the opportunity, individuals will train AI based on their preferences. In this case, they preferred to train AI to be fair. The fact that people are motivated to train AI based on their preferences presents a challenge for AI development because humans are biased. Our results, therefore, suggest that developers of AI algorithms should consider how their training regimes adopt human bias.

Acknowledgements

We would like to thank members of the Control and Decision Making Lab and 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.

References

Acosta-Mitjans, A.; Cruz-Sandoval, D.; Hervas, R.; Johnson, E.; Nugent, C.; and Favela, J. 2019. Affective Embodied Agents and Their Effect on Decision Making. In *13th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI 2019*. MDPI.

Añasco, J.; Navas, B. J. N.; Mora, P. A. P.; and Kramskova, M. A. V. 2023. Simulation of ultimatum game with artificial intelligence and biases. *ACI Avances en Ciencias e Ingenierías*, 15(1).

Andersen, S.; Ertaç, S.; Gneezy, U.; Hoffman, M.; and List, J. A. 2011. Stakes Matter in Ultimatum Games. *American Economic Review*, 101(7): 3427–3439.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23: 77–91.

Avi-Itzhak, B.; and Levy, H. 2004. On measuring fairness in queues. *Advances in Applied Probability*, 36(3): 919–936.

Azizi, M. J.; Vayanos, P.; Wilder, B.; Rice, E.; and Tambe, M. 2018. Designing Fair, Efficient, and Interpretable Policies for Prioritizing Homeless Youth for Housing Resources. In *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, 35–51. Springer International Publishing.

Bansal, G.; Nushi, B.; Kamar, E.; Weld, D. S.; Lasecki, W. S.; and Horvitz, E. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2429–2437.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. S. 2020. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *CoRR*, abs/2006.14779.

Barclay, P.; and Stoller, B. 2014. Local competition sparks concerns for fairness in the ultimatum game. *Biology Letters*, 10(5): 20140213.

Bayati, M.; Braverman, M.; Gillam, M.; Mack, K. M.; Ruiz, G.; Smith, M. S.; and Horvitz, E. 2014. Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study. *PLoS ONE*, 9(10): e109264.

Bornstein, B. H.; and Emler, A. C. 2001. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of Evaluation in Clinical Practice*, 7(2): 97–107.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. PMLR.

Camerer, C. F. 2003. Strategizing in the Brain. *Science*, 300(5626): 1673–1675.

Camerer, C. F. 2011. *Behavioral game theory: Experiments in strategic interaction*. Princeton university press.

Cazes, M.; Franiatte, N.; Delmas, A.; André, J.; Rodier, M.; and Kaadoud, I. C. 2021. 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)*.

Chen, M.; Zhao, Z.; and Lai, H. 2018. The time course of neural responses to social versus non-social unfairness in the ultimatum game. *Social Neuroscience*, 14(4): 409–419.

Chen, Y.; Liu, Y.; and Podimata, C. 2020. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33: 15265–15276.

Cohn, A.; Gesche, T.; and Maréchal, M. A. 2022. Honesty in the Digital Age. *Management Science*, 68(2): 827–845.

Coles, N. A.; and Frank, M. C. 2023. A quantitative review of demand characteristics and their underlying mechanisms. *PsyArXiv*.

Dawes, R. M. 1980. Social dilemmas. *Annual review of psychology*, 31(1): 169–193.

de Kleijn, R.; van Es, L.; Kachergis, G.; and Hommel, B. 2019. Anthropomorphization of artificial agents leads to fair and strategic, but not altruistic behavior. *International Journal of Human-Computer Studies*, 122: 168–173.

de Melo, C.; Gratch, J.; and Carnevale, P. 2014. The Importance of Cognition and Affect for Artificially Intelligent Decision Makers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

de Melo, C.; Marsella, S.; and Gratch, J. 2016. People Do Not Feel Guilty About Exploiting Machines. *ACM Transactions on Computer-Human Interaction*, 23(2): 1–17.

de Melo, C. M.; and Gratch, J. 2015. People show envy, not guilt, when making decisions with machines. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 315–321.

de Melo, C. M.; Marsella, S.; and Gratch, J. 2017. Social decisions and fairness change when people's interests are represented by autonomous agents. *Autonomous Agents and Multi-Agent Systems*, 32(1): 163–187.

