

# Opting In? Information About AI Training Affects Human Behavior

LAUREN S. TREIMAN, Washington University in St. Louis, USA

CHIEN-JU HO\*, Washington University in St. Louis, USA

WOUTER KOOL\*, Washington University in St. Louis, USA

When people shop online or interact with chatbots, their behavior generates training data for artificial intelligence (AI). To regulate the use of such data, privacy regulations increasingly require organizations to disclose AI training and obtain explicit consent for this use. However, such disclosures may themselves alter user behavior and affect the collected data. In particular, three distinct mechanisms within the consent process could drive these changes. First, people who voluntarily consent to AI training may differ systematically from those who do not, potentially introducing selection bias into training data. Second, simply being aware that AI training is occurring may trigger behavioral changes. Third, behavioral changes may depend on users actively reading details about AI training, rather than merely knowing that such training is occurring. Across two experiments, we examined whether behavioral changes in training data arise from consent, mere awareness of AI training, or engagement with information about how their data would be used. To test these hypotheses, participants played the ultimatum game, deciding whether to accept monetary proposals. Some participants were informed about AI training, while others were not. In Experiment 1, we manipulated whether participants could opt into AI training and whether they had the option to read information about how their data would train AI for future participants. Although most participants opted into training, most chose not to read the information. Participants who opted in without reading behaved similarly to those unaware of AI training, while those who read rejected more unfair offers. In Experiment 2, we tested whether the amount of information affected engagement and found that, compared to brief disclosures, detailed disclosures produced similar reading times and behavioral changes. These findings show that informing users about AI training alters training data only when users engage with the information, regardless of how much detail is provided. This creates heterogeneous data in which informed and uninformed users behave systematically differently. This work highlights that disclosure practices intended to protect users may directly shape the data used to train AI, underscoring the need to carefully document and account for these behavioral shifts.

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

Additional Key Words and Phrases: AI training, ultimatum game, decision making, AI information

## ACM Reference Format:

Lauren S. Treiman, Chien-Ju Ho, and Wouter Kool. 2026. Opting In? Information About AI Training Affects Human Behavior. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAcT '26)*, June 25–28, 2026, Montreal, QC, Canada. ACM, New York, NY, USA, 26 pages. <https://doi.org/10.1145/3805689.3812384>

## 1 Introduction

When users chat with ChatGPT, scroll through social media, or rate products on Amazon, their interactions are often collected as training data for AI. For years, much of this data collection occurred without users fully

\*Both authors contributed equally to this research.

Authors' Contact Information: Lauren S. Treiman, [ltreiman@wustl.edu](mailto:ltreiman@wustl.edu), Washington University in St. Louis, St. Louis, Missouri, USA; Chien-Ju Ho, Washington University in St. Louis, St. Louis, Missouri, USA; Wouter Kool, Washington University in St. Louis, St. Louis, Missouri, USA.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

*FAcT '26, June 25–28, 2026, Montreal, QC, Canada*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2596-8/2026/06

<https://doi.org/10.1145/3805689.3812384>

being aware that their behavior was used to shape algorithmic decision making. As concerns about user data protection have grown, privacy regulations such as the European Union's General Data Protection Regulation (GDPR) [76] and the California Consumer Privacy Act (CCPA) [17] have required organizations to disclose when user data are collected and how they will be used, and to obtain explicit consent for such use. These regulations aim to give users greater control over their personal data [21, 64]. Implicit in these approaches is the assumption that disclosure increases user awareness. However, awareness itself may change how people provide training data, with important implications for the data used to train AI. If awareness of AI training alters behavior, the resulting data may no longer reflect baseline human decision making, potentially introducing systematic bias into AI systems and harming the users these regulations aim to protect [3, 13].

Recent empirical evidence supports this concern. Prior studies show that when people are made aware that their behavior will be used to train AI, they systematically modify how they act. For example, Cen et al. [9] found that users adjusted their engagement depending on how recommendation algorithms were described: they liked more content when told recommendations were based on likes, and spent more time on content when told recommendations were based on viewing time. Similarly, Treiman et al. [70, 71, 72] examined behavior in the ultimatum game, where participants accepted or rejected monetary offers from another party. They found that participants rejected more unfair offers when told their decisions would be used to train AI. In these settings, people appear to adjust their behavior to reflect how they believe AI should behave, rather than expressing their baseline preferences. This pattern aligns with well-established findings in social science, including Campbell's Law [8] and Goodhart's Law [26], which show that when people know their behavior is being measured or used as a target, they adjust it accordingly [7, 30, 59]. Across these cases, awareness of how people's behavioral data are used plays a central role in shaping behavior.

In practice, however, consent processes do not reliably produce such awareness. Consent procedures involve multiple components, including awareness that training is occurring, the act of consenting, and engagement with explanatory information. Platforms typically rely on consent forms, implicitly assuming that users read and understand the information before deciding whether to participate. This assumption often fails [19, 50]. Many users consent without reading disclosures [2] or without fully understanding their implications [34], often due to consent fatigue [63]. Psychological research helps explain these patterns: people tend to avoid cognitive effort unless sufficiently motivated [6, 41, 45, 47]. Applied to AI training, users may carefully read consent information when motivated to influence outcomes, but otherwise may opt in without engaging with the information or opt out entirely. Engagement may also depend on how consent information is presented, with simpler disclosures more likely to be read.

As a result, real-world consent regimes introduce sources of variation that go beyond awareness alone. These considerations highlight a gap between consent practices and user awareness, motivating closer examination of how consent mechanisms shape behavior and the data used to train AI. More specifically, while prior work shows that awareness of AI training can change behavior [70–72], these studies directly inform participants about AI training. In contrast, consent-based systems introduce three additional mechanisms that remain largely unexplored. First, consent is typically voluntary, meaning that training data may reflect a self-selected subset of users. Second, consent procedures make users aware that they are training AI, and this awareness alone may lead to behavioral shifts. Finally, reading these consent disclosures may trigger behavioral changes, rather than merely knowing the training is occurring. While prior work demonstrates that awareness changes behavior, it remains unclear which of these mechanisms operates in practice and how they interact.

In this work, we study how specific features of the consent process influence user behavior and the resulting data by addressing two research questions:

**RQ1: When people can opt into AI training without reading information, do they choose to inform themselves, and does this choice affect their behavior?**

**RQ2: Does the format of consent information influence whether people engage with it and how they behave when training AI?**

To study these questions, we used the ultimatum game [27]. 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 accepted, both players receive their allotted amounts. If rejected, both receive nothing. This game is widely used to study how fairness concerns influence decision making and provides a straightforward and controlled framework for answering our research questions [57, 74]. Additionally, this paradigm generalizes to several AI training contexts, which we discuss further in the General Discussion.

Prior work using the ultimatum game showed that people rejected more unfair offers when told their decisions would train AI, compared to those unaware of training [70–72]. However, this prior work did not involve a consent process, as all participants were required to read information about AI training before participating. As a result, it remains unclear whether behavioral changes arise from awareness itself, from the act of consenting, or from their interaction.

We disentangle these mechanisms by incorporating an explicit AI-training consent process that mirrors real-world practice. In Experiment 1, we manipulated whether participants could opt into AI training and, among those who could opt in, whether they could choose to read information about what AI training involved. This design enabled us to disentangle the effects of consent and awareness on training behavior. We found that most participants chose to opt into AI training. However, among those who could decide whether to read the information, the majority opted in without doing so. Critically, behavioral change depended on information engagement rather than mere awareness of AI training: participants who opted in without reading behaved like those never informed about AI training, whereas those who read the information rejected more unfair offers.

We next examined whether the presentation of consent information influenced engagement and behavior. In Experiment 2, participants required to train AI received the same information in different formats, either as three concise sentences or as a longer "pop-up" screen requiring an explicit mouse button click on an "I agree" button, emulating real-world consent interfaces [48, 73]. We measured engagement using time spent reading the information. We found that when participants were required to read the information, their behavior changed similarly regardless of presentation format.

Taken together, these results highlight how disclosure practices systematically shape training data. When users engage with disclosures about AI training, their behavior changes, while users who consent without reading provide unmodified data. As a result, policymakers and systems designers need to consider how to balance user autonomy with the systematic differences in training data that disclosure itself produces. Moreover, from a modeling and technical perspective, this also raises the need to account for potential distributional shifts when training AI on human behavior.

## 2 Related Work

### 2.1 Behavioral Changes from AI Training Awareness

Recent research demonstrates that informing people their behavior will train AI changes how they act. Treiman et al. [70, 71] showed how people modify their behavior when training AI using the ultimatum game [27]. They found that participants rejected significantly more unfair offers when told their decisions would train an AI proposer they would encounter in a follow-up session. This behavioral shift occurred even when they were training AI that only other participants would encounter (i.e., not themselves), suggesting that people are guided by altruistic motivations. In subsequent work, Treiman et al. [72] also demonstrated that people tend to rely on automatic, intuitive information processing rather than deliberate strategic thinking when training AI to play the ultimatum game.

These findings highlight a challenge in AI development: when participants actively engage with information about AI training, they may not provide baseline behavioral data. Instead, they embed beliefs and norms into training data through automatic, intuitive processes. Similar effects occur in other contexts. For example, Cen et al. [9] found that users strategically altered engagement patterns when recommendation algorithms were disclosed, liking more content when told recommendations were based on likes, or spending more time on content when told they were based on viewing time, biasing training data based on the disclosed algorithm.

