# Do People Think Fast or Slow When Training AI?

Lauren S. Treiman
Washington University in St. Louis
St. Louis, Missouri, USA
ltreiman@wustl.edu

Chien-Ju Ho*
Washington University in St. Louis
St. Louis, Missouri, USA
chienju.ho@wustl.edu

Wouter Kool*
Washington University in St. Louis
St. Louis, Missouri, USA
wkool@wustl.edu

## Abstract

Artificial intelligence (AI) plays a crucial role in decision making. In doing so, it often learns to make choices from human behavior, assuming that people provide unbiased training data. However, studies show that people change their behavior when they are aware they are training AI. It remains unknown whether these modifications are intuitive, driven by social norms, or deliberate, aimed at maximizing personal gain. Across three experiments, we investigated the extent people deliberate when training AI using the ultimatum game. In this game, participants decided whether to accept monetary rewards. Some participants were informed they would train an AI to respond to their or other participants' proposals made in a follow-up session, while others were not. Those training AI could intuitively reject unfair offers or deliberately accept them to maximize current and future rewards. We found that participants rejected unfair offers, suggesting they were more inclined to rely on intuition when training AI. This reveals that people often embed their biases into AI, posing a challenge for AI designed to make optimal decisions.

## CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → **Psychology**.

## Keywords

AI training, cognitive processing, ultimatum game, decision making

**ACM Reference Format:**
Lauren S. Treiman, Chien-Ju Ho, and Wouter Kool. 2025. Do People Think Fast or Slow When Training AI?. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3715275.3732177

## 1 Introduction

AI is increasingly used in decision making in various domains, including medicine [8, 36, 49, 59] and law [2, 43, 115]. These AI models are often trained on human decisions, assuming this training data is unbiased [79]. Recent work [105, 106] has challenged this assumption, showing that people change their behavior when aware that AI is learning from them. However, it remains unclear how people choose to modify their behavior when training AI. On the one hand, people may deliberately plan their actions using an internal model of how the AI algorithm works [16]. On the other hand, people may rely on intuitions and heuristics [7, 35, 52, 78] when training AI, which provide quick but often inaccurate solutions. Indeed, humans tend to avoid mentally effortful tasks [20, 60, 61, 63, 80], suggesting they may default to using intuition rather than exerting effort to understand how AI learns. Here, we investigate how people deploy these strategies when modifying their behavior to train AI.

Psychological research on dual-system theories [4, 30, 51, 95, 98] provides a framework for considering how people use heuristics and deliberation when training AI. These theories propose that decision making operates through two distinct systems: a fast, automatic system that relies on instinct or habit and a slow, controlled system that plans toward goals. These systems embody different accuracy-demand trade-offs. The fast system has low computational demands but is less accurate, whereas the slow system is more accurate but more computationally demanding. In this work, we apply these theories to the context of AI training. People may rely on intuition, training AI based on lay theories of how to respond or how AI works without fully simulating the future consequences of its deployment. Alternatively, they may deliberate, carefully adapting their behavior to maximize future rewards for themselves or others.

We aimed to determine how people choose between intuition and deliberation when training AI. Our first research question was:

**Do humans rely more on intuition or deliberation when training AI? (Experiment 1)**

Because we found evidence that people mostly relied on intuition when training AI, we next explored whether participants could be encouraged to use more deliberation:

**Can humans be moved to deliberate during AI training? (Experiments 2 & 3)**

To answer our research questions, we used the ultimatum game [41], following prior work in this domain [105, 106]. 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. If the responder accepts the offer, both players receive payments according to the offer. If the responder rejects the offer, neither player receives anything. While game theoretical analysis suggests that rational responders should accept any nonzero offer, empirical studies show that people often reject unfair offers (e.g., less than 30% of the total amount) [13, 82]. This game is widely used to study how fairness concerns influence decision making [82, 110]. While we recognize that the ultimatum game may not reflect how people typically interact with AI in real-world settings (e.g., human annotation tasks [76]), we believe it provides a straightforward and controlled framework for answering our research questions.

*Both authors contributed equally to this research.


*FAccT '25, June 23–26, 2025, Athens, Greece*

ACM ISBN 979-8-4007-1482-5/2025/06
https://doi.org/10.1145/3715275.3732177

To determine whether people rely on intuition or deliberation, we incorporated AI training into the ultimatum game. This approach builds on the work of Treiman et al. [105, 106], who found that participants were more likely to reject unfair offers when training an AI *proposer* that either they or other participants would encounter in a higher-stakes follow-up session. While this behavior could reflect deliberate planning to teach AI to make fair proposals to them in the follow-up session, it may also stem from an instinct to punish unfairness [14, 37]. Thus, since both intuition and deliberation lead to the same behavior in this prior work, it remains unclear which strategy people use when training AI.

Here, we designed our task so intuition and deliberation predict different behaviors. To do this, some participants were informed they were training an AI *responder* that either they or other participants would make proposals to in a higher-stakes follow-up session. In this task, participants aiming to maximize rewards should accept unfair offers. This increases their reward in both the current and follow-up sessions, as they teach AI to accept their future unfair proposals and reap the rewards of accepting any offer right now. However, this approach requires participants to overcome their innate tendency to punish unfair behavior through rejecting unfair offers [14, 37]. By examining whether participants became more likely to accept or reject unfair offers when training AI, we inferred whether they relied more on intuition or deliberation during the training process. While this design assumes that deliberation leads participants to maximize rewards, we realize that it is possible that participants in this task may deliberately promote fairness by punishing more unfair offers. To address this, we ran additional studies designed to minimize fairness concerns. Thus, across these studies, we reasoned that rejecting unfair offers indicated a greater reliance on intuition, whereas accepting unfair offers suggested a more deliberate (goal-directed) approach.

To investigate whether people use intuition or deliberation when training AI, we conducted three experiments with the paradigm described above. In Experiment 1, we hypothesized that participants would deliberate when training AI, accepting unfair offers to teach AI to accept their future, unfair proposals. However, we found that participants rejected more unfair offers, suggesting that they relied on intuition when training AI. To explore whether we could encourage deliberation, we designed Experiment 2, which was identical to Experiment 1 but included an extensive comprehension test to assess participants' understanding of the task and the training process. We hypothesized that this intervention would prompt participants to recognize that accepting unfair offers would lead to maximizing rewards. Despite this, participants continued to reject more unfair offers when training AI. While we reasoned that rejecting unfair offers reflects a reliance on intuition [14, 37], it is also possible that participants in Experiment 2 deliberately chose to prioritize fairness [48]. To rule out this explanation, we conducted Experiment 3. This experiment followed the same design as before, except participants trained an AI that only they would encounter in the follow-up session. In this context, training the AI for fairness was irrelevant since the AI would not interact with anyone but the participant who performed the training and, therefore, could not promote fairness. We hypothesized that participants would start to accept more unfair offers in this training context, showing signs of deliberation. Yet again, we found that participants rejected more unfair offers when training AI, suggesting that they were still relying on intuition.

Across three experiments, we found that participants not only changed their behavior when training AI, replicating prior work [105, 106], but mostly relied on intuition to do so. This suggests that people may unintentionally instill their biases into training data, which can lead to unreliable and discriminatory AI outcomes [6, 24]. These findings offer valuable insight into current crowdsourcing methods used to train AI and underscore the importance of understanding how people provide training data. They may also help AI developers identify when people are likely to introduce biases into AI algorithms and account for them when designing AI.

## 2 Related Work

### 2.1 Behavior Modification in AI Training

Our approach directly builds on the work of Treiman et al. [105, 106], who used the ultimatum game to provide some of the first evidence that humans change their behavior when training AI. In their study, some participants were informed they were training an AI *proposer* that they or others would encounter in a follow-up session with higher stakes (Figure 1a top row). They found that participants aware of AI training were more inclined to reject unfair offers to teach AI to make fairer proposals. While people may have strategically rejected unfair offers to teach AI to make fair proposals to them in the follow-up session (i.e., using deliberation), there is also an alternative explanation: people may have relied on intuition when rejecting unfair offers. Specifically, people's instinct to punish unfairness [14, 37] may have driven them to reject unfair offers, teaching AI to make fair proposals without considering the consequences of the deployed AI. This strategy closely resembles intuitive reinforcement learning [23, 66], where actions are driven by automatic responses rather than careful deliberation. Because intuition (prioritizing fairness) and deliberation (maximizing rewards) predict the same behavior (punishing unfair offers) (Figure 1b top row) in this task, it is unclear which strategy people used.

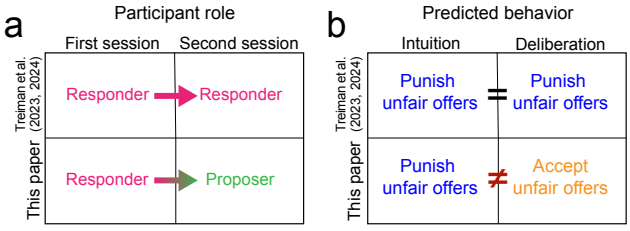


**Figure 1: Task design comparisons between Treiman et al. [105, 106] and this paper on (a) participant's role and (b) predicted behavior based on whether people use intuition or deliberation when training AI.**

Here, we modified this task so these strategies predicted different behaviors, allowing us to assess how people tradeoff between them. In this new task, all participants played as the responder in the first session, but were now told they would switch roles and play as the proposer in the follow-up session (Figure 1a bottom row). As a result, some participants were told they would train an AI *responder*

(instead of a proposer [105, 106]), which they or others would make proposals to in the higher-stakes follow-up session. Participants in the control condition did not receive this information.

In this task, participants could still rely on their innate tendency to reject more unfair offers [14, 37] when training AI. However, contrary to prior work by Treiman et al. [105, 106], in this paradigm this strategy does not maximize reward: rejecting unfair offers teaches AI to reject participants' future unfair proposals. Instead, participants should accept unfair offers during training, teaching AI to accept them as well. This strategy maximizes reward not only in the current session but also in the follow-up session, as participants can exploit AI by making unfair proposals that are likely to be accepted. In this modified task, intuition and deliberation predict different behaviors (Figure 1b bottom row), enabling us to determine whether people deliberate when training AI.

## 2.2 Cost-Benefit Analysis in AI Training

It has been well-established that people flexibly trade off between automatic and controlled strategies using a cost-benefit analysis [62, 92, 93, 112]. People exploit task structures to their advantage if it is worthwhile [12, 54, 62, 75, 86], even in contexts that involve human-AI interactions [22, 28, 55]. However, when the effort cost of planning becomes too high, people shy away from deliberate lines of actions [57, 61, 88]. Consequently, people may apply this same cost-benefit tradeoff when deciding how to train AI. Here, we manipulate both the costs (e.g., mental effort to understand the algorithm) and benefits (e.g., potential rewards) to examine how people choose to train AI.

We hypothesize that the complexity of the AI training algorithm (costs) affects people's motivation to deliberate. Since individuals generally avoid mental effort [57, 61, 88], they may only engage in deliberation if the AI training process aligns with intuitive learning patterns. For example, previous work [105, 106] used reinforcement learning [40, 70], where the AI adapted to participants' behavior to maximize rewards. This method parallels human reinforcement-learning processes [23, 66], thereby lowering deliberation costs due to its intuitive nature. In contrast, this study investigates how people navigate less intuitive AI training processes and whether they rely more on deliberation or intuition.

We also assess how the potential rewards of exploiting AI (benefits) influences people's willingness to engage in goal-directed behavior. Since people engage more in analytical processing when higher rewards are involved [10, 67, 69, 83], they may adjust their deliberation efforts based on the potential benefits of AI training. Here, we manipulate who benefits from the training to determine how people weigh rewards when deciding how to train AI.

## 2.3 Behavior Shifts in Human-AI Interactions

When people are informed about how an algorithm will make decisions, they often change their behavior to "game the system" [19, 42, 84, 102], a phenomenon known as Goodhart's Law [39]. For example, Camacho and Conover [12] showed that people in Colombia reported exaggerated financial needs to just qualify for aid once they learned the rules of social welfare distribution. Similarly, people strategically change their behavior to influence AI recommendations. For instance, Cen et al. [16] showed that people adapted their behavior to match how the AI learns: they liked more content when informed that recommendations were based on likes, and paused longer when told recommendations were based on dwell time. In these cases, people are informed about how the AI algorithm will use their data, eliminating the need for deliberation about optimal training strategies. We extend this research by investigating whether people rely on intuition or deliberation when modifying their behavior during AI training, specifically when they are not explicitly informed about the training process.

## 2.4 Human-AI Interactions in Ultimatum Game

Prior work has investigated how people consider fairness when interacting with AI compared to human counterparts in the context of the ultimatum game [1, 27, 29, 81, 90, 99, 107]. Most of these studies indicate that people are more likely to accept unfair offers from AI than from humans [18, 77, 91, 111]. However, Torta et al. [104] found the opposite, with individuals rejecting unfair computer offers more frequently. Additionally, Treiman et al. [105, 106] found no difference in responses when humans played with an AI compared to a human participant. This discrepancy may stem from a relatively low emphasis on the nature of the partners compared to other studies [77, 91, 104].