Di Dio, C.; Manzi, F.; Itakura, S.; Kanda, T.; Ishiguro, H.; Massaro, D.; and Marchetti, A. 2019. It Does Not Matter Who You Are: Fairness in Pre-schoolers Interacting with Human and Robotic Partners. *International Journal of Social Robotics*, 12(5): 1045–1059.

Dieterich, W.; Mendoza, C.; and Brennan, T. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4): 1.

Engel, C. 2011. Dictator games: a meta study. *Experimental Economics*, 14(4): 583–610.

Erlei, A.; Das, R.; Meub, L.; Anand, A.; and Gadiraju, U. 2022. 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.

Freedman, R.; Borg, J. S.; Sinnott-Armstrong, W.; Dickerson, J.; and Conitzer, V. 2018. Adapting a Kidney Exchange Algorithm to Align With Human Values. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Giordano, C.; Brennan, M.; Mohamed, B.; Rashidi, P.; Modave, F.; and Tighe, P. 2021. Accessing Artificial Intelligence for Clinical Decision-Making. *Frontiers in Digital Health*, 3.

Güth, W.; Schmittberger, R.; and Schwarze, B. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior &amp Organization*, 3(4): 367–388.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

Harris, A.; Young, A.; Hughson, L.; Green, D.; Doan, S. N.; Hughson, E.; and Reed, C. L. 2020. Perceived relative social status and cognitive load influence acceptance of unfair offers in the Ultimatum Game. *PLOS ONE*, 15(1): e0227717.

Hayashi, Y.; and Wakabayashi, K. 2017. 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.

Hu, L.; Immorlica, N.; and Vaughan, J. W. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 259–268.

Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; and Wang, Y. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4): 230–243.

Jo, N.; Tang, B.; Dullerud, K.; Aghaei, S.; Rice, E.; and Vayanos, P. 2022. Fairness in Contextual Resource Allocation Systems: Metrics and Incompatibility Results. arXiv:2212.01725.

Kahneman, D.; Knetsch, J. L.; and Thaler, R. H. 1986. Fairness and the Assumptions of Economics. *The Journal of Business*, 59(S4): S285.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Koh, D.-M.; Papanikolaou, N.; Bick, U.; Illing, R.; Kahn, C. E.; Kalpathi-Cramer, J.; Matos, C.; Martí-Bonmatí, L.; Miles, A.; Mun, S. K.; Napel, S.; Rockall, A.; Sala, E.; Strickland, N.; and Prior, F. 2022. Artificial intelligence and machine learning in cancer imaging. *Communications Medicine*, 2(1).

Kube, A.; Das, S.; and Fowler, P. J. 2019. Allocating Interventions Based on Predicted Outcomes: A Case Study on Homelessness Services. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 622–629.

Lai, V.; Carton, S.; Bhatnagar, R.; Liao, Q. V.; Zhang, Y.; and Tan, C. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. ACM.

Lee, D.; Kanellis, J.; and Mulley, W. R. 2019. Allocation of deceased donor kidneys: A review of international practices. *Nephrology*, 24(6): 591–598.

Marchetti, A.; Castelli, I.; Harlé, K. M.; and Sanfey, A. G. 2011. Expectations and outcome: The role of Proposer features in the Ultimatum Game. *Journal of Economic Psychology*, 32(3): 446–449.

Marwell, G.; and Ames, R. E. 1979. Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem. *American Journal of Sociology*, 84(6): 1335–1360.

Mathur, V.; Stavrakas, Y.; and Singh, S. 2016. Intelligence analysis of Tay Twitter bot. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 231–236. IEEE.

Mendoza, S. A.; Lane, S. P.; and Amodio, D. M. 2014. For Members Only. *Social Psychological and Personality Science*, 5(6): 662–670.

Miller, J. P.; Perdomo, J. C.; and Zrnic, T. 2021. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, 7710–7720. PMLR.

Milli, S.; Miller, J.; Dragan, A. D.; and Hardt, M. 2019. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 230–239.

Moretti, L.; and di Pellegrino, G. 2010. Disgust selectively modulates reciprocal fairness in economic interactions. *Emotion*, 10(2): 169–180.

Mussel, P.; Göritz, A. S.; and Hewig, J. 2013. The value of a smile: Facial expression affects ultimatum-game responses. *Judgment and Decision Making*, 8(3): 381–385.

Narayanan, A. 2018. Tutorial: 21 fairness definition and their politics. Presented at ACM FAT (Fairness, Accountability and Transparency) Conference 2018.