However, across all these studies [9, 70–72], participants were *required* to read information about AI training. This leaves open the critical question our research addresses: what happens when people can *choose* whether to inform themselves about AI training? Here, we investigate how this choice affects how people train AI.

## 2.2 Consent Processes and Information Engagement

Research on digital consent shows that interface design often shapes user decisions more than the information itself. Nouwens et al. [55] found that removing opt-out buttons from the first page of the consent popup increased consent rates by 22-23 percentage points. Similarly, Utz et al. [73] found that approximately 30% of users accepted all cookies when check boxes were preselected, compared to less than 0.1% when nothing was preselected. They also found that users were more likely to interact with cookie notices that appeared in the lower left part of the screen.

However, consent decisions represent only part of the problem: most users spend minimal time engaging with consent information. Utz et al. [73] found that users spent a median of 4–5 seconds on simple consent notices and 7–8 seconds on more detailed notices, far too brief to meaningfully comprehend their implications. Similarly, Bakos et al. [2] showed that only one or two in every thousand users read licensing agreements, and even among those who did, reading times were insufficient for comprehension. Subsequent work replicates this finding, demonstrating that most people do not engage deeply enough with consent materials to understand them [19, 34, 50].

To address this issue, researchers have developed methods to simplify consent forms and improve readability [29, 36, 69]. However, these efforts have largely failed to improve comprehension. For example, Davis et al. [15] found that while people prefer easier consent forms, comprehension does not increase with reduced complexity. This work suggests that improving surface-level readability alone is insufficient to change how people process consent information.

Together, these findings reveal that consent failures arise not from excessive complexity, but from a lack of motivation to engage. Simplified forms do not improve comprehension because users often do not read them in the first place. Despite reporting strong privacy concerns, individuals routinely share personal information without careful consideration, a phenomenon known as the privacy paradox [40]. Repeated exposure to consent requests further exacerbates this issue, leading to consent fatigue, in which users develop habitual acceptance patterns and disengage from the content entirely [63]. As a result, interface design features such as button placement and default selections dominate user behavior, while informational content plays a minimal role [55, 73].

Importantly, the studies discussed above focus on privacy consent contexts in which users perceive little personal benefit from engaging with the information. Contexts where people provide training data for AI may differ in critical ways. Unlike privacy policies that primarily describe passive data collection, AI training consent offers users the opportunity to actively influence downstream AI behavior. Prior work shows that awareness of AI training can meaningfully alter decision making [70–72], suggesting that users may be more motivated to engage with training information than with privacy notices. The current research tests whether patterns of consent fatigue observed in privacy contexts generalize to AI training, and whether the perceived ability to shape AI outcomes changes how, and how much, people engage with consent information.

### 2.3 Training Data Quality

The quality of training data directly affects the fairness and performance of AI models. Biased or unrepresentative training data can lead to discriminatory outcomes [3, 13], particularly in high-stakes domains such as healthcare [25, 39, 56] and law [32, 49]. Studies have documented how AI trained on skewed data perpetuates or amplifies existing inequalities [5, 58]. Kamulegeya et al. [38] found that an algorithm trained on a dataset of predominantly light-skinned patients performed 50% less accurately on darker skin tones. Similarly, Angwin et al. [1] analyzed a recidivism prediction algorithm and found that it incorrectly classified Black defendants as high risk at roughly twice the rate of White defendants.

How people change their behavior when training AI introduces an unexamined source of training data bias. When people are aware their behavior will train AI, they do not provide baseline behavior but instead demonstrate what they believe AI should learn, embedding their normative beliefs and social biases into the training data [70–72]. As a result, AI may learn from strategically modified behavior rather than behavior that reflects baseline human preferences, potentially amplifying biases when deployed in high-stakes contexts.

Our research reveals an additional dimension to this problem: if behavioral modification depends on whether people choose to engage with AI training information, systematic differences may emerge between those who choose to train AI and those who don't. This creates the potential for self-selection bias in training data: the people who opt to train AI may differ systematically from the broader population [51]. Combined with the strategic behavioral modification documented in Section 2.1, consent processes shape who chooses to participate in AI training and how they behave when they do, introducing two compounding sources of bias. Understanding how consent design influences both self-selection and behavioral shifts is therefore critical for ensuring fair and representative AI in high-stakes domains.

## 3 RQ1: When People Can Opt Into AI Training Without Reading Information, Do They Choose to Inform Themselves, and Does This Choice Affect Their Behavior?

We designed Experiment 1 to investigate how consent and information about AI training affect decision making. To do this, we used the ultimatum game [27]. Participants played multiple rounds as responders, deciding whether to accept or reject divisions of money proposed by either AI or human partners. We manipulated whether participants could choose to read information about AI training and whether they could opt into training. Participants were randomly assigned to one of four conditions:

- **Forced reading, forced training:** Participants were required to read detailed information about AI training and train AI. This replicated prior work where consent and information about AI training were mandatory [70–72].
- **Forced reading, optional training:** Participants were required to read detailed information about AI training but could choose whether to opt into training. This condition tested whether participants chose to opt into AI training and how this decision affected their behavior.
- **Optional reading, optional training:** Participants could choose whether to view information about AI training and whether to opt into training. This condition simulated real-world consent processes where both information engagement and participation are voluntary.
- **Control:** Participants received no information about AI training and were not part of training.

This design helps us answer RQ1 by testing four questions. First, when given the choice, do people inform themselves about AI training? Second, do behavioral changes occur only among those who read the information, or also among those who consent without reading? Third, does making a consent decision itself influence behavior? Fourth, does awareness that AI training is occurring, regardless of reading information or consenting, lead to behavioral changes?

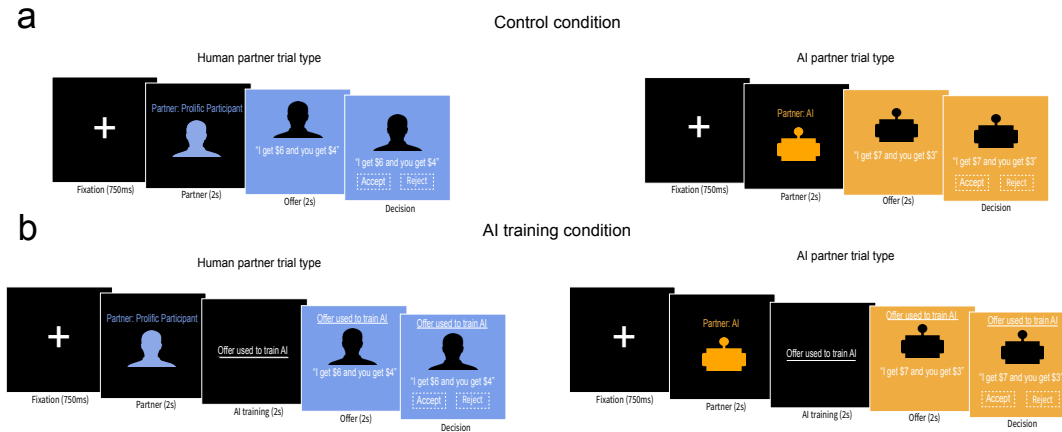


Fig. 1. Example trials for the control (a) and AI training (b) conditions for each partner type (left human participant and right AI). In the AI training condition, participants saw an additional reminder that their responses were training the AI, which was not shown in the control condition. Aside from this reminder, the trial format was identical across conditions. Participants not training AI completed the same task as those in the control condition. Each trial began with a fixation cross (750ms), followed by the partner type (human or AI) (2s). Participants in the AI training condition then saw the reminder screen (2s). All participants then saw the offer amount (2s) 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.

### 3.1 Design

At the start of the experiment, all participants were briefed on the rules of the ultimatum game and told they would play as the responder. They were then randomly assigned to one of four conditions. In the **forced reading, forced training** condition ( $n = 150$ ), participants were required to read that their responses would train an AI proposer that other participants would play with in a follow-up session, and could not opt out of training. In the **forced reading, optional training** condition ( $n = 144$ ), participants received the same information but could choose whether to opt into AI training. In the **optional reading, optional training** condition ( $n = 133$ ), participants made two independent decisions. They could choose whether to view information explaining that their responses would be used to train an AI proposer for future participants, and they could choose whether to opt into training. In the **control** condition ( $n = 161$ ), participants received no information about AI training.

Next, participants played multiple rounds of the ultimatum game (Figure 1). In each round, participants chose whether to accept or reject a proposer's offer of how to allocate a \$10 sum between both partners. Participants played against both AI and human partners, with partner type randomly associated with a color, blue or orange, to help participants distinguish between them.

Each round began with a fixation cross (750ms), followed by a two-second display of the partner type icon (human participant or AI). Participants training AI saw an additional screen with the text "Offer used to train AI" (2s) to remind them that they were training AI. 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 training AI 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 also 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 to ensure that we could balance offers between training conditions using the same amounts.

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.