In our task, we present partner types as anonymous silhouettes, similar to Treiman et al. [105, 106]. This design choice avoids drawing attention to the partner type since our main goal is to understand how people's behavior changes due to AI training, not the type of partner. In fact, we include AI partners to help confirm that AI training is taking place. Although our focus is not on partner effects, this approach allows us to contribute to the literature by exploring how interactions with AI differ from those with humans when partner type is minimally emphasized.

## 2.5 Human's Internal Models of AI Learning

"Theory of mind" refers to the ability to understand and infer the mental states of others [68, 87]. This cognitive process is not only limited to interactions with fellow humans but also extends to interactions with AI [5, 26, 56, 58]. However, people exert less cognitive effort to infer the mental states of AI compared to humans [64, 89]. For example, McCabe et al. [72] used the trust game [50] to show that brain activation in the medial prefrontal cortex (a brain region associated with theory of mind) occurs only when interacting with humans, not AI. Gallagher et al. [34] found analogous results in the game rock-paper-scissors. Our research builds on these findings to explore what internal models of AI people form when they are not directly interacting with it.

# 3 Experiment 1: Do Humans Rely More on Intuition or Deliberation When Training AI?

We investigated whether people use intuition or deliberation when training AI using the ultimatum game [41]. In this task, participants played multiple rounds as the responder, partnered with either AI or another participant. We told all participants that they would be invited to a follow-up session where they would make proposals instead.

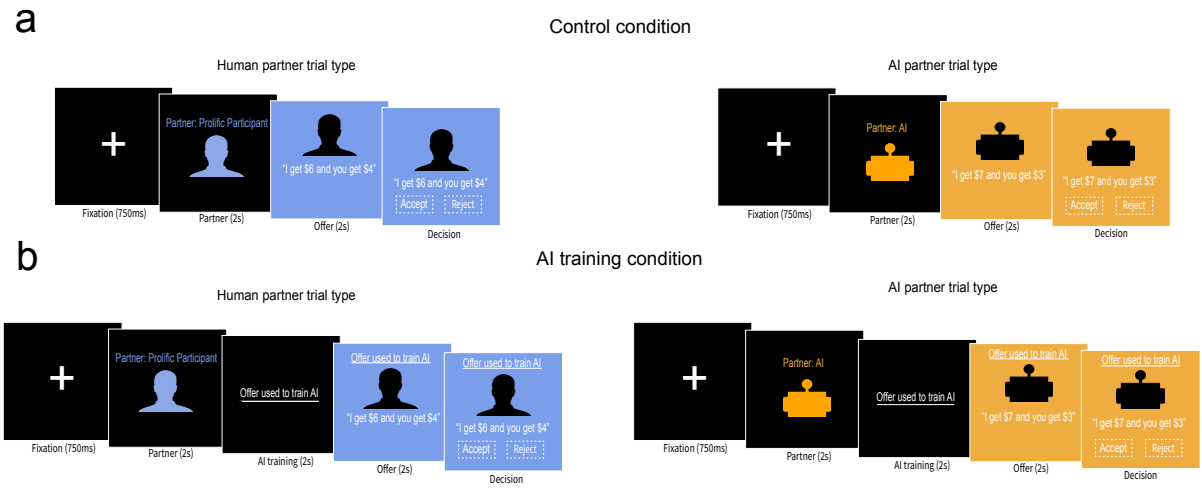


**Figure 2: Example trials for the control (a) and AI training (b) conditions for each partner type (left human participant and right AI). In the AI training condition, participants saw additional text reminding them that their responses were training AI. This text was not shown in the control condition. Aside from this reminder displayed on an additional screen (2s) and when making a choice, the trial format was the same in both conditions. Each trial began with a fixation cross (750ms), followed by the partner type (human or AI) (2s). Participants in the AI training conditions then saw the additional screen reminding them of AI training (2s). They then saw the offer amount for 2 seconds before they could make a choice. Participants had unlimited time to choose. Each participant made multiple choices with varying partner types and offer amounts. Only training condition was varied between participants.**

This study used a $3x2$ design, with partner type (human or AI) as a within-subject factor and training condition as a between-subject factor. There were three training conditions in this experiment: **AI training for self:** Participants were informed they would train an AI responder that they would encounter in a follow-up session. **AI training for others:** Participants were informed they would train an AI responder that other participants, i.e., not themselves, would encounter in a follow-up session. **Control condition:** Participants received no information about AI training. Experiments 2 and 3 closely followed this general set-up, but with some modifications.

## 3.1 Participants

A total of 320 participants (173 female, 6 non-binary; $M$ = 38.70, SD = 12.87) were recruited from Prolific. One participant was removed from the analysis for being exposed to more than one condition since they refreshed the webpage after completing the practice trials. The average completion time was 8 minutes and the median pay rate was approximately $9.50 per hour. In all experiments, participants were paid $8 per hour before receiving a bonus, and provided informed consent before completing each session. The Washington University in St. Louis IRB approved this study.

## 3.2 Design

At the start of the experiment, participants were randomly assigned to the AI training for self ($n$ = 98), AI training for others ($n$ = 100), or control condition ($n$ = 121). They were briefed on the rules of the ultimatum game and told they would play as the responder. Participants were also informed that they would be invited to a follow-up session within the next few weeks, where they would switch roles and play as the proposer.

Participants in the AI training for self condition were told that all their responses would be used to train an AI responder they would later interact with in the follow-up session. Specifically, they were told, *"You will play with the AI that you help train here."* In contrast, participants in the AI training for others condition were informed that they would train an AI responder that other participants—not themselves—would interact with. They were told, *"You will not encounter the AI you train. The AI you help train will only play against other Prolific participants."* Thus, the only difference between the two AI training conditions was whether participants would interact with the AI they trained (further details are provided in Appendix A). In both conditions, participants were not told what the training would entail. The complete set of instructions for all experiments is provided in Appendix B.

Next, participants played multiple rounds of the ultimatum game (Figure 2). In each round, participants chose whether to accept or reject a proposer's offer of how to allocate a $10 sum between both partners. We manipulated partner type within-subject: each participant played against both AI and human partners. To help distinguish between them, each partner type was associated with one of two colors, blue or orange, which were randomly assigned for each participant.

Each round started with a display of a fixation cross ($750ms$). Next, an icon representing the partner type (human participant or AI) was displayed ($2s$). Participants in the AI training condition saw an additional screen with the text "Offer used to train AI responder" ($2s$). This screen served as a reminder that an AI responder would

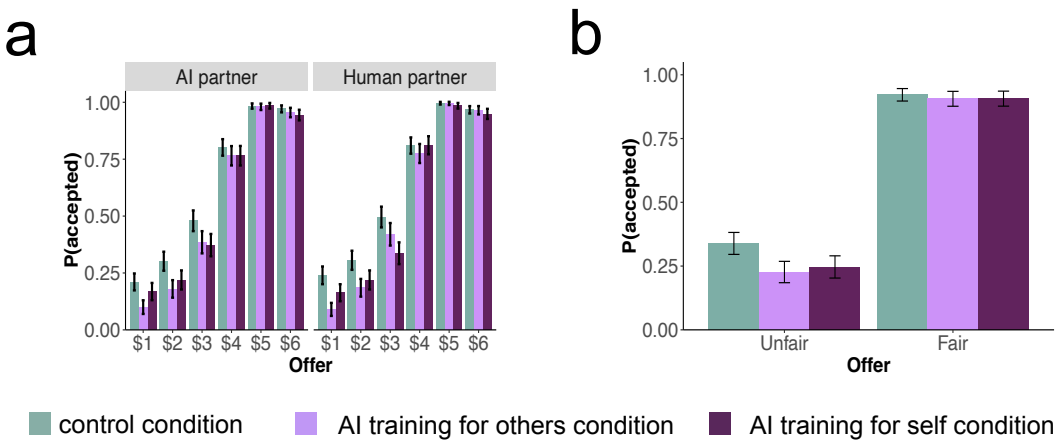


**Figure 3: Results for Experiment 1. Graphs show the proportion of accepting an offer based on (a) offer amount and (b) fairness conditioned on partner type and fairness. Error bars indicate standard error of the mean.**

learn from their responses. Then, participants again saw the partner icon, now accompanied by the offer that was displayed as a line of text indicating the proposed split (e.g., "I get $6 and you get $4). Participants in the AI training conditions also saw the same text as before to remind them of AI training. 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' keys on the keyboard. Participants were given unlimited time to make their decision.

Participants completed 24 rounds of the ultimatum game, playing 12 rounds with each partner type. Offer amounts, ranging from $1 to $6, were presented in a random order. They were balanced across partner types for each participant, ensuring that all participants saw each offer two times from each partner type. For the AI partner trials, we ensured that the offer amounts were the same between conditions. For human partner trials, we recruited enough participants from various studies to ensure that we could balance offers between training conditions using the same amounts. We considered offer amounts $\$1 - \$3$ to be unfair and offer amounts $\$4 - \$6$ to be fair, consistent with prior literature [77].

To incentivize choice behavior, participants were informed that one trial would be randomly selected and resolved at the end of each session. Participants received a bonus of 5% of the amount they earned from the trial. This bonus was increased to 15% for all second sessions to encourage them to return. [1]

After the experiment, participants were asked to describe any strategies they used. Although these responses were not formally analyzed, selected open-ended responses are included in Appendix C. Stimuli, data, and analysis scripts for all experiments can be found on Open Science Framework (OSF) [2].

[1]Since our results suggest that participants mostly relied on intuition when training AI, we expect that they also did not deliberate on how they should make proposals based on their training. For example, they likely did not consider that they should propose fair offers because they trained the AI to punish people for acting unfairly. Therefore, we do not report these results.
[2]Link found here: https://osf.io/2caqy

## 3.3 Analysis

We employed logistic mixed-effects models to assess how partner type, training condition, offer amount, and their interactions predicted participants' acceptance of offers. The models were estimated in R using the lmerTest package. We used this approach for all five experiments.

## 3.4 Results

The results of Experiment 1 are shown in Figure 3. The logistic mixed-effects model revealed that participants accepted more offers as the offer amount increased ($b = 2.07$, SE $= 0.08$, $p < 0.001$), replicating prior findings [77, 91, 111]. However, they responded no differently when partnered with a human compared to an AI partner ($b = -0.06$, SE $= 0.07$, $p = 0.39$).

More importantly, we found that participants in the AI training for others condition rejected more offers than those in the control condition ($b = -0.87$, SE $= 0.40$, $p = 0.03$). The significant interaction effect between offer amount and training condition suggests that participants in the AI training condition for others were more punitive for lower offers than those in the control condition ($b = 0.28$, SE $= 0.13$, $p = 0.026$).

The relationship between the AI training for self condition and control condition was less clear. The logistic mixed-effects model showed neither a main effect of training condition ($b = -0.77$, SE $= 0.40$, $p = 0.056$) nor an interaction effect between offer amount and training condition ($b = -0.04$, SE $= 0.12$, $p = 0.73$). However, visual inspection of Figure 3 suggests that those in the AI training for self condition were more punitive toward unfair offers than those in the control condition. To test this conjecture, we conducted a post-hoc $t$-test comparing the two groups for unfair offers. Although the $t$-test was not significant ($t_{213} = 1.92$, $p = 0.056$), the interaction effect was replicated in all subsequent experiments, suggesting that participants in the AI training for self condition were more punitive toward unfair offers.

We ran another mixed-effects model to assess the relationship between AI training conditions. There was no main effect of training

condition between AI conditions ($b = 0.10,\ \text{SE} = 0.42,\ p = 0.81$). However, there was a significant interaction between offer amount and training condition ($b = -0.32,\ \text{SE} = 0.13,\ p = 0.013$). Specifically, participants in the AI training for others condition were more punitive for lower dollar amounts than those in the AI training for self condition. There were no other significant effects ($ps \geq 0.26$).

### 3.5 Discussion

We found that participants in both AI training conditions rejected more unfair offers than the control condition, consequently training an AI responder to punish people for making unfair proposals. However, participants who trained an AI for themselves rejected less unfair offers than those who trained an AI for others. These findings suggest that people who can directly benefit from AI training engage in more deliberate thinking than those who cannot, but not enough to maximize their rewards. Indeed, participants not only refrained from exploiting AI for personal or others' benefit but also punished themselves and other participants by training AI to be more punitive to their future proposals. This behavior could stem from participants relying on intuition due to their unwillingness to deliberate over their internal model of the AI training process. However, it's also possible that participants had a wrong internal model of the task structure or AI training process. In this case, deliberating over this faulty model would lead to choices (i.e., rejecting unfair offers) that fail to maximize rewards.

## 4 Experiment 2: Does Reducing Mental Costs Encourage Greater Deliberation?

We designed Experiment 2 to assess whether participants would exploit AI if we ensured they understood the task structure and AI training process. The study design was identical to Experiment 1, **but incorporated a comprehension test to evaluate participants' understanding.** In Experiment 2*A*, this test included questions about participants' current and future roles and their partners (AI and human), which they had to answer correctly before proceeding. We reasoned that this test would increase participants' understanding of the task structure and prompt them to consider how the trained AI would respond to their future offers.