Nishio, S.; Ogawa, K.; Kanakogi, Y.; Itakura, S.; and Ishiguro, H. 2012. Do robot appearance and speech affect people's attitude? Evaluation through the Ultimatum Game. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 809–814.

Novakova, J.; and Flegr, J. 2013. 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(4): e60966.

Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.-E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; Kompatsiaris, I.; Kinder-Kurlanda, K.; Wagner, C.; Karimi, F.; Fernandez, M.; Alani, H.; Berendt, B.; Kruegel, T.; Heinze, C.; Broelemann, K.; Kasneci, G.; Tiropanis, T.; and Staab, S. 2020. Bias in Data-driven AI Systems – An Introductory Survey. arXiv:2001.09762.

Oosterbeek, H.; Sloof, R.; and van de Kuilen, G. 2004. Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis. *Experimental Economics*, 7(2): 171–188.

Paolacci, G.; Chandler, J.; and Ipeirotis, P. G. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5): 411–419.

Perdomo, J.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2020. Performative prediction. In *International Conference on Machine Learning*, 7599–7609. PMLR.

Peters, U. 2020. What Is the Function of Confirmation Bias? *Erkenntnis*, 87(3): 1351–1376.

Pillutla, M. M.; and Murnighan, J. 1996. Unfairness, Anger, and Spite: Emotional Rejections of Ultimatum Offers. *Organizational Behavior and Human Decision Processes*, 68(3): 208–224.

Russo, P. A.; Duradoni, M.; and Guazzini, A. 2021. How self-perceived reputation affects fairness towards humans and artificial intelligence. *Computers in Human Behavior*, 124: 106920.

Sandoval, E. B.; Brandstetter, J.; Obaid, M.; and Bartneck, C. 2015. Reciprocity in Human-Robot Interaction: A Quantitative Approach Through the Prisoner's Dilemma and the Ultimatum Game. *International Journal of Social Robotics*, 8(2): 303–317.

Sanfey, A. G.; Rilling, J. K.; Aronson, J. A.; Nystrom, L. E.; and Cohen, J. D. 2003. The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science*, 300(5626): 1755–1758.

Slonim, R.; and Roth, A. E. 1998. Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic. *Econometrica*, 66(3): 569.

Soleimani, M.; Intezari, A.; and Pauleen, D. J. 2021. Mitigating Cognitive Biases in Developing AI-Assisted Recruitment Systems. *International Journal of Knowledge Management*, 18(1): 1–18.

Soleimani, M.; Intezari, A.; Taskin, N.; and Pauleen, D. 2021. Cognitive biases in developing biased Artificial Intelligence recruitment system. *Hawaii International Conference on System Sciences*, 54: 5091–5099.

Swiderska, A.; Krumhuber, E. G.; and Kappas, A. 2019. Behavioral and Physiological Responses to Computers in the Ultimatum Game. *International Journal of Technology and Human Interaction*, 15(1): 33–45.

Torta, E.; van Dijk, E.; Ruijten, P. A. M.; and Cuijpers, R. H. 2013. The Ultimatum Game as Measurement Tool for Anthropomorphism in Human–Robot Interaction. In *Social Robotics*, 209–217. Springer International Publishing.

Tulk, S.; and Wiese, E. 2018. Trust and Approachability Mediate Social Decision Making in Human-Robot Interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1): 704–708.

van Dijk, E.; and Dreu, C. K. D. 2021. Experimental Games and Social Decision Making. *Annual Review of Psychology*, 72(1): 415–438.

van 't Wout, M.; Kahn, R. S.; Sanfey, A. G.; and Aleman, A. 2006. Affective state and decision-making in the Ultimatum Game. *Experimental Brain Research*, 169(4): 564–568.

Wason, P. C.; and Johnson-Laird, P. N. 1972. *Psychology of Reasoning: Structure and Content*. Cambridge, MA, USA: Harvard University Press.

Wolf, M. J.; Miller, K.; and Grodzinsky, F. S. 2017. Why we should have seen that coming: comments on Microsoft's tay" experiment," and wider implications. *Acm Sigcas Computers and Society*, 47(3): 54–64.

Yang, K.; and Stoyanovich, J. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM.

Završnik, A. 2020. Criminal justice, artificial intelligence systems, and human rights. *ERA Forum*, 20(4): 567–583.