### 3.2 Analysis

We employed logistic mixed-effects models to assess how partner type, training condition, offer amount, and their interactions predicted participants' acceptance of offers. Partner type was contrast coded (AI = 1, human = -1) such that the intercept and effect of training condition could be interpreted across both partner types rather than for one partner type alone. Similarly, offer amount was mean-centered such that all effects were interpretable at the average offer amount rather than \$0. Training condition was dummy coded, with the reference category changed across models to investigate differences between groups <sup>1</sup>. The model specification was:

$$P(\text{accept} = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 \times \text{offer} + \beta_2 \times \text{partner} + \beta_3 \times \text{condition} + \beta_4 \times \text{offer} \times \text{partner} + \beta_5 \times \text{offer} \times \text{condition} + \beta_6 \times \text{partner} \times \text{condition} + \beta_7 \times \text{offer} \times \text{partner} \times \text{condition} + u_i)$$

where  $u_i \sim N(0, \sigma_u^2)$  denotes the random intercept for each participant, and the dependent variable was a binary accept/reject decision. Models were estimated in R using the `lme4` package [4], with standard errors computed using the default method in the `lme4` package and  $p$ -values obtained using the `lmerTest` package [43]. Our analysis focused on the statistical significance of differences between conditions rather than the magnitude of effects, as our primary goal was to determine whether participants' behavior differed depending on whether they were informed about and opted into AI training rather than to quantify the size of these differences. We used this approach for both experiments. Complete mixed-effects model results for both experiments are reported in Appendix F.

Following prior literature [54], we considered offers of \$1-\$3 to be unfair and offers of \$4-\$6 to be fair. Stimuli, data, and analysis scripts for all experiments can be found on the Open Science Framework (OSF).<sup>2</sup>

### 3.3 Results

**3.3.1 Participants.** We aimed to recruit at least 100 participants per condition, as prior work has detected significant effects with this sample size [70–72]. Because participants in some conditions could opt out of or read information about AI training, we overrecruited to ensure sufficient sample sizes across conditions. As a result, a total of 591 participants (305 female, 6 non-binary 1 missing;  $M = 38.35$ ,  $SD = 12.88$ ) were recruited from Prolific. Four participants were excluded for completing the task twice. The average completion time was 8 minutes, with a median pay of approximately \$10 per hour (base rate of \$8.50 per hour plus a bonus). For both experiments, all participants provided informed consent, and the study was approved by the Washington University in St. Louis IRB.

**3.3.2 Opt-in and Information Engagement Rates.** We found that participants opted into AI training regardless of whether they were informed about details of the training process. Specifically, among participants who were

<sup>1</sup>Regression specifications with explicit condition names for both experiments are reported in Appendix E.

<sup>2</sup>Link found here: <https://osf.io/pv8ku>

required to read information about AI training, 90% (130 of 144) chose to opt in. Similarly, among participants who could opt in without reading any information, 95% (126 of 133) opted in.

However, participants chose not to read information about AI training. Specifically, 81% (108 of 133) of participants in the optional reading condition opted in without viewing any information about AI training.

Based on participants' decisions, we identified a key distinction that allows us to isolate the effect of information on behavior. Most participants in the optional reading, optional training condition chose not to view information but still opted into training. This created a natural comparison: some participants opted in and were required to read information (from the forced reading, optional training condition), while others opted in without reading any information (from the optional reading, optional training condition).

We therefore defined two analysis conditions based on whether participants read information about AI training: the **informed opt-in training** condition ( $n = 130$ , participants who read and opted in) and the **uninformed opt-in training** condition ( $n = 108$ , participants who opted in without reading). Together with the control ( $n = 161$ ) and forced training ( $n = 150$ ) conditions, this allows us to investigate how information engagement and voluntary participation affect behavior. Table 1 provides a detailed breakdown.<sup>3</sup>

Condition	# participants	Informed opt-in	Informed opt-out	Uninformed opt-in	Uninformed opt-out
Forced reading, forced training	<b>150</b>	-	-	-	-
Forced reading, optional training	144	<b>130</b>	14	-	-
Optional reading, optional training	133	17	3	<b>108</b>	4
Control	<b>161</b>	-	-	-	-

Table 1. Number of participants per condition for Experiment 1. We focused analyses on groups with adequate sample sizes (bolded:  $n = 150$ ,  $n = 130$ ,  $n = 108$ ,  $n = 161$ ), excluding groups with fewer than 20 participants.

**3.3.3 Behavioral Changes by Information.** We found that behavioral change depended on whether participants read information about AI training, not whether they consented (Figure 2). Participants who read information rejected more offers similarly to those forced to train AI. In contrast, participants who opted in without reading behaved identically to the control condition.

A logistic mixed-effects model revealed a main effect of offer amount ( $b = 1.88$ ,  $SE = 0.06$ ,  $p < 0.001$ ), replicating prior work [54, 65, 75]. Participants in the forced training condition rejected more offers than those in the control condition ( $b = 1.04$ ,  $SE = 0.34$ ,  $p = 0.003$ ), aligning with previous findings [70–72].

To address our research questions, we examined how the informed opt-in training and uninformed opt-in training conditions compared to both the control and forced training conditions. Participants in the informed opt-in training condition rejected more offers than those in the control ( $b = 0.79$ ,  $SE = 0.36$ ,  $p = 0.03$ ) and the uninformed opt-in training ( $b = 0.87$ ,  $SE = 0.40$ ,  $p = 0.03$ ) conditions. However, their responses did not differ from those in the forced training condition ( $b = -0.25$ ,  $SE = 0.36$ ,  $p = 0.49$ ).

In contrast, participants in the uninformed opt-in training condition showed the opposite pattern: they responded no differently from those in the control condition ( $b = -0.08$ ,  $SE = 0.38$ ,  $p = 0.84$ ) and rejected fewer offers than those in the forced training condition ( $b = -1.11$ ,  $SE = 0.38$ ,  $p = 0.004$ ).

We found no interactions between training condition and offer amount (all  $ps \geq 0.06$ ), contrasting with prior work [70–72]. However, visual inspection of Figure 2 suggested effects were particularly strong for unfair offers ( $\leq \$3$ ) that the mixed effects model could not capture the interaction effects. Following prior work [70–72], we conducted post-hoc  $t$ -tests on unfair offers only. These tests confirmed that both the forced training and informed opt-in training conditions rejected more unfair offers than the uninformed opt-in training and control

<sup>3</sup>Due to low sample sizes, we excluded participants who opted out of AI training or who viewed information in the optional reading condition but did not opt in.

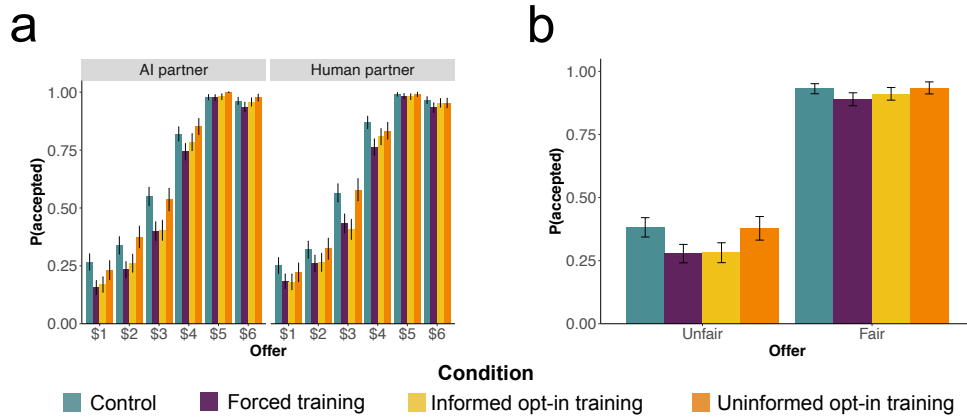


Fig. 2. Experiment 1 results. Graphs show the proportion of accepting an offer based on (a) offer amount and (b) fairness conditioned on partner type and fairness. The informed opt-in training condition refers to participants who read information and opted in (a subset of the forced reading, optional training condition), while the uninformed opt-in training condition refers to participants who opted in without reading (a subset of the optional reading, optional training condition). Error bars indicate standard error.

conditions (see Appendix A for detailed results). These findings suggest that information about AI training increased rejection of unfair offers.

We found that participants responded no differently when partnered with a human compared to an AI partner ( $b = .10$ ,  $SE = 0.08$ ,  $p = 0.21$ ). While the mixed-effects model identified additional interaction effects with partner type, they were not relevant to our research questions and are reported in Appendix D.

### 3.4 Discussion

We found that when given the opportunity to train AI, the vast majority of participants chose to opt into it. However, when given the choice whether to read information about AI training, most people chose to train AI without reading about what the training involved. This lack of engagement had critical consequences for behavior. Participants who consented without reading behaved identically to the control condition, while participants who read information about AI training rejected unfair offers similarly to the forced condition. We replicated this informed opt-in pattern in an independent sample (see Appendix B). Together, these findings demonstrate that behavioral modification depends on actual engagement with information about the AI training process, not merely knowing that training is occurring or voluntarily participating in AI training.

## 4 RQ2: Does the Format of Consent Information Influence Whether People Engage With It and How They Behave When Training AI?

Experiment 1 showed that reading information about AI training led people to change their behavior when training AI. However, it remains unclear how information presentation influences participant engagement. In practice, consent information appears in many formats: some platforms present brief, easily skippable text, whereas others require explicit acknowledgment through mechanisms such as clicking “I agree.” These differences in format may shape both attention to the information and subsequent behavioral change.