We also considered that participants may understand the task structure but not the AI training process, which could prevent them from exploiting AI. Experiment 2*B* addressed this concern by adding two questions to the comprehension test from Experiment 2*A* on how the AI learns to accept or reject offers based on observed responses. Participants had to answer all questions correctly before proceeding with the experiment. We hypothesized that by explicitly making participants understand the task structure and AI training process, they would be more likely to exploit the AI.

### 4.1 Participants

In Experiment 2*A*, 350 participants (196 female, 5 non-binary; $M = 42.31$, SD = 12.63) were recruited from Prolific. Two participants were excluded from the analysis because they refreshed the webpage and were exposed to a different condition. This experiment took 10 minutes to complete, and the median pay rate was approximately $9 per hour.

In Experiment 2*B*, 391 participants (213 female, 3 non-binary; $M = 41.12$, SD = 12.74) were recruited from Prolific. Four participants were excluded for the same reason. This experiment took 11 minutes to complete, with a median pay of approximately $8.50 per hour.

### 4.2 Design

The design of Experiment 2 was identical to Experiment 1 (Figure 2), with the addition of a comprehension test consisting of 5 multiple-choice questions to assess participants' understanding of the task. All participants were tested on the types of partners they would encounter, as well as their current role and their future role in the follow-up session. In Experiment 2*A*, participants in the AI training for others ($n = 100$) and AI training for self ($n = 119$) conditions were required to answer two additional questions. These questions tested participants' knowledge of the type of AI they were training (proposer vs. responder), and whether they would encounter this AI in the follow-up session. Participants in the control condition ($n = 118$) did not engage in AI training, so they did not see these final two questions.

In Experiment 2*B*, participants completed the same task as Experiment 2*A* except for one critical change. Participants assigned to the AI training for others condition ($n = 97$) and the AI training for self condition ($n = 132$) were told how the AI responder would be trained. They were informed that the AI would learn to accept similar offers they accepted and reject similar offers they rejected. Specifically, they were told, *"The AI will learn to respond by copying how you respond to offers. In other words, the AI will learn to accept the offer amounts you accept. Similarly, the AI will learn to reject the offer amounts you reject. Therefore, you can teach the AI which offers it should accept and which it should reject."* To assess participants' understanding, we included two comprehension questions about how the AI would mirror their responses. These questions tested whether participants understood that accepting a $2 offer teaches the AI to accept a $2 offer; and that rejecting a $2 offer teaches the AI to reject a $2 offer. Participants in the control condition ($n = 149$) completed the same task as in Experiment 2A.

All participants had to answer all questions correctly before they could start the experiment. Participants were instructed that if they missed any question, they would be required to re-read the instructions. This motivated participants to learn the task structure (Experiments 2*A* & 2*B*) and AI training process (Experiment 2*B*) to avoid re-reading them. However, participants who needed many attempts may have passed the comprehension test by using a process of elimination strategy. Thus, we removed participants who failed the comprehension test at least 3 times from the analysis. All questions and multiple-choice options can be found in Appendix D.

### 4.3 Results

**Experiment 2A.** In Experiment 2*A*, 71% passed the comprehension test on their first attempt, and 97% passed within three attempts. Eleven participants were removed from the analysis. A detailed analysis of the comprehension test pass rate per question for all experiments can be found in the Appendix E..

The results of Experiment 2*A* were clear: even though they showed an understanding of the task structure, participants in

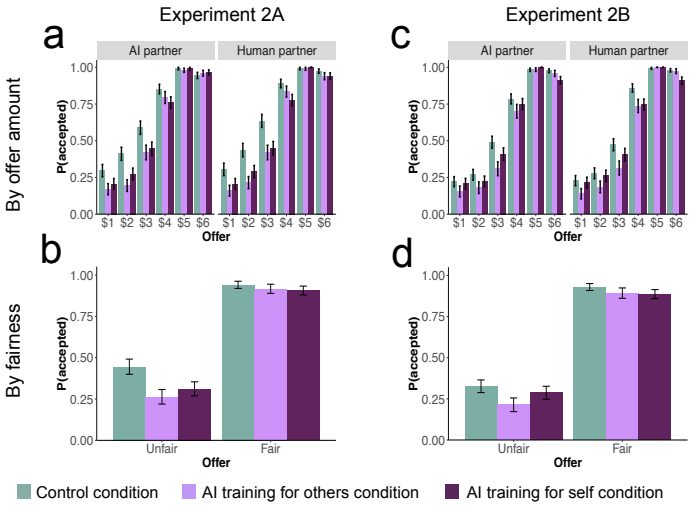


**Figure 4: Results for Experiment 2. All figures show the proportion of acceptance based on offer amount (Top row: a, c) and by fairness (Bottom row: b,d) Error bars indicate standard error of the mean.**

both AI training conditions rejected more unfair offers than those in the control condition (Figures 4a and 4b).

A logistic mixed-effects model revealed a main effect of offer amount ($b = 1.81$, SE = 0.08, $p < 0.001$) and training condition, with participants in the AI training for others ($b = -1.16$, SE = 0.46, $p = 0.012$) and AI training for self ($b = -1.02$, SE = 0.44, $p = 0.02$) conditions rejecting more offers than those in the control condition. Two significant interaction effects between offer amount and training condition qualified these main effects, indicating that participants in the AI training for others ($b = 0.58$, SE = 0.13, $p < 0.001$) and AI training for self ($b = 0.29$, SE = 0.11, $p = 0.009$) conditions were more likely to reject lower offer amounts than those in the control condition.

When comparing AI training conditions, the mixed-effects model revealed no main effect of training condition ($b = 0.14$, SE = 0.45, $p = 0.76$). However, it showed an interaction between training condition and offer amount ($b = -0.29$, SE = 0.13, $p = 0.028$), indicating that participants in the AI training for others condition rejected more offers as the offer amount decreased. These findings replicate the results from Experiment 1.

The mixed-effects model revealed a main effect of partner type ($b = -0.22$, SE = 0.08, $p = 0.005$), with participants more likely to accept offers made by human participants than by AI. Additionally, a significant interaction effect between partner type and training condition was found between the AI training for self and control conditions ($b = 0.21$, SE = 0.11, $p = 0.05$). Post hoc paired $t$-tests showed that participants in the control condition accepted more offers from humans than from AI ($t_{117} = -2$, $p = 0.04$), whereas those in the AI training for self condition showed no difference ($t_{118} = -0.39$, $p = 0.7$). No additional significant interaction effects were observed ($p \geq 0.12$).

**Experiment 2B.** Experiment 2*B* followed the same design as Experiment 2*A*, with the addition of an extra question in the comprehension test. Thus, we ran the same analyses for Experiment 2*B*. First, we found that 78% passed the comprehension test on their first attempt, with 98% passing within three attempts. Ten participants were removed from the analysis for taking too many attempts.

The results of Experiment 2*B* were mostly consistent with those from Experiment 2*A* (Figures 4c and 4d). The logistic mixed-effects model revealed main effects of offer amount ($b = 2.28$, SE = 0.09, $p < 0.001$) and training condition, with participants in both the AI training for others ($b = -1.09$, SE = 0.43, $p = 0.012$) and AI training for self ($b = -0.79$, SE = 0.40, $p = 0.048$) conditions rejecting more offers than those in the control condition.

When comparing the AI training for self and control conditions, we found an interaction between training condition and offer amount ($b = -0.47$, SE = 0.11, $p < 0.001$), with participants in the AI training for self condition rejecting lower offers than those in the control condition.

When comparing the AI training for others and control condition, we found no interaction effect between training condition and offer amount ($b = 0.15$, SE = 0.13, $p = 0.26$). However, inspection of Figure 4c suggests that the main effect of AI training was so strong for offers $\leq$ \$3 that the model could not capture the interaction effect. Thus, we conducted a post-hoc $t$-test comparing the two groups on unfair offers and confirmed this conjecture ($t_{219} = 2.50$, $p = 0.01$).

When comparing between training conditions, we replicated the results from both Experiments 1 and 2*A*. Specifically, a mixed-effects model showed no main effect between training conditions ($b = 0.30$, SE = 0.44, $p = 0.49$), but revealed an interaction effect between training condition and offer amount ($b = -0.62$, SE = 0.12, $p < 0.001$). Hence, participants in the AI training for others condition rejected lower dollar amounts than those in the AI training for self condition.

A mixed-effects model revealed a main effect of partner type ($b = -0.14$, SE = 0.07, $p = 0.042$), showing participants were more likely to accept offers from humans than AI. This finding replicates the result from Experiment 2*A* and is discussed in the General Discussion. There were no other significant interactions ($ps \geq 0.06$).

## 4.4 Discussion

Participants in both AI training conditions continued to reject unfair offers even after demonstrating an understanding of the task structure (Experiments 2*A* & 2*B*) and AI training process (Experiment 2*B*), once again training the AI to punish their future offers. However, participants training an AI for themselves reject unfair offers less often than those training an AI for others. This suggests that participants training an AI for themselves deliberated more, but not enough to train the AI to accept their future, unfair proposals.

However, there is an alternative explanation. Participants may have understood the task yet deliberately chose to train AI to punish unfair behavior [48] rather to maximize rewards. In this case, participants deliberately chose to prioritize fairness over personal gain [108], even to the extent of intentionally sacrificing both current and future rewards to ensure the AI punishes unfair behavior.

## 5 Experiment 3: Do Humans Rely on Intuition When Fairness Concerns Are Irrelevant?

We designed Experiment 3 to test whether participants use goal-directed deliberation to train AI for fairness. The design followed the previous experiments, but with one key difference: **participants were told they would be the only ones to encounter the AI they trained in a follow-up session.** With this change, training the AI to punish unfair behavior became irrelevant, as it would no longer affect others. If participants were engaging in deliberation, they should recognize this and train the AI to maximize their own rewards. In contrast, if participants trained the AI to punish only their future offers, this would suggest they were relying on intuition. To test this, we ran two versions: Experiment 3*A*, which did not include a comprehension test, and Experiment 3*B*, which included a comprehension test to encourage participants to deliberate. We hypothesized that, in the absence of fairness concerns, participants would accept unfair offers, indicating that they engaged in deliberation when training the AI.

### 5.1 Participants

In Experiment 3*A*, 217 participants (111 female, 1 non-binary; $M = 38.09$, $SD = 12.33$) were recruited from Prolific. One participant was excluded from the analysis because they were exposed to both conditions. This experiment took 8 minutes to complete, and the median pay rate was approximately $9.40 per hour.

In Experiment 3*B*, 223 participants (127 female, 1 non-binary; $M = 36.85$, $SD = 11.49$) were recruited from Prolific. Two participants were excluded for refreshing the webpage and being exposed to both conditions. This experiment took 11 minutes to complete, and the median pay rate was approximately $9.10 per hour.

### 5.2 Design

The design was similar to Experiment 1 (Figure 2) except participants were randomly assigned to only two conditions: 'AI training for self' (now referred to as 'AI training') (Experiment 3*A*: $n = 93$) and control (Experiment 3*A*: $n = 123$). Additionally, participants in the AI training condition were informed that only they would encounter the AI they trained in the follow-up session. Specifically, they were told, *"You will be invited to participate in a follow-up experiment where you will make proposals. In this follow-up experiment, you will play with the AI that you train here. You will be the only person to interact with the AI you train. In other words, no other participant will encounter the AI you are training here."* This feature of the experimental design ensured that training the AI for fairness became meaningless, as no other participant would interact with the trained AI. Additional details about how this training condition differs from the training conditions in Experiments 1 and 2 are provided in Appendix A.

For Experiment 3*B*, participants in the AI training condition ($n = 96$) completed the same comprehension test as in Experiment 2*B*, along with an additional question about whether other participants would encounter the AI they trained. This ensured that participants knew that no one else would be affected by their training. Participants in the control condition ($n = 117$) completed the same comprehension test as in Experiments 2. Similar to the previous experiments, participants had to answer all questions correctly before proceeding and those who failed the comprehension test within 3 attempts were removed from the analysis.

### 5.3 Results

The findings of Experiment 3 are clear (Figures 5a and 5b). Participants in the AI training condition forewent their rewards to train the AI to be fair (Experiments 3*A* & 3*B*), even after showing an understanding of the task (Experiment 3*B*).

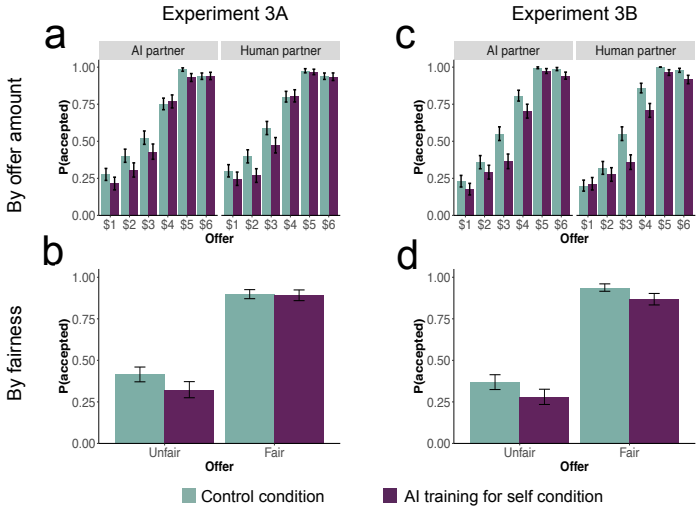


**Figure 5: Results for Experiment 3. All figures show the proportion of acceptance based on offer amount (Top row: a, c) and by fairness (Bottom row: b,d) Error bars indicate standard error of the mean.**

**Experiment 3A.** A logistic mixed-effects model revealed a main effect of offer amount ($b = 1.58$, SE = 0.05, $p < 0.001$). Although there was no main effect of training condition ($b = -0.20$, SE = 0.20, $p = 0.32$), there was an interaction between training condition and offer amount ($b = 0.14$, SE = 0.05, $p = 0.002$), showing that participants in the AI training condition were more punitive for lower offer amounts than those in the control condition.