Experiment 2 was designed to test whether the format of AI training information affects how long participants engage with it and how they subsequently behave. As in Experiment 1, participants played the ultimatum game as

responders. Participants in the AI training conditions were always presented with information about AI training, but we varied its format. Some received three sentences of brief text displayed on the screen, while others received a detailed three-paragraph popup message requiring them to click an 'I agree' button. To assess engagement, we measured how long participants spent viewing the instructions. Participants were randomly assigned to one of three conditions:

- **Forced training–brief:** Participants were presented with three sentences describing AI training. They could not opt out of training. This condition replicated the forced training condition from Experiment 1.
- **Forced training–detailed:** Participants viewed a popup message describing AI training in three paragraphs and were required to click “I agree” to proceed. They could not opt out of training. This format mirrors common real-world consent interfaces.
- **Control:** Participants received no information about AI training and were not part of training.

#### 4.1 Design

The design was nearly identical to Experiment 1 (see Figure 1), with two key modifications. First, because Experiment 1 showed that participants overwhelmingly consented to AI training when given a choice, we required all participants in the training conditions to train the AI (i.e., they could not opt out). This modification allowed us to isolate the effect of information format on engagement and behavior. Second, we introduced a new consent format in which AI training information was presented as a detailed popup requiring explicit acknowledgment.

Participants were randomly assigned to one of three conditions. In the **forced training–brief** condition ( $n = 142$ ), participants viewed information about AI training presented in three sentences, identical to the forced training condition in Experiment 1. In the **forced training–detailed** condition ( $n = 153$ ), participants received the same core information expanded into three paragraphs, presented in a popup window. Participants were required to click “I agree” to proceed. The AI training information text for both conditions is provided in Appendix C. In the **control** condition ( $n = 149$ ), participants received no information about AI training, identical to the control condition of Experiment 1.

To assess engagement, we recorded how long participants spent viewing the AI training information. In the **forced training–brief** condition, this was the time spent on the AI training information screen. In the **forced training–detailed** condition, this was the time the popup remained open before participants clicked “I agree.”

The ultimatum game procedure, partner types, offer amounts, and incentive structure were identical to Experiment 1.

#### 4.2 Results

**4.2.1 Participants.** A total of 448 participants (227 female, 7 non-binary, 1 missing;  $M = 41.53$ ,  $SD = 13.10$ ) were recruited from Prolific. Three participants were excluded: two for completing the task twice and one for refreshing the webpage and being exposed to multiple conditions. The average completion time was 7.5 minutes, with a median pay of approximately \$10.25 per hour (base rate of \$9.25 per hour plus a bonus).

**4.2.2 Reading Time by Format.** We first examined whether format affected how long participants spent reading AI training information. Despite the detailed format containing substantially more text (three paragraphs vs. three sentences), participants in the forced training–detailed ( $M = 17.4$  seconds,  $SE = 2.2$ ) and forced training–brief conditions ( $M = 12.7$  seconds,  $SE = 1.8$ ) spent similar time reading as those in the forced training–brief condition ( $W = 11448$ ,  $p = 0.63$ ). This suggests participants did not fully read the longer, more detailed format.

To confirm this pattern reflected differences in AI training information specifically rather than differences in reading speed across conditions, we examined time spent on general instructions unrelated to AI training. Pairwise

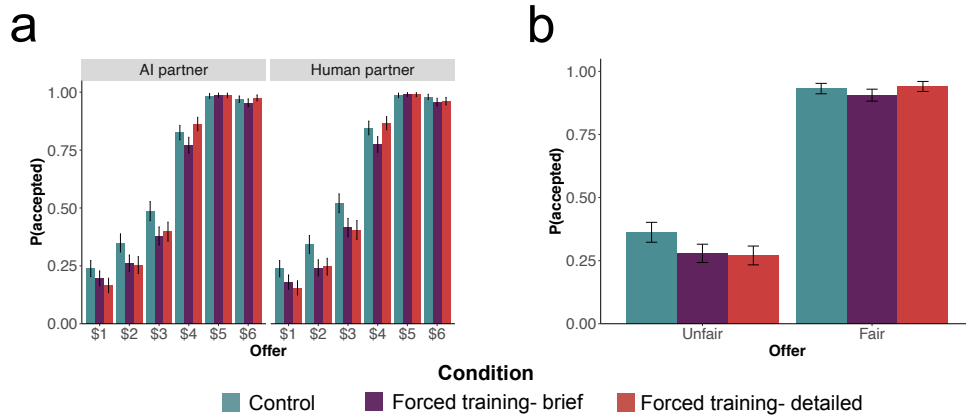


Fig. 3. Results for Experiment 2. Graphs show the proportion of accepting an offer as a function of (a) offer amount and (b) fairness, conditioned on partner type. Error bars indicate standard error.

Wilcoxon rank-sum tests revealed no differences between conditions ( $ps \geq 0.24$ ),<sup>4</sup> confirming comparable baseline reading speeds.

**4.2.3 Behavioral Changes in AI Training by Format.** We next examined how the information format affected participants' training behavior. Both formats produced similar behavioral changes when training AI (Figure 3). Participants in both the forced training–brief and forced training–detailed conditions rejected more unfair offers than the control condition, and the two training conditions did not differ from each other.

A logistic mixed-effects model revealed a main effect of offer amount ( $b = 2.36$ ,  $SE = 0.09$ ,  $p < 0.001$ ), replicating prior work [54, 65, 75]. When comparing the forced training–brief and control conditions, the model found neither a main effect of training condition ( $b = -0.78$ ,  $SE = 0.43$ ,  $p = 0.07$ ) nor an interaction effect with offer amount ( $b = 0.08$ ,  $SE = 0.13$ ,  $p = 0.54$ ). However, visual inspection of Figure 3 suggested participants in the forced training–brief condition were more punitive toward unfair offers. A  $t$ -test confirmed this pattern ( $t_{441} = 2.15$ ,  $p = 0.03$ ).

Participants in the forced training–detailed condition showed the same pattern. They rejected more offers than the control condition ( $b = -0.96$ ,  $SE = 0.42$ ,  $p = 0.02$ ) but did not differ from the forced training–brief condition ( $b = -0.18$ ,  $SE = 0.42$ ,  $p = 0.66$ ). Post-hoc  $t$ -tests confirmed they were more punitive for unfair offers than the control condition ( $t_{441} = 1.99$ ,  $p = 0.047$ ) but did not differ from the forced training–brief condition ( $t_{441} = 0.20$ ,  $p = 0.84$ ).

Partner type did not affect responses ( $b = -0.11$ ,  $SE = 0.07$ ,  $p = 0.13$ ), with no interaction effects found ( $ps \geq 0.17$ ).

### 4.3 Discussion

Experiment 2 showed that providing substantially different amounts of information, three sentences versus three paragraphs, did not meaningfully change how long participants spent reading AI training instructions. This similarity in engagement time was mirrored in behavior: participants in the forced training–brief and forced training–detailed conditions exhibited comparable rejection patterns relative to the control condition.

<sup>4</sup>Control vs. forced training–detailed:  $W = 12025$ ,  $p = 0.75$ ; control vs. forced training–brief:  $W = 10364$ ,  $p = 0.38$ ; forced training–brief vs. forced training–detailed:  $W = 10217$ ,  $p = 0.24$ .

These results suggest that participants did not engage more deeply with the detailed popup information, instead spending just enough time to register the core fact that their behavior would be used to train an AI model. As a result, both formats produced similar behavioral changes. This pattern indicates that awareness of AI training itself, rather than engagement with or comprehension of detailed explanatory content, is sufficient to elicit behavioral modification.

## 5 General Discussion

### 5.1 Recap

As privacy regulations now require organizations to disclose when they collect human data for AI training, it is important to understand how these disclosures shape the training data itself. If disclosure changes how people behave, the data used to train AI may no longer reflect baseline human decision making. This research examined how consent processes and information engagement influence behavior when training AI. Across two experiments, participants played the ultimatum game as responders, deciding whether to accept or reject offers from AI and human partners. Some participants were informed their decisions would train an AI proposer for future participants, while others could choose whether to train AI.

Experiment 1 tested whether consenting to AI training and choosing whether to inform oneself about it affects how people train AI. We manipulated whether participants could opt into AI training and whether they could choose to read information about the training process. We found that most participants ( $\geq 90\%$ ) opted for AI training, but the majority (81%) did so without reading information about what AI training involved. Critically, only participants who read information modified their behavior, rejecting more unfair offers than the control group. Participants who consented without reading behaved identically to controls, demonstrating that consent alone does not drive behavioral change: only information engagement does.

Experiment 2 examined whether the format of consent information influences engagement and behavior. We compared two formats: brief text (three sentences) versus a detailed popup (three paragraphs). Despite the substantial difference in text length, participants spent similar time reading both formats and showed similar behavioral changes. Both training conditions rejected more unfair offers than controls, suggesting that providing any information about AI training drives behavioral modification.

Across both experiments, we found main effects of training condition rather than interactions with offer amount, unlike prior work [70–72]. However, exploratory post-hoc *t*-tests confirmed that participants who received information about AI training were more punitive toward unfair offers. Together, these findings reveal that information engagement, not consent decisions or presentation format, determines whether people modify their behavior when training AI.

### 5.2 Behavioral Changes Depend on the Presence of Information

Our findings show that behavioral shifts when training AI depend on reading information about the training, not simply knowing it exists or opting into it. In Experiment 1, participants who consented to AI training without reading any explanatory information behaved indistinguishably from the control group, despite being aware that AI training existed. This pattern suggests that awareness alone is insufficient to drive behavioral change. Behavioral modification emerged only when participants engaged with information describing how their behavior would be used to train AI.

Experiment 2 clarifies that it is the presence of this information, rather than the amount of information processed, that drives behavioral change. Participants modified their behavior regardless of whether AI training information was presented in a brief or detailed format. Although the detailed format contained substantially more text, participants spent similar time reading both versions. This pattern indicates that participants in the

detailed condition did not fully read the text but still extracted the core message: that they were training an AI to play against future participants.

Several mechanisms could explain this efficient extraction. Information foraging theory [60, 61], which describes how people optimize information gathering by stopping once they've found what they seek, would predict that participants read sequentially and stopped once encountering the AI training message. Alternatively, research on skim reading [20] shows that readers focus on early portions of text. Applied to our results, participants may have engaged in strategic skimming of the detailed popup, bypassing peripheral details about data privacy and storage because the core AI training message appeared in the first half. Finally, guided visual search theory [77], which explains how people scan for specific target information, would predict that participants scan for keywords related to AI training while bypassing peripheral details. Eye-tracking research may distinguish among these explanations by revealing whether participants focus early in the text (skimming), stop after key information (foraging), or scan non-sequentially (visual search).

Together, these findings have important implications for AI training contexts. Providing more detailed consent information does not ensure deeper engagement or comprehension. Instead, users extract minimal information needed to recognize that AI training is occurring and adjust their behavior accordingly. The framing of the core message appears to be the primary factor shaping how people behave when training AI.

### 5.3 Implications for Transparency and AI Training

Our findings reveal an important dynamic at the heart of transparency requirements for AI training. Informing users that their behavior will train AI changes the very behavior being collected. Prior frameworks such as Goodhart's Law [26] and Campbell's Law [8] focus on how people game metrics once they know they are being evaluated. Our results reveal a related dynamic at an earlier stage, as people modify their behavior when generating data used to train AI. As a result, when organizations disclose AI training practices, they may introduce systematic variation into their training data as users modify their behavior in response. While transparency has been advocated as essential for accountability and trust in AI systems [16, 52], our findings suggest that full transparency about AI training may introduce additional biases into training data that require careful consideration.

Notably, participants modified their behavior even though we provided minimal detail about the AI training process. We informed participants that their responses would train an AI proposer that would play against future participants, but we did not explain how the AI would learn from their behavior or what algorithms would be used. Despite this limited information, participants still changed how they responded. This suggests that behavioral modification does not require technical understanding of AI training. Instead, participants appear to rely on their own assumptions and beliefs about AI when deciding how to behave [72].

This finding connects to broader research on user strategization: the tendency for people to adjust their behavior to influence algorithmic outcomes [10, 18, 31, 68]. Prior work has shown that users modify behavior when recommendation algorithms are disclosed [9]. Our research further demonstrates that strategization occurs even without explicit information about algorithmic mechanisms. People form beliefs about how AI learns and adjust their behavior, embedding their expectations about fairness, appropriate AI behavior, and social norms into the training data.

However, these behavioral adjustments are not necessarily problematic. In some cases, they may reflect users expressing normative preferences about fairness and attempting to steer AI behavior in socially desirable directions [70, 71]. For example, participants in our experiment may have modified their behavior to instill fairness into the AI, teaching it to make fair offers to future participants. Whether these shifts constitute bias or a form of democratic correction depends on what one considers representative and fair training data.

Regardless of the underlying motives, such shifts should be documented and made transparent. One way to support this transparency is to adopt practices similar to "Datasheets for Datasets" [24], where all aspects of the training process, including how consent procedures may have shaped participation and behavior, are documented and made available to policymakers.

Our findings also raise considerations for AI developers. When training data are collected under transparency requirements, some users will read the disclosures and modify their behavior while others will not. This creates systematic variation in the training data depending on whether users engage with information. As a result, models trained on this mixed data may not accurately reflect the baseline behavior they are intended to capture. Developers should account for this variation when building and evaluating AI trained on human behavior.

#### 5.4 Sociotechnical Implications

Our findings have important implications for how different communities engage with the consent process. Research on informed consent has documented distrust of consent processes among marginalized groups, stemming from a legacy of mistreatment and research abuses [12, 23]. In focus groups, marginalized groups have expressed beliefs that consent processes are designed to protect researchers rather than participants [23]. In AI training contexts, marginalized communities may similarly perceive consent requests as extractive or predatory, designed to benefit developers rather than users. As a result, marginalized groups may be less likely to opt into AI training, producing training sets that do not represent the broader population and risk amplifying existing inequities [3, 13]. Rather than treating this as a self-selection bias problem, developers should consider how the consent processes may reproduce existing power imbalances [14] and work to design consent procedures that are culturally sensitive and build trust within the communities [62].

#### 5.5 Why People Don't Read Information About AI Training

In Experiment 1, we found that most participants opted into AI training, yet the majority did so without viewing information about the training process. This finding parallels research on cookie consent, where users routinely accept notices while ignoring details [28, 37, 46, 73]. Participants may not have cared enough about AI training to invest effort in reading, consistent with the privacy paradox [40]. Additionally, participants accustomed to clicking through consent forms in online contexts may have treated AI training consent the same way, reflecting consent fatigue [63].

Two explanations could account for why participants avoided reading AI training information. First, they may have believed they already understood what AI training entailed based on prior knowledge, making additional details seem unnecessary. Second, participants may have conducted an implicit cost-benefit analysis [41, 66, 67], weighing the cognitive effort required to read against the perceived value of the information. In our low-stakes study, participants likely did not care strongly about AI training implications. As a result, the cost of reading may have outweighed the perceived benefit. Future research could test these explanations by directly asking participants why they avoided information or by manipulating decision stakes to increase motivation [6, 45].

#### 5.6 Generalization of the Ultimatum Game

We used the ultimatum game, a well-established paradigm for measuring fairness concerns [57, 74], to answer our research questions. In our version of this, participants made *explicit* evaluative judgments about AI proposals. Therefore, our results may generalize to AI training contexts in which users provide direct feedback signals such as rating content or reinforcement learning from human feedback [44, 53, 78]. However, our findings may not extend to different AI training contexts, such as imitation learning [33] or systems that rely on *implicit* feedback [35]. In these settings, achieving the same underlying goal may require users to behave differently during training. For example, in the ultimatum game, a user who wants to teach an AI to make fair decisions

can reject unfair offers in a labeling task, but would instead need to switch roles and propose fair offers under imitation learning. Future work could examine whether the training paradigm itself leads users to systematically modify their behavior and whether those modifications differ in magnitude or direction.

### 5.7 Effect of Partner Type

Participants in our experiments responded to offers from both human and AI partners. This manipulation primarily served to remind participants that AIs are actively participating, but also allowed us to examine whether people respond differently to AI versus human partners. We found no difference in responses to partner types, replicating prior work [70–72]. This contrasts with several studies [11, 54, 65, 75] that report higher acceptance rates for unfair offers from AI proposers. Several factors may explain this discrepancy. First, participants may have felt less interpersonal connection with human proposers, as they were represented by abstract silhouettes in our study and those by Treiman et al. [70, 71, 72]. Additionally, our study was conducted entirely online, while previous studies were conducted in person [11, 54, 65, 75]. Finally, framing the AI models as capable of learning may have led participants to treat them more similarly to human partners. Future research could test these explanations by systematically varying data collection method (online vs. in-person) and partner representation while holding other aspects of the paradigm constant.

### 5.8 Limitations

Our findings should be interpreted with appropriate caution. While the ultimatum game provided a scientifically rigorous method for answering our research questions, behavioral responses may differ in real-world contexts. For example, AI is increasingly used in high-stakes domains such as allocating resources to people experiencing homelessness [42] or kidneys to patients [22]. When people know their training behavior affects such consequential outcomes, they may modify their responses differently than in our experimental setting. Nevertheless, these findings provide important insights into whether people choose to inform themselves when training AI and how this choice affects their behavior. Future work could apply this framework to test information engagement effects in more applied contexts.

We should also note that the stakes in this study were relatively low, with participants earning only 5% of the amount from a single negotiation. This setup may have led participants to rush through the task or opt into AI training without much consideration, as they had little incentive to carefully evaluate their decisions. However, using relatively low stakes better reflects many real-world interactions with AI, where the stakes are often minimal in both value and impact. While individual low-stakes decisions might seem insignificant, they can accumulate into high-stakes consequences. Therefore, understanding human behavior during AI training is crucial, even when the stakes are relatively low.

## 6 Conclusion

This research demonstrates that informing users about AI training changes the very behavior being collected. When given the choice, most people consent to AI training without reading details, and only those who engage with information modify their behavior. Moreover, the format of information presentation matters less than whether users engage with it at all. As a result, privacy regulations requiring transparency about AI training may introduce systematic variation into training data. Users who choose to inform themselves modify their behavior to reflect what they believe AI should learn, while those who consent without reading provide unmodified responses. Thus, when designing disclosures, both user autonomy and the behavioral changes that occur during training deserve attention. AI developers should document key aspects about the training process to ensure that those who use these systems are making informed decisions.

## Acknowledgments

We would like to thank members of the Control and Decision Making Lab and the Ho Lab for their advice and assistance.

## Generative AI Usage Statement

Artificial intelligence tools (Claude and ChatGPT) were used to assist with minor editing tasks, formatting tables, and debugging code. All scientific content, experimental design, data analysis, and interpretations were developed by the authors, who take full responsibility for the content of the paper.