The mixed-effects model also revealed that participants were more likely to accept offers made by human participants than by AI ($b = -0.12$, SE = 0.05, $p = 0.02$). While this replicates the results of Experiment 2, it differs from the findings of Experiment 1, which we discuss in the General Discussion. No other significant interactions were found ($ps \geq 0.32$).

**Experiment 3B.** In Experiment 3*B*, 74% of participants passed the comprehension test on their first attempt, with 96% passing within three attempts. Eight participants were excluded for taking too many attempts to pass the test.

The results of Experiment 3*B* closely resemble those of Experiment 3*A* (Figures 5c and 5d). A logistic mixed-effects model showed a main effects of offer amount ($b = 1.95$, SE = 0.06, $p < 0.001$) and training condition ($b = -0.61$, SE = 0.21, $p = 0.004$), as well as an interaction between them ($b = -0.23$, SE = 0.06, $p < 0.001$).

Once again, participants in the AI training condition rejected lower offers than those in the control condition.

When considering partner effects, the mixed-effects model did not show a main effect ($b = -0.008$, SE = 0.05, $p = 0.89$). However, a three-way interaction between partner type, offer amount, and training condition was found ($b = 0.09$, SE = 0.04, $p = 0.02$). To interpret this interaction, we ran two additional mixed-effects models conditioned on training condition. These models found no main effect of partner type or interaction between offer amount and partner type ($ps \geq 0.07$), so we do not report these further. No other significant interactions were found ($ps \geq 0.54$).

### 5.4 Discussion

We found that participants training AI continued to prioritize fairness, even though the AI would only punish them for acting unfairly in the follow-up session. This strongly suggests that people rely on intuition when training AI, and that it is difficult for them to overcome this reliance and use more deliberate decision making.

## 6 General Discussion

### 6.1 Recap and Interpretation

In this study, we examined whether people rely on intuition or deliberation when training AI. To do this, participants played the ultimatum game as responders, with some told they were training an AI responder that they would make proposals to in a higher-stakes follow-up session. To maximize rewards, participants training AI needed to accept unfair offers, requiring them to override their intuitive response to punish unfairness [14, 37]. We reasoned that rejecting unfair offers reflects participants relying on intuition, while accepting unfair offers suggests they engaged in deliberation.

Across three experiments, participants training AI rejected more unfair offers than those unaware of AI training, suggesting they relied on an intuitive internal model of how the AI works when making decisions. This reliance on intuition was surprisingly hard to counteract. These findings suggest that people rely on fast and automatic decision making (i.e., intuition) when training AI.

Research on cost-benefit analyses during decision making provides a useful framework for understanding these results. As a reminder, this framework is based on the idea that people attach a cost to mental effort and generally avoid tasks requiring significant mental work [11, 63, 114]. However, this cost can be offset by the perceived benefits [10, 67, 69, 83], leading people to use a cost-benefit analysis to decide whether to use intuition or deliberation [11, 61, 62, 92, 94]. Applying this framework to our case, participants training an AI may have found that the mental effort required to consider the consequences of the trained AI (costs) outweighed the personal gains of exploiting the AI for themselves or the satisfaction of doing so for others (benefits). We also found that participants who could directly benefit from the AI training were less likely to reject unfair offers than those who could not, suggesting that participants who could benefit from AI training were more likely to deliberate. This suggests that these participants placed more value on the potential benefits, but not enough to offset the costs of deliberately training AI. Future research could explore how much reward needs to be increased or how much mental effort must be reduced (e.g., simplifying the task to make it more intuitive to understand) for the benefits to outweigh the costs, so people choose to exploit AI for themselves or for others.

### 6.2 Effect of Partner Type

To reassure participants that AI training was occurring, all experiments included AI partners. This also allowed us to explore the effects of partner types when they were either not emphasized or somewhat emphasized through the comprehension test. When participants were not tested on the types of partners they would encounter (Experiments 1 and 3*A*), results were mixed. In Experiment 1, behavior did not change based on partner type, replicating Treiman et al. [105, 106], while in Experiment 2*A*, participants accepted more offers from humans than AI. These findings suggest a weak sensitivity to partner type, even when not emphasized.

When partner types were explicitly tested (Experiments 2 and 3*B*), we also found mixed results. In Experiment 2, participants accepted more offers from humans, indicating they were more likely to consider how their responses affected others when prompted. However, this effect was absent in Experiment 3*B*, indicating that the effect may be weak.

In Experiments 2 and 3*A*, participants accepted more offers from human partners than from AI, replicating previous findings [104]. However, this contrasts with other research showing that people tend to reject more offers from humans than from AI [18, 77, 91, 111]. A potential explanation lies in the experimental design: both our study and Torta et al. [104], where participants accepted more human offers, were conducted online. In contrast, studies showing a preference for AI offers were conducted in traditional in-person laboratory settings [18, 77, 91, 111]. Another possibility is that the in-person studies often used actors ("confederates") as the partner with which participants would interact, potentially leading to a stricter adherence to social norms (which may be reduced in our study where partners were displayed as abstract silhouettes). Future research may explicitly test this hypothesis by using the exact same paradigm and varying the method of data collection and the presentation of partners.

### 6.3 Fairness Considerations

In our first two experiments, we assumed participants rejected unfair offers because they have an instinct to punish unfair behavior [14, 37]. However, they may have deliberately chosen to reject unfair offers to promote fairness [48]. To test this conjecture, we conducted Experiment 3, where participants trained an AI that only they would encounter. If participants had engaged in deliberate reasoning, they would have recognized that training the AI for fairness had no impact on others and would have exploited it for personal gain. However, participants continued to prioritize fairness, suggesting their behavior was driven by intuition and not deliberate reasoning.

While Experiment 3 was designed to eliminate fairness concerns by ensuring participants had no incentive to prioritize fairness when training AI, we cannot completely rule out the possibility that some still chose to do so. For instance, participants may have deliberately chosen to train AI for fairness because they felt a moral responsibility [100] or wanted to maximize moral payoffs [15, 17]. Nonetheless, it's important to note that there were no broader social

welfare benefits for training AI for fairness in this experiment, as participants' actions would only affect their own outcomes.

There are a myriad of reasons why punishing unfair offers may be an intuitive response. As noted in the Introduction, people may have an inherent drive to punish unfairness [14, 37]. Additionally, they may be motivated by a desire to maintain a positive self-image [9] or conform to social norms [38, 45]. These motivations may vary depending on contextual factors, such as whether their actions are being observed [85]. Future research could explore the motivations behind intuitive punishment of unfairness, such as by manipulating the visibility of observers.

## 6.4 Impact of Stakes in AI Training

We should note that the stakes in this study were relatively low, with participants earning only 5% of the amount from a single negotiation and 15% of the amount in the follow-up session. This setup may have encouraged participants to rush through the task, relying on intuitive strategies. To address this, we included a comprehension test that required participants to re-read the instructions if they missed a question. This made it more efficient for participants to read the instructions carefully, as failing the test would require more time than passing on the first attempt. Thus, the comprehension tests were designed to encourage deliberation, ensuring that participants could not simply rush through the task despite the low incentives. While this design promoted deliberation, we should note that the overall low-stakes setup closely mirrors real-world practices such as crowdsourcing for training data collection [113]. Therefore, in parallel to improving data quality from crowdsourcing [32, 33, 46, 47, 103], it is equally important to understand how people provide training data in similar training conditions.

Additionally, using relatively low stakes better reflects real-world interactions with AI, where the stakes are often minimal in both value and impact. For example, a poor recommendation from AI on a social media platform may only waste a few seconds. While individual low-stakes decisions might seem insignificant, they can accumulate into high-stakes consequences. For instance, overlooking human behavior during AI training could result in substantial social impacts, like the echo chamber effect [21], or make learning infeasible [101]. Therefore, understanding human behavior during AI training is crucial, even when the stakes are relatively low.

## 6.5 Perceptions of AI Algorithms

We informed participants training AI that it would learn to make decisions by mimicking their choices. However, we deliberately kept details about the AI's learning process somewhat vague to mirror real-world applications. Specifically, people are rarely given explicit explanations of how AI systems learn [31, 71] and therefore tend to rely on their own perceptions and assumptions about how the AI operates. While this design choice better reflects real-world AI training scenarios, it is important to consider how people's priors about AI may have shaped their decisions. For instance, although participants correctly identified how the AI would learn from their choices, their assumptions about the strength of that influence likely varied. Some participants may have assumed that AI is resilient to changes, leading them to exaggerate their responses. Alternatively, others may have believed that AI is highly sensitive to training data, prompting them to be more cautious when making responses. Therefore, it is crucial to consider how people's perceptions of AI lead to different behavioral changes. Future research could explore this by assessing people's perceptions about the AI training process.

## 6.6 Limitations

This study only provides behavioral evidence that people mostly rely on intuition when training AI. Cognitive neuroscience has identified several neurophysiological markers that indicate increased mental effort. For example, it is well-established that greater activation in the dorsolateral prefrontal cortex is linked to more controlled, goal-directed thinking and decision making [73, 74, 97]. Future research could use functional magnetic resonance imaging to investigate whether deliberation occurs during AI training. For causal evidence, future research could use transcranial magnetic stimulation on this part of the frontal cortex, which has been shown to reduce goal-directed deliberation [96] and thus may lead to more intuitive AI training. Finally, pupil dilation is a less expensive physiological marker of effort exertion, as pupils dilate when people exert effort [44, 53, 65, 109]. Therefore, this measure may be used to predict trial-by-trial fluctuations in the degree to which people use deliberation to train AI.

Additionally, we show that people rely on intuitive decision-making strategies when training AI in the context of the ultimatum game, a low-stakes scenario where the AI's decisions only affect small monetary outcomes (up to $1.20). Our results may not extend to other contexts where AI training has high-stakes consequences, such as in medical triage [3, 25] or criminal justice [2, 43] applications. In these high-stakes settings, people may shift toward more deliberative decision-making strategies when training AI. Future research could adapt the framework used here to explore whether people continue to rely on intuition or adopt more deliberative decision-making strategies in such high-stakes environments.

# 7 Conclusion

We found that people chose to forgo both current and future rewards to train AI for fairness, indicating a greater reliance on intuition over deliberation. This reliance on intuition was surprisingly difficult to offset, preventing participants from fully exploiting AI to maximize rewards. This tendency to favor intuitive over deliberative decision-making has broader implications for crowdsourcing methods used in AI training. Specifically, as people alter their behavior to train AI, they often embed their biases into the algorithm. Consequently, these behavioral shifts can lead to unreliable and discriminatory outcomes in AI systems, posing a challenge for AI development. To improve AI systems, AI developers should consider the cognitive strategies people use when training AI and how these strategies influence the biasing of the training process. This approach can lead to better AI systems that improve AI assisted decision making.


# 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.

## References


[1] Adrian Acosta-Mitjans, Dagoberto Cruz-Sandoval, Ramon Hervas, Esperanza Johnson, Chris Nugent, and Jesus Favela. 2019. Affective embodied agents and their effect on decision making. In *13th International Conference on Ubiquitous Computing and Ambient Intelligence*. MDPI, 71.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[3] Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, Daniel Mullarkey, Davinder Sangar, Mobasher Butt, Arnold DoRosario, and Saurabh Johri. 2020. A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. *Frontiers in artificial intelligence* 3 (2020), 543405.

[4] Bernard W Balleine and John P O'doherty. 2010. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* 35, 1 (2010), 48–69.

[5] Jaime Banks. 2021. Of like mind: The (mostly) similar mentalizing of robots and humans. *Technology, Mind, and Behavior* (2021).

[6] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[7] Jonathan Baron. 2014. Heuristics and biases. *The Oxford handbook of behavioral economics and the law* (2014), 3–27.

[8] Mohsen Bayati, Mark Braverman, Michael Gillam, Mack, Mark S. Smith, and Eric Horvitz. 2014. Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study. *PLoS ONE* 9, 10 (2014), e109264. https://doi.org/10.1371/journal.pone.0109264

[9] Roland Bénabou and Jean Tirole. 2006. Incentives and prosocial behavior. *American economic review* 96, 5 (2006), 1652–1678.

[10] Matthew Botvinick and Todd Braver. 2015. Motivation and cognitive control: from behavior to neural mechanism. *Annual review of psychology* 66 (2015), 83–113.

[11] Matthew M Botvinick, Stacy Huffstetler, and Joseph T McGuire. 2009. Effort discounting in human nucleus accumbens. *Cognitive, affective, & behavioral neuroscience* 9, 1 (2009), 16–27.