## References

- [1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.
- [2] Yannis Bakos, Florencia Marotta-Wurgler, and David R. Trossen. 2009. Does Anyone Read the Fine Print? Testing a Law and Economics Approach to Standard Form Contracts. *New York University Law and Economics Working Papers* 195 (2009). [https://lsr.nellco.org/nyu\\_lewp/195](https://lsr.nellco.org/nyu_lewp/195)
- [3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.
- [4] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- [5] Kelli A Bird, Benjamin L Castleman, and Yifeng Song. 2025. Are algorithms biased in education? Exploring racial bias in predicting community college student success. *Journal of Policy Analysis and Management* 44, 2 (2025), 379–402.
- [6] Matthew Botvinick and Todd Braver. 2015. Motivation and cognitive control: from behavior to neural mechanism. *Annual review of psychology* 66 (2015), 83–113.
- [7] Adriana Camacho and Emily Conover. 2011. Manipulation of social program eligibility. *American Economic Journal: Economic Policy* 3, 2 (2011), 41–65.
- [8] Donald T Campbell. 1979. Assessing the impact of planned social change. *Evaluation and program planning* 2, 1 (1979), 67–90.
- [9] 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).
- [10] Sarah H Cen, Andrew Ilyas, and Aleksander Madry. 2023. User strategization and trustworthy algorithms. *arXiv preprint arXiv:2312.17666* (2023).
- [11] 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>
- [12] Giselle Corbie-Smith, Stephen B Thomas, and Diane Marie M St George. 2002. Distrust, race, and research. *Archives of internal medicine* 162, 21 (2002), 2458–2463.
- [13] 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.
- [14] Christopher L Dancy and P Khalil Saucier. 2021. AI and blackness: toward moving beyond bias and representation. *IEEE Transactions on Technology and Society* 3, 1 (2021), 31–40.
- [15] Terry C Davis, Hans J Berkel, Randall F Holcombe, Sumona Pramanik, and Stephen G Divers. 1998. Informed consent for clinical trials: a comparative study of standard versus simplified forms. *JNCI: Journal of the National Cancer Institute* 90, 9 (1998), 668–674.
- [16] Karl de Fine Licht and Jenny de Fine Licht. 2020. Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & society* 35 (2020), 917–926.
- [17] Lydia de la Torre. 2018. A guide to the california consumer privacy act of 2018. *Available at SSRN 3275571* (2018).
- [18] Michael A DeVito, Jeffrey T Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. 2018. The algorithm and the user: How can HCI use lay understandings of algorithmic systems?. In *Extended Abstracts of the 2018 CHI Conference on human factors in Computing Systems*. 1–6.
- [19] Benjamin D Douglas, Emma L McGorray, and Patrick J Ewell. 2021. Some researchers wear yellow pants, but even fewer participants read consent forms: Exploring and improving consent form reading in human subjects research. *Psychological methods* 26, 1 (2021), 61.
- [20] Geoffrey B Duggan and Stephen J Payne. 2009. Text skimming: The process and effectiveness of foraging through text under time pressure. *Journal of experimental psychology: Applied* 15, 3 (2009), 228.
- [21] Grace Fox, Theo Lynn, and Pierangelo Rosati. 2022. Enhancing consumer perceptions of privacy and trust: a GDPR label perspective. *Information Technology & People* 35, 8 (2022), 181–204.

- [22] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P Dickerson, and Vincent Conitzer. 2020. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence* 283 (2020), 103261.
- [23] Vicki S Freimuth, Sandra Crouse Quinn, Stephen B Thomas, Galen Cole, Eric Zook, and Ted Duncan. 2001. African Americans' views on research and the Tuskegee Syphilis Study. *Social science & medicine* 52, 5 (2001), 797–808.
- [24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.
- [25] 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>
- [26] Charles AE Goodhart and CAE Goodhart. 1984. *Problems of monetary management: the UK experience*. Springer.
- [27] 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](https://doi.org/10.1016/0167-2681(82)90011-7)
- [28] Hana Habib, Megan Li, Ellie Young, and Lorrie Cranor. 2022. “Okay, whatever”: An evaluation of cookie consent interfaces. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–27.
- [29] Kristie B Hadden, Latrina Y Prince, Tina D Moore, Laura P James, Jennifer R Holland, and Christopher R Trudeau. 2017. Improving readability of informed consents for research at an academic medical institution. *Journal of clinical and translational science* 1, 6 (2017), 361–365.
- [30] 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.
- [31] Andreas Haupt, Dylan Hadfield-Menell, and Chara Podimata. 2023. Recommending to strategic users. *arXiv preprint arXiv:2302.06559* (2023).
- [32] 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>
- [33] Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems* 29 (2016).
- [34] Mark Hochhauser. 2004. Informed consent: reading and understanding are not the same: subjects may read consent forms, but they don't always understand them. *Applied Clinical Trials* 13, 4 (2004), 42–47.
- [35] Dietmar Jannach, Lukas Lerche, and Markus Zanker. 2018. Recommending based on implicit feedback. In *Social information access: systems and technologies*. Springer, 510–569.
- [36] Michael Jefford and Rosemary Moore. 2008. Improvement of informed consent and the quality of consent documents. *The lancet oncology* 9, 5 (2008), 485–493.
- [37] Nikhil Jha, Martino Trevisan, Marco Mellia, Daniel Fernandez, and Rodrigo Irrarrazaval. 2024. Privacy Policies and Consent Management Platforms: Growth and Users' Interactions over Time. *arXiv preprint arXiv:2402.18321* (2024).
- [38] Louis Kamulegeya, John Bwanika, Mark Okello, Davis Rusoke, Faith Nassiwa, William Lubega, Davis Musinguzi, and Alexander Börve. 2023. Using artificial intelligence on dermatology conditions in Uganda: a case for diversity in training data sets for machine learning. *African Health Sciences* 23, 2 (2023), 753–63.
- [39] 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>
- [40] Spyros Kokolakis. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & security* 64 (2017), 122–134.
- [41] 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.
- [42] Amanda Kube, Sanmay Das, and Patrick J. Fowler. 2019. Allocating Interventions Based on Predicted Outcomes: A Case Study on Homelessness Services. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (2019), 622–629. <https://doi.org/10.1609/aaai.v33i01.3301622>
- [43] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82, 13 (2017), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- [44] Nathan Lambert. 2025. Reinforcement learning from human feedback. *arXiv preprint arXiv:2504.12501* (2025).
- [45] 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.
- [46] Jie Li. 2025. What If We Don't Accept the Cookies? *Interactions* 32, 1 (2025), 19–21.
- [47] 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.>

- [48] Ewa Luger, Stuart Moran, and Tom Rodden. 2013. Consent for all: revealing the hidden complexity of terms and conditions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2687–2696.
- [49] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.
- [50] Louise-Anne McNutt, Eve Waltermaurer, Robert A Bednarczyk, Bonnie E Carlson, Jerroo Kotval, Jeanne McCauley, Jacquelyn C Campbell, and Daniel E Ford. 2008. Are we misjudging how well informed consent forms are read? *Journal of Empirical Research on Human Research Ethics* 3, 1 (2008), 89–97.
- [51] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.
- [52] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 2053951716679679.
- [53] Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.
- [54] Laura Moretti and Giuseppe Di Pellegrino. 2010. Disgust selectively modulates reciprocal fairness in economic interactions. *Emotion* 10, 2 (2010), 169.
- [55] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
- [56] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
- [57] 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.
- [58] Mahmut Özer, Matjaz Perc, and H Eren Suna. 2024. Artificial intelligence bias and the amplification of inequalities in the labor market. *Journal of Economy Culture and Society* 69 (2024), 159–168.
- [59] Juan Perdomo, Tijana Zrnic, Celestine Mender-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.
- [60] Peter Pirolli and Stuart Card. 1995. Information foraging in information access environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 51–58.
- [61] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.
- [62] Sandra Crouse Quinn, Mary A Garza, James Butler, Craig S Fryer, Erica T Casper, Stephen B Thomas, David Barnard, and Kevin H Kim. 2012. Improving informed consent with minority participants: results from researcher and community surveys. *Journal of Empirical Research on Human Research Ethics* 7, 5 (2012), 44–55.
- [63] Robert Ranisch. 2021. Consultation with Doctor Twitter: consent fatigue, and the role of developers in digital medical ethics. *The American Journal of Bioethics* 21, 7 (2021), 24–25.
- [64] Viviane Reding. 2011. Your data, your rights: Safeguarding your privacy in a connected world. Speech at Privacy Platform "The Review of the EU Data Protection Framework". Vice-President of the European Commission, EU Justice Commissioner.
- [65] 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>
- [66] 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.
- [67] 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.
- [68] Donghee Shin. 2020. How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. *Computers in human behavior* 109 (2020), 106344.
- [69] Alan R Tait, Terri Voepel-Lewis, Shobha Malviya, and Sandra J Philipson. 2005. Improving the readability and processability of a pediatric informed consent document: effects on parents' understanding. *Archives of pediatrics & adolescent medicine* 159, 4 (2005), 347–352.
- [70] 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.
- [71] 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.
- [72] Lauren S Treiman, Chien-Ju Ho, and Wouter Kool. 2025. Do people think fast or slow when training AI?. In *ACM Conference on Fairness, Accountability, and Transparency*.
- [73] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un) informed consent: Studying GDPR consent notices in the field. In *Proceedings of the 2019 acm sigsac conference on computer and communications security*. 973–990.
- [74] Eric van Dijk and Carsten KW De Dreu. 2021. Experimental games and social decision making. *Annual Review of Psychology* 72 (2021), 415–438.