[12] Adriana Camacho and Emily Conover. 2011. Manipulation of social program eligibility. *American Economic Journal: Economic Policy* 3, 2 (2011), 41–65.

[13] Colin F. Camerer. 2003. Strategizing in the Brain. *Science* 300, 5626 (jun 2003), 1673–1675. https://doi.org/10.1126/science.1086215

[14] Alexander W Cappelen, Ulrik H Nielsen, Bertil Tungodden, Jean-Robert Tyran, and Erik Wengström. 2016. Fairness is intuitive. *Experimental Economics* 19 (2016), 727–740.

[15] Valerio Capraro and David G Rand. 2018. Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making* 13, 1 (2018), 99–111.

[16] Sarah H Cen, Andrew Ilyas, Jennifer Allen, Hannah Li, and Aleksander Madry. 2024. Measuring Strategization in Recommendation: Users Adapt Their Behavior to Shape Future Content. *arXiv preprint arXiv:2405.05596* (2024).

[17] Gary Charness and Matthew Rabin. 2002. Understanding social preferences with simple tests. *The quarterly journal of economics* 117, 3 (2002), 817–869.

[18] Mingliang Chen, Zhen Zhao, and Hongxia Lai. 2018. The time course of neural responses to social versus non-social unfairness in the ultimatum game. *Social Neuroscience* 14, 4 (jul 2018), 409–419. https://doi.org/10.1080/17470919.2018.1486736

[19] Yatong Chen, Wei Tang, Chien-Ju Ho, and Yang Liu. 2024. Performative Prediction with Bandit Feedback: Learning through Reparameterization. In *International Conference on Machine Learning*. PMLR, 7298–7324.

[20] Trevor T-J Chong, Matthew Apps, Kathrin Giehl, Annie Sillence, Laura L Grima, and Masud Husain. 2017. Neurocomputational mechanisms underlying subjective valuation of effort costs. *PLoS biology* 15, 2 (2017), e1002598.

[21] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.

[22] Alain Cohn, Tobias Gesche, and Michel André Maréchal. 2022. Honesty in the digital age. *Management Science* 68, 2 (2022), 827–845.

[23] Anne Gabrielle Eva Collins. 2019. Reinforcement learning: bringing together computation and cognition. *Current Opinion in Behavioral Sciences* 29 (2019), 63–68.

[24] Bart Custers. 2013. Data dilemmas in the information society: Introduction and overview. In *Discrimination and privacy in the information society: Data mining and profiling in large databases*. Springer, 3–26.

[25] Adebayo Da'Costa, Jennifer Teke, Joseph E Origbo, Ayokunle Osonuga, Eghosasere Egbon, and David B Olawade. 2025. Ai-driven triage in emergency departments: A review of benefits, challenges, and future directions. *International Journal of Medical Informatics* (2025), 105838.

[26] Maartje M.A. de Graaf and Bertram F. Malle. 2019. People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 239–248. https://doi.org/10.1109/HRI.2019.8673308

[27] Celso M. de Melo and Jonathan Gratch. 2015. People show envy, not guilt, when making decisions with machines. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 315–321. https://doi.org/10.1109/ACII.2015.7344589

[28] Celso M. de Melo, Stacy Marsella, and Jonathan Gratch. 2016. People Do Not Feel Guilty About Exploiting Machines. *ACM Trans. Comput.-Hum. Interact.* 23, 2, Article 8 (may 2016), 17 pages. https://doi.org/10.1145/2890495

[29] C. Di Dio, F. Manzi, S. Itakura, T. Kanda, H. Ishiguro, D. Massaro, and A. Marchetti. 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 (feb 2019), 1045–1059. https://doi.org/10.1007/s12369-019-00528-9

[30] Anthony Dickinson. 1985. Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 308, 1135 (1985), 67–78.

[31] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

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

[33] Xiaoni Duan, Chien-Ju Ho, and Ming Yin. 2022. The influences of task design on crowdsourced judgement: A case study of recidivism risk evaluation. In *Proceedings of the ACM Web Conference 2022*. 1685–1696.

[34] Helen L Gallagher, Anthony I Jack, Andreas Roepstorff, and Christopher D Frith. 2002. Imaging the intentional stance in a competitive game. *Neuroimage* 16, 3 (2002), 814–821.

[35] Thomas Gilovich, Dale Griffin, and Daniel Kahneman. 2002. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.

[36] Chris Giordano, Meghan Brennan, Basma Mohamed, Parisa Rashidi, François Modave, and Patrick Tighe. 2021. Accessing Artificial Intelligence for Clinical Decision-Making. *Frontiers in Digital Health* 3 (2021). https://doi.org/10.3389/fdgth.2021.645232

[37] Sam Goldstein and Robert B Brooks. 2021. Measured Fairness. *Tenacity in Children: Nurturing the Seven Instincts for Lifetime Success* (2021), 111–124.

[38] Russell Golman. 2023. Acceptable discourse: Social norms of beliefs and opinions. *European Economic Review* 160 (2023), 104588.

[39] Charles AE Goodhart and CAE Goodhart. 1984. *Problems of monetary management: the UK experience*. Springer.

[40] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems* 26 (2013).

[41] Werner Güth, Rolf Schmittberger, and Bernd Schwarze. 1982. An Experimental Analysis of Ultimatum Bargaining. *Journal of economic behavior & organization* 3, 4 (1982), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

[42] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. 111–122.

[43] Yugo Hayashi and Kosuke Wakabayashi. 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. https://doi.org/10.1145/3022198.3026338

[44] Eckhard H Hess and James M Polt. 1960. Pupil size as related to interest value of visual stimuli. *Science* 132, 3423 (1960), 349–350.

[45] E Tory Higgins. 1992. Achieving'shared reality'in the communication game: A social action that create; meaning. *Journal of Language and Social Psychology* 11, 3 (1992), 107–131.

[46] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*. 419–429.

[47] Chien-Ju Ho and Jennifer Vaughan. 2012. Online task assignment in crowdsourcing markets. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 26. 45–51.

[48] Guy Hochman, Shahar Ayal, and Dan Ariely. 2015. Fairness requires deliberation: the primacy of economic over social considerations. *Frontiers in psychology* 6 (2015), 747.

[49] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. 2017. Artificial Intelligence in Healthcare: Past, Present and Future. *Stroke and Vascular Neurology* 2, 4 (2017), 230–243. https://doi.org/10.1136/svn-2017-000101

[50] Noel D Johnson and Alexandra A Mislin. 2011. Trust games: A meta-analysis. *Journal of economic psychology* 32, 5 (2011), 865–889.

[51] Daniel Kahneman. 1973. Attention and effort.

[52] Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

[53] Daniel Kahneman and Jackson Beatty. 1966. Pupil diameter and load on memory. *Science* 154, 3756 (1966), 1583–1585.

[54] Ata B. Karagoz, Zachariah M. Reagh, and Wouter Kool. 2023. The Construction and Use of Cognitive Maps in Model-Based Control. *Journal of Experimental*

*Psychology: General* (2023). https://doi.org/10.1037/xge0001491
[55] Jurgis Karpus, Adrian Krüger, Julia Tovar Verba, Bahador Bahrami, and Ophelia Deroy. 2021. Algorithm exploitation: Humans are keen to exploit benevolent AI. *Iscience* 24, 6 (2021).
[56] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' Mental Models of AI: An Item Response Theory Approach. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM. https://doi.org/10.1145/3593013.3594111
[57] Mehdi Keramati, Peter Smittenaar, Raymond J Dolan, and Peter Dayan. 2016. Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences* 113, 45 (2016), 12868–12873.
[58] Sara Kiesler and Jennifer Goetz. 2002. Mental models of robotic assistants. In *CHI'02 extended abstracts on Human Factors in Computing Systems*. 576–577.
[59] Dow-Mu Koh, Nickolas Papanikolaou, Ulrich Bick, Rowland Illing, Charles E. Kahn, Jayshree Kalpathi-Cramer, Matos, Anne Miles, Seong Ki Mun, Napel, Evis Sala, Nicola Strickland, and Fred Prior. 2022. Artificial Intelligence and Machine Learning in Cancer Imaging. *Communications Medicine* 2, 1 (2022). https://doi.org/10.1038/s43856-022-00199-0
[60] Wouter Kool and Matthew Botvinick. 2014. A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General* 143, 1 (2014), 131–141.
[61] Wouter Kool and Matthew Botvinick. 2018. Mental labour. *Nature human behaviour* 2, 12 (2018), 899–908.
[62] Wouter Kool, Samuel J. Gershman, and Fiery A. Cushman. 2017. Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science* 28, 9 (2017), 1321–1333. https://doi.org/10.1177/0956797617708288
[63] Wouter Kool, Joseph T McGuire, Zev B Rosen, and Matthew M Botvinick. 2010. Decision making and the avoidance of cognitive demand. *Journal of experimental psychology: general* 139, 4 (2010), 665.
[64] Sören Krach, Frank Hegel, Britta Wrede, Gerhard Sagerer, Ferdinand Binkofski, and Tilo Kircher. 2008. Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS one* 3, 7 (2008), e2597.
[65] Krzysztof Krejtz, Andrew T Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. 2018. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS one* 13, 9 (2018), e0203629.
[66] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences* 40 (2017), e253.
[67] Lauren A Leotti and Tor D Wager. 2010. Motivational influences on response inhibition measures. *Journal of Experimental Psychology: Human Perception and Performance* 36, 2 (2010), 430.
[68] Alan M Leslie, Ori Friedman, and Tim P German. 2004. Core mechanisms in 'theory of mind'. *Trends in cognitive sciences* 8, 12 (2004), 528–533.
[69] R Libby and Mg Lipe. 1992. Incentives, Effort, And The Cognitive-Processes Involved In Accounting-Related Judgments. *Journal of Accounting Research* 30, 2 (1992), 249–273. https://doi.org/10.
[70] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. 2020. A Review on Interactive Reinforcement Learning From Human Social Feedback. *IEEE Access* 8 (2020), 120757–120765. https://doi.org/10.1109/ACCESS.2020.3006254
[71] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
[72] Kevin McCabe, Daniel Houser, Lee Ryan, Vernon Smith, and Theodore Trouard. 2001. A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the national academy of sciences* 98, 20 (2001), 11832–11835.
[73] Joseph T McGuire and Matthew M Botvinick. 2010. Prefrontal cortex, cognitive control, and the registration of decision costs. *Proceedings of the national academy of sciences* 107, 17 (2010), 7922–7926.
[74] Earl K Miller and Jonathan D Cohen. 2001. An integrative theory of prefrontal cortex function. *Annual review of neuroscience* 24, 1 (2001), 167–202.
[75] George A. Miller, Eugene Galanter, and Karl H. Pribram. 1960. *Plans and the Structure of Behavior*. Henry Holt and Co. https://doi.org/10.1037/10039-000
[76] Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.
[77] Laura Moretti and Giuseppe Di Pellegrino. 2010. Disgust selectively modulates reciprocal fairness in economic interactions. *Emotion* 10, 2 (2010), 169.
[78] Carey K Morewedge and Daniel Kahneman. 2010. Associative processes in intuitive judgment. *Trends in cognitive sciences* 14, 10 (2010), 435–440.
[79] Carey K. Morewedge, Sendhil Mullainathan, Naushan, Jon Kleinberg, Manish Raghavan, and Jens O. Ludwig. 2023. Human Bias in Algorithm Design. *Nature Human Behaviour* 7, 11 (2023), 1822–1824. https://doi.org/10.1038/s41562-023-01724-4
[80] David Navon and Daniel Gopher. 1979. On the economy of the human-processing system. *Psychological review* 86, 3 (1979), 214.
[81] Shuichi Nishio, Kohei Ogawa, Yasuhiro Kanakogi, Shoji Itakura, and Hiroshi Ishiguro. 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. https://doi.org/10.1109/ROMAN.2012.6343851
[82] Hessel Oosterbeek, Randolph Sloof, and Gijs Van De Kuilen. 2004. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental economics* 7 (2004), 171–188.
[83] Srikanth Padmala and Luiz Pessoa. 2011. Reward reduces conflict by enhancing attentional control and biasing visual cortical processing. *Journal of cognitive neuroscience* 23, 11 (2011), 3419–3432.
[84] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.
[85] Jutta Peterburs, Rolf Voegler, Roman Liepelt, Anna Schulze, Saskia Wilhelm, Sebastian Ocklenburg, and Thomas Straube. 2017. Processing of fair and unfair offers in the ultimatum game under social observation. *Scientific reports* 7, 1 (2017), 44062.
[86] Thomas Pouncy, Pedro Tsividis, and Samuel J. Gershman. 2021. What Is the Model in Model-Based Planning? *Cognitive Science* 45, 1 (2021), e12928. https://doi.org/10.1111/cogs.12928
[87] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.
[88] M Richter, GHE Gendolla, and RA Wright. 2016. Three decades of research on motivational intensity theory: What we have learned about effort and what we still don't know. *Advances in motivation science* 3 (2016), 149–186.
[89] James K Rilling, David A Gutman, Thorsten R Zeh, Giuseppe Pagnoni, Gregory S Berns, and Clinton D Kilts. 2002. A neural basis for social cooperation. *Neuron* 35, 2 (2002), 395–405.
[90] Eduardo Benítez Sandoval, Jürgen Brandstetter, Mohammad Obaid, and Christoph Bartneck. 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 (dec 2015), 303–317. https://doi.org/10.1007/s12369-015-0323-x
[91] Alan G. Sanfey, James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom, and Jonathan D. Cohen. 2003. The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science (New York, N.Y.)* 300, 5626 (2003), 1755–1758. https://doi.org/10.1126/science.1082976
[92] Amitai Shenhav, Matthew M Botvinick, and Jonathan D Cohen. 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* 79, 2 (2013), 217–240.
[93] Amitai Shenhav, David G Rand, and Joshua D Greene. 2017. The relationship between intertemporal choice and following the path of least resistance across choices, preferences, and beliefs. *Judgment and Decision making* 12, 1 (2017), 1–18.
[94] Nura Sidarus, Stefano Palminteri, and Valérian Chambon. 2019. Cost-benefit trade-offs in decision-making and learning. *PLoS computational biology* 15, 9 (2019), e1007326.
[95] Steven A Sloman. 1996. The empirical case for two systems of reasoning. *Psychological bulletin* 119, 1 (1996), 3.
[96] Peter Smittenaar, Thomas HB FitzGerald, Vincenzo Romei, Nicholas D Wright, and Raymond J Dolan. 2013. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron* 80, 4 (2013), 914–919.
[97] Alexander Soutschek and Philippe N Tobler. 2020. Causal role of lateral prefrontal cortex in mental effort and fatigue. *Human Brain Mapping* 41, 16 (2020), 4630–4640.
[98] Keith E. Stanovich and Richard F. West. 2000. Advancing the rationality debate. *Behavioral and Brain Sciences* 23, 5 (2000), 701–717. https://doi.org/10.1017/S0140525X00623439
[99] Aleksandra Swiderska, Eva G. Krumhuber, and Arvid Kappas. 2019. Behavioral and Physiological Responses to Computers in the Ultimatum Game. *International Journal of Technology and Human Interaction* 15, 1 (jan 2019), 33–45. https://doi.org/10.4018/ijthi.2019010103
[100] Matthew Talbert. 2016. *Moral responsibility: an introduction*. John Wiley & Sons.
[101] Wei Tang and Chien-Ju Ho. 2019. Bandit Learning with Biased Human Feedback.. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. 1324–1332.
[102] Wei Tang, Chien-Ju Ho, and Yang Liu. 2021. Linear models are robust optimal under strategic behavior. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2584–2592.
[103] Wei Tang, Ming Yin, and Chien-Ju Ho. 2019. Leveraging peer communication to enhance crowdsourcing. In *The World Wide Web Conference*. 1794–1805.
[104] Elena Torta, Elisabeth van Dijk, Peter AM Ruijten, and Raymond H Cuijpers. 2013. 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, 209–217.

[105] Lauren S Treiman, Chien-Ju Ho, and Wouter Kool. 2023. Humans forgo reward to instill fairness into AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. 152–162.
[106] Lauren S Treiman, Chien-Ju Ho, and Wouter Kool. 2024. The consequences of AI training on human decision-making. *Proceedings of the National Academy of Sciences* 121, 33 (2024), e2408731121.
[107] Stephanie Tulk and Eva Wiese. 2018. Trust and Approachability Mediate Social Decision Making in Human-Robot Interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62, 1 (sep 2018), 704–708. https://doi.org/10.1177/1541931218621160
[108] Carmelo Joseph Turillo, Robert Folger, James J Lavelle, Elizabeth E Umphress, and Julie O Gee. 2002. Is virtue its own reward? Self-sacrificial decisions for the sake of fairness. *Organizational behavior and human decision processes* 89, 1 (2002), 839–865.
[109] Pauline van der Wel and Henk Van Steenbergen. 2018. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic bulletin & review* 25 (2018), 2005–2015.
[110] Eric van Dijk and Carsten KW De Dreu. 2021. Experimental games and social decision making. *Annual Review of Psychology* 72 (2021), 415–438.
[111] Mascha van 't Wout, René S. Kahn, Alan G. Sanfey, and André Aleman. 2006. Affective state and decision-making in the Ultimatum Game. *Experimental Brain Research* 169, 4 (feb 2006), 564–568. https://doi.org/10.1007/s00221-006-0346-5
[112] Eliana Vassena, James Deraeve, and William H Alexander. 2020. Surprise, value and control in anterior cingulate cortex during speeded decision-making. *Nature Human Behaviour* 4, 4 (2020), 412–422.
[113] Jennifer Wortman Vaughan. 2018. Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research* 18, 193 (2018), 1–46.
[114] Andrew Westbrook, Daria Kester, and Todd S Braver. 2013. What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PloS one* 8, 7 (2013), e68210.
[115] Aleš Završnik. 2020. Criminal Justice, Artificial Intelligence Systems, and Human Rights. *ERA Forum* 20, 4 (2020), 567–583. https://doi.org/10.1007/s12027-020-00602-0

# A Further Details about AI Training Conditions

## A.1 AI Training Self vs AI Training Others

Because people tend to deliberate more when higher rewards are at stake [10, 67, 69, 83], we tested whether the opportunity to directly benefit from the AI they trained would lead to more deliberate training. To do so, we created two AI training conditions that differed in one key way: **whether participants would interact with the AI they trained in the follow-up session.** One condition trained an AI that they would encounter again in the follow-up session ("AI training for self"). The second condition trained an AI that only other participants (i.e., not themselves) would encounter in the follow-up session. Aside from this distinction, all instructions were identical across the two conditions.

Across both experiments, participants in the AI for self condition were more likely to accept lower offers compared to those in the AI for others condition, suggesting a greater willingness to deliberate when personal rewards were at stake. However, both AI training groups still rejected more low offers than the control group, indicating that neither group fully prioritized reward-maximizing behavior during AI training.

## A.2 New AI Training Condition in Experiment 3

In the AI training conditions in Experiments 1 and 2, participants trained an AI that other participants would encounter. As a result, they may have been motivated to train the AI to punish unfair behavior in order to promote fairness. To remove this incentive, we introduced a new training condition in Experiment 3, where participants trained an AI that only they would encounter in the follow-up session. In other words, **no one else would ever interact with the AI they trained.** In this condition, there is no reason to prioritize fairness since they would only be punishing their own rewards. Therefore, if participants were engaging in deliberate reasoning, they should have trained the AI to maximize their own rewards rather than prioritize fairness.

# B Task Instructions

The following are the exact instructions shown to participants across all experiments. Instructions were presented as a series of individual slides, with each bullet point corresponding to a single slide. The content varied slightly depending on the AI training condition and whether participants completed a comprehension check. All condition-specific variations are noted within the instructions.

At the beginning of the experiment, all participants read instructions about the ultimatum game, the identity of their partner (either another human or AI), and how to respond to offers. In Experiments 2A, 2B, and 3B, participants who were required to complete a comprehension test before proceeding also read an additional line explaining the roles of the proposer and responder, as this information was included in the comprehension test. After reading these instructions, participants completed 2 practice trials.

- Welcome to the experiment! Please read the instructions carefully in order to understand the task. Press **next** or **spacebar** to continue.
- For this experiment, you will play with partners. For each trial, you will play with one partner that could be either another human participant recruited on Prolific or artificial intelligence (AI). The AI is a computer program that has been trained to make choices by observing other human participants do the same task that you are going to complete.
- On each trial, you and your partner will decide how to split $10 amongst yourselves. Your partner has already decided how to divide the $10 and you can choose whether to **accept** or **reject** this proposal.
- In other words, your partner is the **proposer** since they are proposing the offer. Similarly, you are the **responder** since you are responding to their offers.
- If you **accept** the offer, then you and your partner will receive the amount your partner proposed. For example, if your partner proposed to give you $7 and keep $3 for themself, then by accepting this offer you would earn $7 and they would earn $3.
- If you **reject** the offer, then you and your partner will both receive nothing. For example, if your partner proposed to give you $7 and keep $3 for themself, then by rejecting this offer you would both earn $0.
- On each trial, you will see a screen displaying whether your partner is another human participant or the AI. This screen will appear for 2 seconds.
- If your partner is another human participant, you will see this icon on the screen: *human icon.*
- If your partner is the AI, you will see this icon on the screen: *AI icon.*
- You will then see your partner's offer and will choose whether to **accept** or **reject** this offer. To **accept** the offer, press the **F Key**. To **reject** the offer, press the **J Key**.
- You will now complete 2 practice trials. Since this is just a practice trial, there will be no icon displayed. You will not receive any money from these trials. Press **next** or **spacebar** to continue.

## Control Condition

Participants in the control condition were then given a recap of what they just read, details about the follow-up session, and information about potential bonuses. We slightly differed the instructions when a comprehension test was present, as participants completed a comprehension test before proceeding with the experiment.

**Experiments without comprehension test (1 and 3A)**

- After you make all your offers, 1 offer will be randomly selected and resolved. You will earn a bonus of 5% of what you received from that offer.
- Within the next few weeks, you will be invited back to make proposals. For coming back to participate in the follow-up experiment, you will earn a bonus that is 3$X$ as much money than this experiment.
- To recap: 1. You are going to choose between accepting and rejecting offers. 2. Sometimes, human participants made these offers, other times an AI made them. 3. Within the next few weeks, you will be invited to participate in a follow-up experiment where you will make proposals.

- You are now ready for the experiment. You will complete 24 trials. Press space or click the next button to begin!

**Experiments with comprehension test (2A, 2B, and 3B)**

- After you make all your offers, 1 offer will be randomly selected and resolved. You will earn a bonus of 5% of what you received from that offer.
- Within the next few weeks, **you will be invited back to play as the proposer.** This means that you will make the proposals in the follow-up experiment.
- In other words, you are switching roles from what you are doing here. We will recruit participants to respond to your proposals you make in the follow-up experiment.
- For coming back to participate in the follow-up experiment, you will earn a bonus that is $3X$ as much money than this experiment.
- To recap: 1. You are going to choose between accepting and rejecting offers. 2. Sometimes, human participants made these offers, other times an AI made them. 3. Within the next few weeks, you will be invited to participate in a follow-up experiment where you will make proposals.
- Before you complete the experiment, you will be asked a few questions about the rules of the experiment. You must get all questions right before you can proceed. **If you give the wrong answer to any question, then you will be required to read the instructions again.**
- You are now ready for the experiment. You will complete 24 trials. Press space or click the next button to begin!

## AI Training Conditions

Participants in the AI training condition were informed about potential bonuses and that they would be invited to a follow-up session where they would play as the proposer. Instructions varied depending on whether they completed a comprehension test and how their responses would be used for AI training.

**Experiment 1.** Participants were instructed that they were training an AI that either they (AI training for self) or other participants (AI training for others) would make proposals to in a follow-up session. Because participants' responses in the AI training for self condition were used differently than in the AI training for others condition, the instructions varied slightly between the two conditions.

- Before you start the experiment, we need to explain one more aspect.
- Your responses will be used to train an AI to respond to offers, just like you are doing here. This AI will learn by observing your responses.
- You will be invited to participate in a follow-up experiment where you will make proposals. In this follow-up experiment,
  - *(AI training for self)* **you will play with the AI that you help train here.**
  - *(AI training for others)* **you will not encounter the AI you train. The AI you help train will only play against other Prolific participants.**
- To remind you that an AI is observing your choices, you will see a screen with the following text before each offer: **Offer used to train AI responder**
- After you make all your offers, 1 offer will be randomly selected and resolved. You will earn a bonus of 5% of what you received from that offer.
- Within the next few weeks, you will be invited back to make proposals. For coming back to participate in the follow-up experiment, you will earn a bonus that is $3X$ as much money than this experiment.
  - *(AI training for self)* **In this follow-up experiment, you will play with the AI that is trained using your responses in this experiment.**
  - *(AI training for others)* **The AI you train will play against other Prolific participants. You will not encounter the AI you train in the follow-up experiment.**

  .
- To recap: 1. You are going to choose between accepting and rejecting offers. 2. Sometimes, human participants made these offers, other times an AI made them. 3. Your responses will be used to train an AI Responder that...
  - *(AI training for self)* you will encounter in a follow-up session.
  - *(AI training for others)* will play with other Prolific users in future experiments.

  4. Within the next few weeks, you will be invited to participate in a follow-up experiment where you will make proposals.
- You are now ready for the experiment. You will complete 24 trials. Press space or click the next button to begin!

**Experiments 2A and 2B.** Participants in Experiments 2A and 2B completed the same task as those in Experiment 1, but they also completed a comprehension test. As a result, the instructions differed slightly from those in Experiment 1. The only difference between Experiments 2A and 2B was that participants in 2B also received a message explaining how the AI would learn to respond to offers. Aside from this addition, all other instructions were identical.