- [75] 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>
- [76] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.
- [77] Jeremy M Wolfe. 2021. Guided Search 6.0: An updated model of visual search. *Psychonomic bulletin & review* 28, 4 (2021), 1060–1092.
- [78] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (2022), 364–381.

## A Experiment 1 *t*-Test Results

To examine whether training conditions were more punitive toward unfair offers specifically, we conducted exploratory *t*-tests comparing acceptance rates for offers  $\leq$  \$3 across conditions. Participants in the forced training and informed opt-in training conditions rejected significantly more unfair offers than those in the control and uninformed opt-in training conditions. Participants in the uninformed opt-in training condition did not differ from controls. Responses in the forced training and informed opt-in training conditions did not differ from each other. Table 2 presents detailed results.

Comparison	Estimate	SE	df	<i>t</i>	<i>p</i>	Sig.
Control vs. Forced training	0.104	0.041	545	2.57	0.010	*
Control vs. Informed opt-in	0.101	0.042	545	2.39	0.017	*
Control vs. Uninformed opt-in	0.004	0.044	545	0.09	0.930	
Forced vs. Informed opt-in	-0.004	0.043	545	-0.09	0.932	
Forced vs. Uninformed opt-in	-0.100	0.045	545	-2.23	0.026	*
Informed vs. Uninformed opt-in	-0.097	0.046	545	-2.08	0.038	*

Table 2. Pairwise *t*-test comparisons for acceptance rates of unfair offers ( $\leq$  \$3) in Experiment 1. \* indicates  $p < 0.05$ .

## B Supplementary Study: Replication of Informed Opt-In Pattern

### B.1 Method

The method was identical to Experiment 1 with one exception: we did not include the optional reading, optional training condition. Participants were randomly assigned to one of three conditions: forced reading, forced training ( $n = 111$ ), forced reading, optional training ( $n = 224$ ), or control ( $n = 107$ ).

### B.2 Results

A total of 446 participants (257 female, 2 non-binary, 1 missing;  $M = 38.26$ ,  $SD = 12.18$ ) were recruited from Prolific. Two participants were excluded: one for completing the task twice and another for refreshing the page and being exposed to multiple conditions. The average completion time was 8 minutes, with a median pay of approximately \$10 per hour (base rate of \$8.50 per hour plus a bonus). All participants provided informed consent, and the study was approved by the Washington University in St. Louis IRB.

Among participants in the forced reading, optional training condition, 86% (192 of 224) chose to opt into AI training. Due to low sample size, we excluded participants who opted out of training. As in Experiment 1, we refer to participants who read and opted in as the *informed opt-in training* condition and those required to train as the *forced training* condition.

The results are shown in Figure 4. A logistic mixed-effects model revealed a main effect of offer amount ( $b = 1.84$ ,  $SE = 0.08$ ,  $p < 0.001$ ). Additionally, participants in the forced condition rejected more offers than those in the control condition ( $b = -0.83$ ,  $SE = 0.40$ ,  $p = 0.04$ ). Although there was no interaction between training condition and offer amount ( $b = 0.16$ ,  $SE = 0.11$ ,  $p = 0.16$ ), an exploratory post-hoc *t*-test on unfair offers revealed the expected pattern. Participants in the forced training condition rejected significantly more unfair offers than those in the control condition ( $t_{407} = 2.42$ ,  $p = 0.02$ ), replicating Experiment 1.

Critically, participants in the informed opt-in training condition rejected more offers than those in the control condition ( $b = -1.35$ ,  $SE = 0.36$ ,  $p < 0.001$ ) and did not differ from the forced training condition ( $b = -0.52$ ,  $SE = 0.35$ ,  $p = 0.13$ ). Post-hoc *t*-tests on unfair offers confirmed this pattern: participants in the informed opt-in training condition rejected significantly more unfair offers than controls ( $t_{407} = 3.72$ ,  $p < 0.001$ ) and did not differ

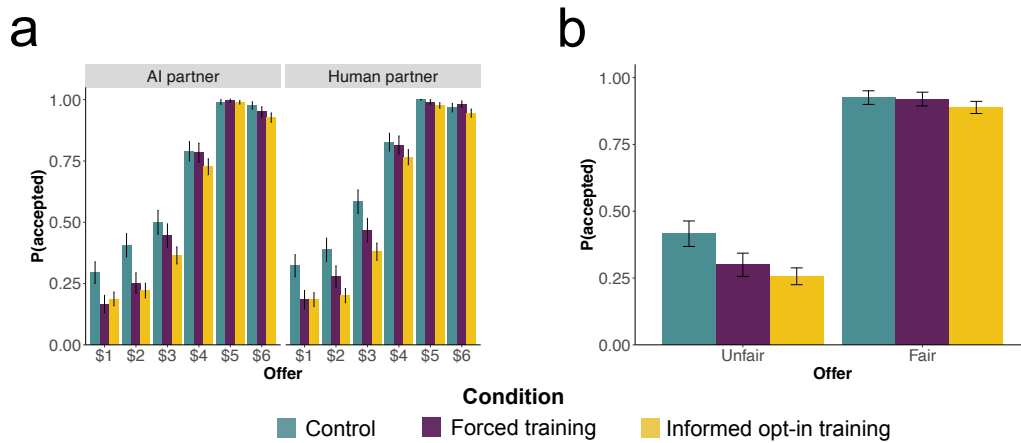


Fig. 4. Results of the replication experiment. Graphs show the proportion of accepting an offer based on (a) offer amount and (b) fairness conditioned on partner type and fairness. The informed opt-in training condition refers to participants who read information and opted in (a subset of the forced reading, optional training condition). Error bars indicate standard error.

from the forced training condition ( $t_{407} = 1.02, p = 0.31$ ). These findings replicate the informed opt-in pattern observed in Experiment 1, demonstrating that information engagement produces robust behavioral changes.

The model found no difference in partner type ( $b = -0.15, SE = 0.08, p = 0.06$ ) or any other significant interactions ( $ps \geq 0.07$ ).

### C Information Formats for Experiment 2

#### C.1 Forced Training—Brief Condition

Participants in the forced training—brief condition saw the following text displayed on screen:

Your responses will be used to train an AI to propose offers.

This AI will learn by observing your responses.

**The AI you train will play against other Prolific participants in future experiments.**

#### C.2 Forced Training—Detailed Condition

Participants in the forced training—detailed condition first saw a brief message with a link:

Your responses will be used to train an AI that proposes offers.

[Click here for more information](#)

Clicking the link opened a popup window with the following text:

### Details on AI Training Participation and Data Usage

This document provides you with important information regarding the AI training component of this study and how your data will be used in this context. As part of this study, you will contribute to the training of an AI model designed to propose offers. Your responses and interactions will be used by the AI to learn and enhance its ability to generate strategies aligned with the patterns of human decision making.

Please note that the AI model trained on your data will be utilized in future studies involving other participants recruited through Prolific. Any data used to train the AI will be treated with strict confidentiality and will be used solely for the purposes outlined in this study.

Your privacy and data protection are our primary concerns, and all data handling will comply with applicable data protection standards. Data relevant to your interactions in this study will be utilized for AI training purposes, and these interactions will be stored and managed in accordance with applicable data protection standards.

The AI training component is an integral part of this research study. By continuing with the experiment, you consent to AI training.

Participants were required to click "I agree" to close the popup and proceed with the experiment.

### D Opponent Effects

The mixed-effects model revealed interactions involving opponent type, training condition, and offer amount. Specifically, the model identified a two-way interaction between opponent type and training condition when comparing the forced and uninformed conditions ( $b = 0.21$ ,  $SE = 0.10$ ,  $p = 0.045$ ), a two-way interaction between opponent type and training condition when comparing the uninformed and control conditions ( $b = -0.21$ ,  $SE = 0.11$ ,  $p = 0.045$ ), and a three-way interaction between opponent type, offer amount, and training condition when comparing the forced and control conditions ( $b = -0.15$ ,  $SE = 0.07$ ,  $p = 0.023$ ).

We first examined the two-way interactions by comparing differences in acceptance rates for AI versus human partners across training conditions. When comparing the forced and uninformed training conditions, participants in the forced condition showed greater sensitivity to opponent type than those in the uninformed condition ( $t_{254} = -2.17$ ,  $p = 0.031$ ). Specifically, while both groups accepted more offers by humans, participants in the forced condition showed a larger difference in acceptance rates between human and AI partners than did participants in the uninformed condition. In contrast, when comparing the control and uninformed conditions, there was no difference in sensitivity to opponent type ( $t_{265} = -1.38$ ,  $p = 0.17$ ).

We next examined the three-way interaction between training condition, opponent type, and offer amount by estimating the interaction between opponent type and offer amount separately within each training condition. Within the forced training condition, acceptance rates increased strongly with offer amount ( $b = 1.84$ ,  $SE = 0.07$ ,  $p < 0.001$ ), but there was no main effect of opponent type ( $b = -0.10$ ,  $SE = 0.06$ ,  $p = 0.10$ ) or significant interaction between opponent type and offer amount ( $b = 0.05$ ,  $SE = 0.04$ ,  $p = 0.27$ ). Thus, participants in the forced condition did not adjust their responses to AI versus human partners depending on the offer amount.