- Before you start the experiment, we need to explain one more aspect.
- Your responses will be used to train an AI to respond to offers, just like you are doing here. The AI will learn by mimicking how you respond to offers.
- *(For participants in Experiment 2B)*: **Important message: Please read carefully!** The AI will learn to respond by copying how you respond to offers. In other words, the AI will learn to **accept** the offer amounts you **accept**. Similarly, the AI will learn to **reject** the offer amounts you **reject**. Therefore, you can teach the AI which offers it should accept and which offers it should reject.
- To remind you that an AI is observing your choices, you will see a screen with the following text before each offer: **Offer used to train AI responder**
- After you make all your offers, 1 offer will be randomly selected and resolved. You will earn a bonus of 5% of what you received from that offer.
- Within the next few weeks, **you will be invited back to play as the proposer**. This means that you will make the proposals in the follow-up experiment. In other words, you are switching roles from what you are doing here. We will

recruit participants to respond to your proposals you make in the follow-up experiment.
- For coming back to participate in the follow-up experiment, you will earn a bonus that is 3$X$ as much money than this experiment.
  - *(AI training for self)* **In this follow-up experiment, you will play with the AI that is trained using your responses in this experiment.**
  - *(AI training for others)* **The AI you train will play against other Prolific participants. You will not encounter the AI you train in the follow-up experiment.**
- To recap: 1. You are going to choose between accepting and rejecting offers. 2. Sometimes, human participants made these offers, other times an AI made them. 3. Your responses will be used to train an AI Responder that...
  - *(AI training for self)* you will encounter in a follow-up session.
  - *(AI training for others)* will play with other Prolific users in future experiments.

  4. Within the next few weeks, you will be invited to participate in a follow-up experiment where you will switch roles and play as the proposer.
- Before you complete the experiment, you will be asked a few questions about the rules of the experiment. You must get all questions right before you can proceed.**If you give the wrong answer to any question, then you will be required to read the instructions again.**
- You are now ready for the experiment. You will complete 24 trials. Press space or click the next button to begin!

**Experiments 3A and 3B**. Participants in Experiment 3A completed the same task as those in Experiment 1, while participants in Experiment 3B completed the same task as those in Experiment 2B. The only difference was that participants training the AI were explicitly told that only they (i.e., no one else) would interact with the AI they trained.

In Experiment 3A, the instructions were identical to those in Experiment 1, except that the following sentence:

"You will be invited to participate in a follow-up experiment where you will make proposals. In this follow-up experiment,

- *(AI training for self)* **you will play with the AI that you help train here.**
- *(AI training for others)* **you will not encounter the AI you train. The AI you help train will only play against other Prolific participants.**"

was replaced with:

"You will be invited to participate in a follow-up experiment where you will make proposals. In this follow-up experiment, **you will play with the AI that you train here. You will be the only person to interact with the AI you train.** In other words, no other participant will encounter the AI you are training here."

Similarly, the instructions in Experiment 3B were identical to those in Experiment 2B, except for the inclusion of the following additional messages:

- You will be invited to participate in a follow-up experiment where you will make proposals. In this follow-up experiment, **you will play with the AI that you train here.**
- **You will be the only person to interact with the AI you train.** In other words, no other participant will encounter the AI you are training here.

In addition, we revised the original recap message:

"To recap: 1. You are going to choose between accepting and rejecting offers. 2. Sometimes, human participants made these offers, other times an AI made them. 3. Your responses will be used to train an AI Responder that...

- *(AI training for self)* you will encounter in a follow-up session.
- *(AI training for others)* will play with other Prolific users in future experiments.

4. Within the next few weeks, you will be invited to participate in a follow-up experiment where you will switch roles and play as the proposer."

with the following revised version:

"To recap: 1. You are going to choose between accepting and rejecting offers. 2. Sometimes, human participants made these offers, other times an AI made them. 3. Your responses will be used to train an AI responder that you will encounter in a follow-up experiment. 4. The AI responder you train will learn to respond to offers by mimicking your responses. 5. Within the next few weeks, you will be invited to participate in this follow-up experiment where you will switch roles and play as the proposer."

## C Open-Ended Responses

At the end of the experiment, participants were asked to describe any strategies they used while completing the task. Below, we present all responses specifically related to AI training, organized by experiment. These are direct quotes and have not been edited for grammar or spelling.

### C.1 Experiment 1

- if AI is truly being trained based on my responses I though it made sense to train it to always "accept" since the roles will be reversed in a future study. Also there is no real reward for "rejecting" the offer. We both receive $0 when I reject. There were times I felt "petty" enough to want to reject the $1 offer but then I would receive nothing. I would be punishing both myself and my somewhat greedy opponent.
- I wanted offers from a human to be fair, didn't want to reward them for being selfish. I didn't care about this so much with an AI because they can't be selfish but still found myself resistant to extremely unfair offers.
- I thought about training the AI for the future, but then I decided to maximize my current bonus so I accepted all offers.

- Yes, I would have rejected any offers that gave my opponent the full amount and left me with nothing - training the AI to take lower offers was in my best interest for future trials so I accepted all non-zero splits
- if they were either split 50/50 or 60/40 then I accepted, since it was training the AI and I wanted it to be more fair in the future.
- I wanted to maximize my own bonus, but I just felt that keeping $9 and only offering $1 was too greedy, so that was my threshold for choosing to reject. I wanted to teach the AI to punish that level of greed as well.

## C.2 Experiment 2A

- Yes, I wanted to teach the AI to give me more money, but accepted regardless of what the person offered.
- Eventually I decided it mattered little if I accept all the AI since no one is facing any consequence. I thought I perhaps should have accepted all for the AI in the future. Oh well.
- I didn't want to accept anything below $3 as it seems very one sided. While I lost out it hopefully was able to teach the AI something about being fair and sharing.
- NO STRATEGIES JUST ALL BASED ON INSTINCT
- I wanted to teach AI to take offers that were better for me
- At first, I was focused on fairness due to the human aspect, so I was aiming for accepting anything between $4-6. Mid-way through I thought about how I was training the AI Robot for the next round and what would potentially be more beneficial to me. At the end, with the final offer of a random trial selection, I realized I probably should've gone with accepting every offer to guarantee "winning" that final offer. Oops. I should've thought through it more before starting.
- I want to accept even low offers so that I can get the AI to accept my offers when I get the follow up
- I rather accept any offer than get nothing and also training the AI to accept every offer.
- I rejected really unfair offers as I stood to gain almost nothing and I wanted to teach the AI observer
- Yes I generally was playing to train the AI to accept lower offers for the next part of the experiment.

## C.3 Experiment 2B

- I would rather have some money than no money, so I accepted pretty much everything. If they were a d*** and tried to keep all $10, I rejected so we both got nothing (: I thought a little about the AI being trained, or I might have decided to reject $9/$1 as well.
- Yes. At first, I thought about accept EVERY offer just so I could train the AI to always do the same thing. Then when I do the follow up study I would be the proposer and I could make horrible offers like I get $9 and the AI gets $1 and it would always accept it. It's not a person so I don't really feel bad about not being fair... Then at some point I realized I should be trying to maximize my bonus for THIS study and thought that maybe I actually should reject those $1 offers so I did that a few times. Then I realized that I would probably have a chance to make more money off the follow up study so I started acceptin gthose low offers anyway. Basically I am trying to maximize earnings for the next study instead of this one
- Part of me wanted to try to be "fair" and only select offers that leaned towards a 50/50 split. Other times, I was tempted to accept all offers so I could train the AI to accept all offers, including bad ones for AI and good ones for me in the future.
- I just refused to accept anything less then $4 I thought everything else was unfair. I know your trying to get people to accept the poor offers and then in 2 weeks make those offers since AI should accept them but I refuse to do that. I will be fair and offer a 50/50 split when I get the chance to decide the offers.
- I wanted to train the AI to say yes to everything so that in the future trial, I could get more. So I a greed to things I wouldn't agree with if I was actually working with/against other humans
- I was looking at the them vs me amount - training AI not to make low offers
- Since I was "training" I thought I should be only accepting "good" offers so I only accepted when we got equal or I got more. However, I wish I had gone with my initial thought and accepted ALL offers because anything would have been better than nothing.
- I didn't see any reason to not accept an offer as long as it didn't offer me $0. It is beneficial for my bonus and to train the AI for the next round when it is the responder.

## C.4 Experiment 3A

- I wanted at least 40% of the total. I didn't accept lower offers because I hoped this would train the future AI to make offers of 40% or more to me which would also maximize it's earnings.
- accept most offers to train the Ai to be more lenient for the next time

## C.5 Experiment 3B

- I wanted to train the AI to accept the lower offer so that next week I get higher as the proposer
- Accepting an offer would always result in me earning money, while rejecting one would ensure I didn't. Another advantage to accepting all offers was that when I encounter this AI in the follow-up, I can give it "unfair" offers and it will be trained to accept them all.
- I wanted to accept all offers because I am training the AI to accept all offers
- I can see no rational reason whatsoever to reject any offer as long as all of those offers offer me any money and every offer is being used to train an AI that that will accept or reject my future offers and that I get a greater bonus for what I offer in the future. I could only see myself rejecting if I had received no money OR if some of the trial runs were not being used to train AI. But every trial run(including those from the human participant) here indicated it was being used to train the AI.
- Train the AI to want nothing

- I was preparing the AI for the next phase in order to capitalize my earnings
- Since the intent is to train an AI, I only accepted offers in which I received 50% or more of the total.
- Since I get at least $1 when I accept and get $0 when I reject, it would be best to accept every offer. Since it is used to train the AI that I will be proposing to, if I accept the they get $9 and I get $1, when I propose, I could just propose all I get $9 and they get $1 and the AI will accept all of my offers, gaining me much more money in the future.
- I accepted all offers so that the AI will learn to accept whatever offer I send it in the follow up trial.

## D Comprehension Test Questions

The comprehension test consisted of 8 multiple-choice questions. Participants in the control condition answered only questions $1-3$. Participants in the AI training condition answered questions $1-3$ as well as a subset of questions $4-8$, depending on the experiment. Specifically, in Experiment 2*A*, participants answered the first 5 questions; in Experiment 2*B*, participants answered all questions except question 6; and in Experiment 3*B*, participants answered all 8 questions.

Participants were required to answer all questions correctly before proceeding with the experiment. If participants missed any questions, they had to reread the instructions. The responses for each multiple-choice question were randomly shuffled each time participants took the quiz. The correct answers to each question are in bold, except for question 5, as the correct answer depends on the AI training condition. In this case, the correct answer for each AI training condition is italicized.

**Questions all participants (control and AI training conditions) completed across all experiments.**

(1) In *this task*, will you be proposing or responding to offers?
   (a) I will be proposing offers.
   (b) **I will be responding to offers.**
(2) In the *follow-up task*, will you be proposing or responding to offers?
   (a) **I will be proposing offers.**
   (b) I will be responding to offers.
(3) Who will you be playing with in this task?
   (a) I will only play with human participants recruited from a separate Prolific experiment.
   (b) I will only play with AI.
   (c) **I will play with both AI and human participants.**

**Questions completed only by participants in the AI training condition (listed by experiment in parentheses).**

(4) What type of AI will you be training? (Exps. 2*A*, 2*B*, 3*B*)
   (a) I will be training an AI to propose offers.
   (b) **I will be training an AI to respond to offers.**
(5) Will you encounter the AI you are training in the follow-up session? (Exps. 2*A*, 2*B*, 3*B*)
   (a) Yes *(Answer for AI training for self condition)*
   (b) No *(Answer for AI training for others condition)*
(6) Will other participants encounter the AI you are training in the follow-up session? (Exp. 3*B*)
   (a) Yes
   (b) **No**
(7) If you *accept* a $6 offer, what will the AI responder learn?* (Exps. 2*B*, 3*B*)
   (a) **The AI responder will learn to accept $6**
   (b) The AI responder will learn to reject $6
   (c) The AI responder will not learn anything.
(8) If you *reject* a $4 offer, what will the AI responder learn?* (Exps. 2*B*, 3*B*)
   (a) The AI responder will learn to accept $4
   (b) **The AI responder will learn to reject $4**
   (c) The AI responder will not learn anything.

*Offer amounts were randomly selected between $1 – $6 for each participant each time they took the test.

# E Comprehension Test Results for Experiments 2 and 3B

Tables 1 to 6 summarize comprehension check performance for Experiments 2A, 2B, and 3B. For each experiment, we report the number of participants who passed on each attempt and the cumulative percentage who passed (Tables 1, 3, and 5) and the number of incorrect responses per question across attempts (Tables 2, 4, and 6).

## E.1 Experiment 2A

**Table 1: Participant Performance Across Attempts.**

| Attempt | Passed | Total Passed | Total Passed (%) |
|---|---|---|---|
| 1 | 248 | 248 | 71.26 |
| 2 | 68 | 316 | 90.80 |
| 3 | 21 | 337 | 96.84 |
| 4 | 5 | 342 | 98.28 |
| 5 | 3 | 345 | 99.14 |
| 6 | 2 | 347 | 99.71 |
| 7 | 1 | 348 | 100.00 |

**Table 2: Incorrect Answers per Question Across Attempts.**

| Attempt | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| 1 | 5 | 30 | 12 | 54 | 27 |
| 2 | 1 | 8 | 12 | 11 | 5 |
| 3 | 1 | 3 | 5 | 1 | 1 |
| 4 | 1 | 3 | 1 | 4 | 0 |
| 5 | 0 | 1 | 2 | 0 | 1 |
| 6 | 0 | 0 | 1 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 |

## E.2 Experiment 2B

**Table 3: Participant Performance Across Attempts.**

| Attempt | Passed | Total Passed | Total Passed (%) |
|---|---|---|---|
| 1 | 303 | 303 | 78.29 |
| 2 | 64 | 367 | 94.83 |
| 3 | 11 | 378 | 97.67 |
| 4 | 4 | 382 | 98.71 |
| 5 | 3 | 385 | 99.48 |
| 6 | 1 | 386 | 99.74 |
| 7 | 1 | 387 | 100.00 |

**Table 4: Incorrect Answers per Question Across Attempts.**

| Attempt | Q1 | Q2 | Q3 | Q4 | Q5 | Q7 | Q8 |
|---|---|---|---|---|---|---|---|
| 1 | 3 | 31 | 18 | 25 | 24 | 6 | 5 |
| 2 | 1 | 5 | 5 | 3 | 6 | 1 | 2 |
| 3 | 1 | 3 | 4 | 2 | 1 | 0 | 1 |
| 4 | 1 | 2 | 3 | 0 | 1 | 0 | 0 |
| 5 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## E.3 Experiment 3B

**Table 5: Participant Performance Across Attempts.**

| Attempt | Passed | Total Passed | Total Passed (%) |
|---|---|---|---|
| 1 | 165 | 165 | 74.32 |
| 2 | 41 | 206 | 92.79 |
| 3 | 8 | 214 | 96.40 |
| 4 | 4 | 218 | 98.20 |
| 5 | 1 | 219 | 98.65 |
| 6 | 1 | 220 | 99.10 |
| 7 | 0 | 220 | 99.10 |
| 8 | 2 | 222 | 100.00 |

**Table 6: Incorrect Answers per Question Across Attempts.**

| Attempt | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 19 | 13 | 7 | 12 | 21 | 1 | 4 |
| 2 | 0 | 5 | 8 | 4 | 2 | 4 | 0 | 2 |
| 3 | 1 | 4 | 3 | 0 | 1 | 2 | 1 | 1 |
| 4 | 1 | 1 | 3 | 0 | 0 | 2 | 0 | 0 |
| 5 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| 6 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 7 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

# F Mixed-Effects Regression Results

## F.1 Mixed Effects Model Results for Experiment 1

**Table 7: Reference Level: control condition**

| Predictor | Estimate | Std. Error | $z$-value | $Pr(>\|z\|)$ | Significance |
|---|---|---|---|---|---|
| (Intercept) | 1.8689 | 0.2747 | 6.80 | $< 0.001$ | *** |
| AI Opponent | −0.0643 | 0.0745 | −0.86 | 0.388 | |
| Offer | 2.0743 | 0.0848 | 24.45 | $< 0.001$ | *** |
| AI training for others condition | −0.8728 | 0.4032 | −2.17 | 0.030 | * |
| AI training for self condition | −0.7740 | 0.4048 | −1.91 | 0.056 | . |
| AI Opponent:Offer | 0.0335 | 0.0547 | 0.61 | 0.540 | |
| AI Opponent:AI training for others condition | −0.0340 | 0.1099 | −0.31 | 0.757 | |
| AI Opponent:AI training for self condition | 0.0414 | 0.1069 | 0.39 | 0.699 | |
| Offer:AI training for others condition | 0.2824 | 0.1265 | 2.23 | 0.026 | * |
| Offer:AI training for self condition | −0.0399 | 0.1168 | −0.34 | 0.733 | |
| AI Opponent:Offer:AI training for others condition | −0.0969 | 0.0860 | −1.13 | 0.260 | |
| AI Opponent:Offer:AI training for self condition | −0.0843 | 0.0818 | −1.03 | 0.303 | |

*Signif. codes*: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

**Table 8: Reference level: AI training for others condition**

| Predictor | Estimate | Std. Error | $z$-value | $Pr(>\|z\|)$ | Significance |
|---|---|---|---|---|---|
| (Intercept) | 0.9961 | 0.2963 | 3.36 | 0.00078 | *** |
| AI Opponent | −0.0983 | 0.0808 | −1.22 | 0.22350 | |
| Offer | 2.3568 | 0.1013 | 23.27 | $< 0.001$ | *** |
| Control condition | 0.8728 | 0.4032 | 2.16 | 0.03041 | * |
| AI training for self condition | 0.0988 | 0.4204 | 0.24 | 0.81415 | |
| AI Opponent:Offer | −0.0634 | 0.0664 | −0.96 | 0.33919 | |
| AI Opponent:Control condition | 0.0340 | 0.1099 | 0.31 | 0.75678 | |
| AI Opponent:AI training for self condition | 0.0754 | 0.1114 | 0.68 | 0.49829 | |
| Offer: Control condition | −0.2825 | 0.1265 | −2.23 | 0.02556 | * |
| Offer:AI training for self condition | −0.3223 | 0.1294 | −2.49 | 0.01273 | * |
| AI Opponent:Offer: Control condition | 0.0969 | 0.0860 | 1.13 | 0.25982 | |
| AI Opponent:Offer:AI training for self condition | 0.0126 | 0.0899 | 0.14 | 0.88820 | |

*Signif. codes*: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## F.2 Mixed Effects Model Results for Experiment 2A

**Table 9: Reference Level: Control condition**

| Predictor | Estimate | Std. Error | $z$-value | $Pr(>\|z\|)$ | Significance |
|---|---|---|---|---|---|
| (Intercept) | 2.6291 | 0.3190 | 8.24 | $< 0.001$ | *** |
| AI Opponent | −0.2217 | 0.0794 | −2.79 | 0.0053 | ** |
| Offer | 1.8113 | 0.0769 | 23.55 | $< 0.001$ | *** |
| AI training for others condition | −1.1583 | 0.4603 | −2.52 | 0.0119 | * |
| AI training for self condition | −1.0219 | 0.4402 | −2.32 | 0.0203 | * |
| AI Opponent:Offer | −0.0754 | 0.0546 | −1.38 | 0.1675 | |
| AI Opponent:AI training for others condition | 0.1709 | 0.1158 | 1.48 | 0.1398 | |
| AI Opponent:AI training for self condition | 0.2121 | 0.1081 | 1.96 | 0.0498 | * |
| Offer:AI training for others condition | 0.5830 | 0.1286 | 4.53 | $< 0.001$ | *** |
| Offer:AI training for self condition | 0.2918 | 0.1117 | 2.61 | 0.0090 | ** |
| AI Opponent:Offer:AI training for others condition | 0.0828 | 0.0884 | 0.94 | 0.3493 | |
| AI Opponent:Offer:AI training for self condition | 0.1191 | 0.0772 | 1.54 | 0.1229 | |

*Signif. codes*: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

**Table 10: Reference Level: AI training for others condition**

| Predictor | Estimate | Std. Error | $z$-value | $Pr(>\|z\|)$ | Significance |
|---|---|---|---|---|---|
| (Intercept) | 1.47139 | 0.33622 | 4.38 | $< 0.001$ | *** |
| AI Opponent | −0.05075 | 0.08424 | −0.60 | 0.547 | |
| Offer | 2.39501 | 0.10821 | 22.13 | $< 0.001$ | *** |
| Control condition | 1.16538 | 0.46110 | 2.53 | 0.011 | * |
| AI training for self condition | 0.13645 | 0.45412 | 0.30 | 0.764 | |
| AI Opponent:Offer | 0.00738 | 0.06956 | 0.11 | 0.916 | |
| AI Opponent:Control condition | −0.17198 | 0.11590 | −1.48 | 0.138 | |
| AI Opponent:AI training for self condition | 0.04120 | 0.11175 | 0.37 | 0.712 | |
| Offer:Control condition | −0.57580 | 0.12887 | −4.47 | $< 0.001$ | *** |
| Offer:AI training for self condition | −0.29134 | 0.13236 | −2.20 | 0.028 | * |
| AI Opponent:Offer:Control condition | −0.08334 | 0.08855 | −0.94 | 0.347 | |
| AI Opponent:Offer:AI training for self condition | 0.03629 | 0.08837 | 0.41 | 0.681 | |

*Signif. codes*: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## F.3 Mixed Effects Model Results for Experiment 2B

**Table 11: Reference Level: Control condition**

| Predictor | Estimate | Std. Error | $z$-value | $Pr(>\|z\|)$ | Significance |
|---|---|---|---|---|---|
| (Intercept) | 2.0196 | 0.2768 | 7.30 | $< 0.001$ | *** |
| AI Opponent | −0.1437 | 0.0706 | −2.03 | 0.042 | * |
| Offer | 2.2835 | 0.0860 | 26.54 | $< 0.001$ | *** |
| AI training for others condition | −1.0898 | 0.4339 | −2.51 | 0.012 | * |
| AI training for self condition | −0.7868 | 0.3977 | −1.98 | 0.048 | * |
| AI Opponent:Offer | −0.0667 | 0.0533 | −1.25 | 0.211 | |
| AI Opponent:AI training for others condition | 0.0512 | 0.1086 | 0.47 | 0.637 | |
| AI Opponent:AI training for self condition | 0.1166 | 0.0950 | 1.23 | 0.220 | |
| Offer:AI training for others condition | 0.1505 | 0.1330 | 1.13 | 0.258 | |
| Offer:AI training for self condition | −0.4672 | 0.1055 | −4.43 | $< 0.001$ | *** |
| AI Opponent:Offer:AI training for others condition | −0.0444 | 0.0876 | −0.51 | 0.612 | |
| AI Opponent:Offer:AI training for self condition | 0.1149 | 0.0712 | 1.61 | 0.107 | . |

**Table 12: Reference Level: AI training for others condition**

| Predictor | Estimate | Std. Error | $z$-value | $Pr(>\|z\|)$ | Significance |
|---|---|---|---|---|---|
| (Intercept) | 0.9298 | 0.3357 | 2.77 | 0.0056 | ** |
| AI Opponent | −0.0925 | 0.0825 | −1.12 | 0.2617 | |
| Offer | 2.4340 | 0.1089 | 22.35 | $< 0.001$ | *** |
| Control condition | 1.0898 | 0.4341 | 2.51 | 0.0121 | * |
| AI training for self condition | 0.3031 | 0.4413 | 0.69 | 0.4922 | |
| AI Opponent:Offer | −0.1111 | 0.0696 | −1.60 | 0.1102 | |
| AI Opponent:Control condition | −0.0512 | 0.1086 | −0.47 | 0.6373 | |
| AI Opponent:AI training for self condition | 0.0654 | 0.1041 | 0.63 | 0.5300 | |
| Offer:Control condition | −0.1505 | 0.1330 | −1.13 | 0.2578 | |
| Offer:AI training for self condition | −0.6177 | 0.1248 | −4.95 | $< 0.001$ | *** |
| AI Opponent:Offer:Control condition | 0.0444 | 0.0876 | 0.51 | 0.6123 | |
| AI Opponent:Offer:AI training for self condition | 0.1593 | 0.0841 | 1.89 | 0.0582 | . |

## F.4 Mixed Effects Model Results for Experiment 3

**Table 13: Mixed Effects Model Results for Experiment 3A**

| Predictor | Estimate | Std. Error | $z$-value | $Pr(>\|z\|)$ | Significance |
|---|---|---|---|---|---|
| (Intercept) | 1.60523 | 0.20173 | 7.96 | $< 0.001$ | *** |
| AI Opponent | −0.11605 | 0.04985 | −2.33 | 0.0199 | * |
| Offer | 1.57531 | 0.04939 | 31.89 | $< 0.001$ | *** |
| AI training | −0.19689 | 0.19972 | −0.99 | 0.3242 | |
| AI Opponent:Offer | −0.01020 | 0.03544 | −0.29 | 0.7735 | |
| AI Opponent:AI training | −0.00398 | 0.04983 | −0.08 | 0.9363 | |
| Offer:AI training | 0.14419 | 0.04649 | 3.10 | 0.0019 | ** |
| AI Opponent:Offer:AI training | −0.01890 | 0.03544 | −0.53 | 0.5938 | |

*Signif. codes*: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

**Table 14: Mixed Effects Model Results for Experiment 3B**

| Predictor | Estimate | Std. Error | $z$-value | $Pr(>\|z\|)$ | Significance |
|---|---|---|---|---|---|
| (Intercept) | 1.61810 | 0.21285 | 7.60 | $< 0.001$ | *** |
| AI Opponent | −0.00753 | 0.05471 | −0.14 | 0.8906 | |
| Offer | 1.95430 | 0.06306 | 30.99 | $< 0.001$ | *** |
| AI training | −0.60944 | 0.21027 | −2.90 | 0.0038 | ** |
| AI Opponent:Offer | −0.01358 | 0.04003 | −0.34 | 0.7345 | |
| AI Opponent:AI training | 0.03337 | 0.05471 | 0.61 | 0.5419 | |
| Offer:AI training | −0.22911 | 0.05734 | −4.00 | $< 0.001$ | *** |
| AI Opponent:Offer:AI training | 0.09306 | 0.04007 | 2.32 | 0.0202 | * |

*Signif. codes*: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1