Within the control condition, we also found a main effect of offer amount ( $b = 1.95$ ,  $SE = 0.07$ ,  $p < 0.001$ ) and no main effect on opponent type ( $b = -0.11$ ,  $SE = 0.07$ ,  $p = 0.10$ ). In contrast to the forced condition, we observed a significant interaction between opponent type and offer amount ( $b = -0.10$ ,  $SE = 0.05$ ,  $p = 0.039$ ). The negative interaction coefficient indicates that as offers increased, participants' relative acceptance of AI

versus human offers decreased. Specifically, at lower offers, participants were more likely to accept offers from AI partners, whereas at higher offers, they were more likely to accept offers from human partners. This pattern indicates that participants in the control condition adjusted their responses to AI and human opponents differently depending on offer amount.

These opponent-specific effects are reported for completeness. Prior work using similar paradigms has found that differences between AI and human opponents, including interaction patterns with offer amounts, are inconsistent and do not reliably replicate across studies [70–72]. Accordingly, these results should be viewed as exploratory and informative for future research rather than as robust effects.

## E Regression Specifications

### E.1 Experiment 1

For Experiment 1, we manipulated the reference category to be the forced training condition, informed training condition, or uninformed training condition. Below is the entire equation when the reference level is the forced training condition.

$$\begin{aligned}
 P(\text{accept} = 1) = \text{logit}^{-1} & (\beta_0 + \beta_1 \times \text{offer} + \beta_2 \times \text{partner} + \beta_3 \times \text{control} + \beta_4 \times \text{informed training} \\
 & + \beta_5 \times \text{uninformed training} + \beta_6 \times \text{offer} \times \text{partner} + \beta_7 \times \text{offer} \times \text{control} \\
 & + \beta_8 \times \text{offer} \times \text{informed training} + \beta_9 \times \text{offer} \times \text{uninformed training} + \beta_{10} \times \text{partner} \times \text{control} \\
 & + \beta_{11} \times \text{partner} \times \text{informed training} + \beta_{12} \times \text{partner} \times \text{uninformed training} \\
 & + \beta_{13} \times \text{offer} \times \text{partner} \times \text{control} + \beta_{14} \times \text{offer} \times \text{partner} \times \text{informed training} \\
 & + \beta_{15} \times \text{offer} \times \text{partner} \times \text{uninformed training} + u_i)
 \end{aligned}$$

### E.2 Experiment 2

For Experiment 2, we manipulated the reference category to be the forced training-brief condition or the control condition. Below is the entire equation when the reference level is the forced training-brief condition.

$$\begin{aligned}
 P(\text{accept} = 1) = \text{logit}^{-1} & (\beta_0 + \beta_1 \times \text{offer} + \beta_2 \times \text{partner} + \beta_3 \times \text{forced training-detailed} + \beta_4 \times \text{control} \\
 & + \beta_5 \times \text{offer} \times \text{partner} + \beta_6 \times \text{offer} \times \text{forced training-detailed} + \beta_7 \times \text{offer} \times \text{control} \\
 & + \beta_8 \times \text{partner} \times \text{forced training-detailed} + \beta_9 \times \text{partner} \times \text{control} \\
 & + \beta_{10} \times \text{offer} \times \text{partner} \times \text{forced training-detailed} + \beta_{11} \times \text{offer} \times \text{partner} \times \text{control} + u_i)
 \end{aligned}$$

## F Mixed-Effects Regression Results

Table 3. Experiment 1. Reference Level: Forced training

Predictor	Estimate	Std. Error	z-value	p
Intercept	1.103	0.244	4.53	< .001***
Offer	1.879	0.064	29.21	< .001***
Opponent (AI)	-0.101	0.060	-1.68	.093
Control	1.037	0.343	3.03	.002**
Informed training	0.246	0.358	0.69	.493
Uninformed training	1.112	0.383	2.90	.004**
Offer × Opponent (AI)	0.051	0.045	1.13	.259
Offer × Control	0.040	0.091	0.44	.662
Offer × Informed training	0.083	0.095	0.88	.382
Offer × Uninformed training	0.198	0.107	1.85	.065
Opponent (AI) × Control	-0.005	0.089	-0.06	.951
Opponent (AI) × Informed training	0.064	0.090	0.71	.477
Opponent (AI) × Uninformed training	0.206	0.103	2.01	.045*
Offer × Opponent (AI) × Control	-0.148	0.065	-2.28	.023*
Offer × Opponent (AI) × Informed training	-0.034	0.069	-0.50	.618
Offer × Opponent (AI) × Uninformed training	-0.018	0.074	-0.24	.812

Note. Estimates are log-odds from a mixed-effects logistic regression with random intercepts for participant. \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

Table 4. Experiment 1. Reference level: Informed training

Predictor	Estimate	Std. Error	z-value	p
Intercept	1.349	0.263	5.12	< .001***
Offer	1.962	0.073	27.03	< .001***
Opponent (AI)	-0.037	0.067	-0.56	.575
Control	0.791	0.356	2.22	.026*
Forced training	-0.246	0.358	-0.69	.493
Uninformed training	0.866	0.395	2.19	.028*
Offer × Opponent (AI)	0.017	0.052	0.32	.747
Offer × Control	-0.044	0.096	-0.45	.652
Offer × Forced training	-0.083	0.095	-0.88	.382
Offer × Uninformed training	0.115	0.112	1.02	.306
Opponent (AI) × Control	-0.069	0.093	-0.74	.457
Opponent (AI) × Forced training	-0.064	0.090	-0.71	.477
Opponent (AI) × Uninformed training	0.142	0.107	1.33	.184
Offer × Opponent (AI) × Control	-0.113	0.070	-1.63	.104
Offer × Opponent (AI) × Forced training	0.034	0.069	0.50	.618
Offer × Opponent (AI) × Uninformed training	0.017	0.078	0.22	.829

Note. Estimates are log-odds from a mixed-effects logistic regression with random intercepts for participant. \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

Table 5. Experiment 1. Reference level: Uninformed training

Predictor	Estimate	Std. Error	z-value	p
Intercept	2.215	0.296	7.48	< .001***
Offer	2.077	0.089	23.44	< .001***
Opponent (AI)	0.104	0.083	1.26	.209
Control	-0.075	0.381	-0.20	.843
Forced training	-1.112	0.383	-2.91	.004**
Informed training	-0.866	0.395	-2.19	.028*
Offer × Opponent (AI)	0.034	0.058	0.58	.564
Offer × Control	-0.158	0.108	-1.46	.144
Offer × Forced training	-0.198	0.107	-1.85	.065
Offer × Informed training	-0.115	0.112	-1.02	.306
Opponent (AI) × Control	-0.211	0.105	-2.00	.045*
Opponent (AI) × Forced training	-0.206	0.103	-2.01	.045*
Opponent (AI) × Informed training	-0.142	0.107	-1.33	.184
Offer × Opponent (AI) × Control	-0.130	0.074	-1.75	.081
Offer × Opponent (AI) × Forced training	0.018	0.074	0.24	.812
Offer × Opponent (AI) × Informed training	-0.017	0.078	-0.22	.829

Note. Estimates are log-odds from a mixed-effects logistic regression with random intercepts for participant. \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

Table 6. Experiment 2. Reference level: control

Predictor	Estimate	Std. Error	z-value	p
Intercept	2.487	0.308	8.08	< .001***
Offer	2.363	0.093	25.50	< .001***
Opponent (AI)	-0.112	0.074	-1.52	.130
Forced training-detailed	-0.963	0.423	-2.28	.023*
Forced training-brief	-0.778	0.432	-1.80	.072
Offer × Opponent (AI)	-0.057	0.056	-1.02	.308
Offer × Forced training-detailed	-0.089	0.119	-0.75	.455
Offer × Forced training-brief	0.077	0.125	0.62	.539
Opponent (AI) × Forced training-detailed	0.085	0.099	0.86	.391
Opponent (AI) × Forced training-brief	0.136	0.103	1.31	.189
Offer × Opponent (AI) × Forced training-detailed	-0.001	0.077	-0.02	.986
Offer × Opponent (AI) × Forced training-brief	0.059	0.082	0.73	.466

Note. Estimates are log-odds from a mixed-effects logistic regression with random intercepts for participant. \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

Table 7. Experiment 2. Reference level: Forced training-brief

Predictor	Estimate	Std. Error	z-value	p
Intercept	1.710	0.306	5.58	< .001***
Offer	2.440	0.092	26.43	< .001***
Opponent (AI)	0.024	0.073	0.33	.739
Control	0.778	0.432	1.80	.072
Forced training-detailed	-0.185	0.424	-0.44	.662
Offer × Opponent (AI)	0.002	0.059	0.03	.974
Offer × Control	-0.077	0.125	-0.62	.539
Offer × Forced training-detailed	-0.166	0.120	-1.39	.166
Opponent (AI) × Control	-0.136	0.103	-1.31	.189
Opponent (AI) × Forced training-detailed	-0.051	0.098	-0.52	.603
Offer × Opponent (AI) × Control	-0.059	0.082	-0.73	.466
Offer × Opponent (AI) × Forced training-detailed	-0.061	0.079	-0.77	.442

Note. Estimates are log-odds from a mixed-effects logistic regression with random intercepts for participant. \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .