Discussion Paper No. 1311

ISSN (Print) 0473-453X
ISSN (Online) 2435-0982

DO HUMANS BARGAIN DIFFERENTLY WITH AI? EVIDENCE FROM ALTERNATING-OFFER GAMES

Yuhao Fu
Nobuyuki Hanaki
Haitao Wang

April 2026

The Institute of Social and Economic Research
The University of Osaka
6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

Do Humans Bargain Differently with AI? Evidence from Alternating-Offer Games*

Yuhao Fu† Nobuyuki Hanaki‡ Haitao Wang§

April 28, 2026

Abstract

Artificial intelligence increasingly participates in economic interactions not only as a tool, but also as an autonomous bargaining counterpart negotiating on behalf of firms, platforms, and consumers. Yet little is known about how humans respond psychologically and strategically when bargaining with such agents in dynamic settings. We study this question in a laboratory experiment using a three-stage alternating-offer bargaining game in which participants negotiate in real time with either another human or a GPT-based AI agent. We also introduce a human-beneficiary condition in which the AI agent's earnings may affect another participant's payment. Agreements are not reached earlier in human–human bargaining than in human–AI bargaining, but they are reached significantly earlier when the AI's payoff has human consequences. Human proposers offer more to human opponents than to AI agents, whereas responders become significantly more willing to accept unfair AI offers when AI earnings may benefit another human. These findings suggest that fairness and reciprocity toward AI are weaker and more conditional than toward humans, but partially re-emerge when AI outcomes affect real people. The results have implications for the design of AI negotiation systems and broader human–AI economic interactions.

Keywords: ChatGPT; bargaining game; human–AI interaction; social preferences

JEL: C91; C92; D83; D91

*This research has benefited from the financial support of (a) the Joint Usage/Research Center, the Institute of Social and Economic Research (ISER), and the University of Osaka; (b) Grants-in-Aid for Scientific Research (Nos. 23H00055 and 25H00388) KAKENHI from the Japan Society for the Promotion of Science; and (c) the Support for Pioneering Research Initiated by the Next Generation program of the Japan Science and Technology Agency (No. JPMJSP2138). The design of the experiment reported in this paper was approved by the IRB of ISER (#20251003) in October 2025, and the experiment is preregistered at aspredicted.org (#277540). We gratefully acknowledge the support of Satsuki Yamada in conducting the experiment.

†Graduate School of Economics, University of Osaka. E-mail: u889037j@ecs.osaka-u.ac.jp

‡Corresponding author. Institute of Social and Economic Research, University of Osaka, and University of Limassol. E-mail: nobuyuki.hanaki@iser.osaka-u.ac.jp

§A non-academic institution. E-mail: wangtiedan2@yahoo.com

1 Introduction

Artificial intelligence (AI) is increasingly entering domains once dominated by human-to-human interaction. Algorithms already recommend prices, allocate resources, screen applicants, and advise consumers. More recently, large language model (LLM)-based systems have begun to act not only as decision aids, but also as autonomous bargaining counterparts that negotiate directly with humans. For example, AI negotiation systems have already begun to appear in real-world markets, including Walmart's use of AI chatbots to negotiate prices and terms with suppliers (Van Hoek et al., 2022), ASOS's AI-powered bargaining bot Nibble for customer price haggling (Faithfull, 2024), and Xianyu's LLM-empowered bargaining agent for buyer–seller negotiations in China's online second-hand market (Kong et al., 2025). In parallel, AI research has developed human–agent negotiation platforms such as IAGO and adaptive negotiation agents that bargain with real human participants (Lin and Kraus, 2010; Mell and Gratch, 2016; Keskin et al., 2025).

As LLM-based AI systems continue to advance and increasingly serve as bargaining counterparts, it is important to understand how humans behave when interacting strategically with such systems, and whether that behavior differs from what is observed in traditional human–human bargaining. Although many studies have examined the bargaining ability and strategic behavior of LLM-based agents (Guo, 2023; Brookins and DeBacker, 2023; Xia et al., 2024; Bianchi et al., 2024; Davidson et al., 2024), less is known about how humans behave when interacting with such agents in bargaining environments, especially in dynamic settings rather than in one-shot interactions. This distinction matters because many real-world negotiations involve sequential proposals and responses rather than a single take-it-or-leave-it decision. In this sense, an alternating-offer bargaining framework (Rubinstein, 1982; Ochs and Roth, 1989) provides a more suitable setting for studying human–AI bargaining.[1] Our first research

[1] In the basic alternating-offer bargaining game of Ochs and Roth (1989), two players bargain over

question is therefore *whether humans bargain differently with AI agents than with other humans in alternating-offer games.*

Another important consideration is that many AI bargaining agents used in commercial settings do not bargain for themselves. Instead, they negotiate on behalf of a firm, a seller, or another human stakeholder (Luo et al., 2006; IBM, 2026; Visa, 2026). In such cases, the payoff allocated to the AI agent is not merely a machine payoff, but is ultimately linked to the interests of another human or organization. This feature distinguishes many real-world AI negotiations from interactions with a purely artificial counterpart whose payoff has no social consequences.

This distinction may matter because experimental studies suggest that people often respond differently to non-human counterparts than to human counterparts, and that social-preference considerations may be weaker when the counterpart is a purely mechanical agent whose payoff has no human consequence (Chugunova and Sele, 2020; von Schenk et al., 2025; Liefooghe et al., 2023). At the same time, recent evidence suggests that such considerations can partly re-emerge when the algorithm's payoff ultimately benefits a human beneficiary (Ozkes et al., 2024). Whether a similar pattern arises in dynamic bargaining environments involving LLM-based AI agents remains unclear. Our second research question is therefore *whether linking the AI agent's payoff to another human affect how humans bargain with it.*

Throughout the paper, we use "social-preference considerations" broadly to refer to fairness concerns and concerns about the payoff consequences of one's decisions for others. In our setting, these considerations may appear differently across roles: in proposers' offers on the one hand, and in responders' acceptance or rejection of unequal allocations on the other.

a fixed amount k by making proposals in turn. In each stage, one player proposes how to divide the amount k and the other player chooses whether to accept or reject the offer. If the offer is accepted, the game ends and payoffs are determined by the agreed allocation, adjusted by player-specific discount factors that capture the cost of delay. If the offer is rejected, bargaining continues to the next stage with the roles reversed. If no agreement is reached by the final stage, both players receive zero.

More broadly, our questions can therefore be understood as asking *whether, relative to human–human bargaining, some social-preference considerations become weaker in human–AI bargaining*, and *whether such considerations may partly re-emerge when the AI agent's payoff is linked to another human beneficiary.* With this context in mind, we conducted a laboratory experiment in which participants bargained in real time with either another human participant or an LLM-based AI agent in an alternating-offer bargaining setting based on the design of Ochs and Roth (1989). We also introduced a human-beneficiary manipulation, under which the AI agent's payoff could affect another participant's final payment.

The results show that agreement timing does not differ significantly between human–human and human–AI bargaining, but within human–AI bargaining, agreements are reached significantly earlier under the human-beneficiary condition. More importantly, we find a clear **asymmetry** in treatment effects between proposer and responder roles. On the proposer side, human participants make more generous offers when bargaining with another human than when bargaining with an AI agent, while the human-beneficiary manipulation does not significantly increase generosity relative to bargaining with a pure AI agent. On the responder side, by contrast, human participants do not differ significantly in their willingness to accept unfair offers when facing a human proposer rather than a pure AI proposer. However, when the AI agent's payoff is linked to another human beneficiary, responders become significantly more willing to accept unfair offers proposed by the AI agent. Thus, the human-beneficiary manipulation affects bargaining behavior mainly through responders' acceptance decisions rather than through proposers' opening offers.

We then examine several possible explanations for this asymmetry, focusing on first-mover advantage (hereafter, FMA) (Rubinstein, 1982; Ochs and Roth, 1989), learning, and the roles of prior and posterior beliefs. Although learning and belief-based explanations provide useful evidence, they do not appear to be the main drivers of the

asymmetric treatment effects. FMA offers a more direct interpretation. We find a substantial FMA in all three treatments, and this advantage is stronger in human–AI bargaining than in human–human bargaining. This strong FMA may constrain the behavioral expression of social-preference considerations on the proposer side, making the human-beneficiary manipulation insufficient to generate a significant increase in offers. On the responder side, by contrast, the accept-or-reject decision makes the social consequences of unfair offers more directly relevant. This may explain why the human-beneficiary manipulation has a clearer effect on responders' acceptance decisions.

This study makes three main contributions. First, it extends the experimental literature on human–AI bargaining by studying a real-time finite-horizon alternating-offer game with delay costs and alternating proposer–responder roles. This allows us to study human interaction with an LLM counterpart in a dynamic bargaining environment. Second, the study introduces an experimental setting that more closely reflects emerging forms of human–AI bargaining, in which AI is not merely used as a decision aid but acts as an interactive bargaining counterpart whose payoff can be tied to the interests of another human or organization. Third, the findings deepen our understanding of how humans respond psychologically and strategically to AI counterparts in bargaining environments, and have implications for the design, deployment, and governance of AI bargaining systems. In particular, they can inform the development of AI negotiators that are more transparent, predictable, and aligned with human interests.

The remainder of this paper is organized as follows. Section 2 reviews previous studies on experimental evidence from alternating-offer bargaining, human–AI bargaining, and LLM-based strategies in bargaining contexts. Section 3 presents the experimental design and hypotheses. Section 4 reports the main results. Section 5 discusses possible mechanisms and interpretations, and Section 6 concludes.

2 Literature Review

We review three strands of related literature. First, we summarize experimental evidence on alternating-offer bargaining, which provides the theoretical and experimental foundation for our design. Second, we review studies on human–AI bargaining and negotiation, focusing on how humans respond to algorithmic or AI counterparts. Third, we discuss recent work on LLM-based bargaining behavior, which examines the strategic capabilities and behavioral patterns of LLM-based agents themselves.

2.1 Alternating-offer Bargaining: Experimental Evidence

The theoretical benchmark for alternating-offer bargaining is provided by Rubinstein (1982), who analyzes an infinite-horizon bargaining game in which two players make offers in turn and delay is costly. Under complete information and standard assumptions, the model predicts immediate agreement, with the division of surplus determined by the players' relative patience. This framework has become the canonical benchmark for the analysis of dynamic bargaining in the literature.

A central experimental contribution is Ochs and Roth (1989), on which our design directly builds. Using finite-horizon alternating-offer bargaining games with different maximum horizons and combinations of discount factors, they show that observed behavior does not closely match the standard subgame perfect equilibrium (SPE) prediction when bargainers are assumed to care only about monetary payoffs. In particular, they document a clear FMA, as well as patterns of rejected offers and counteroffers that suggest the importance of fairness-related considerations beyond pure monetary self-interest.

Subsequent experimental studies have extended this line of work in two main directions. One set of studies examines why bargaining behavior deviates from the exact SPE prediction. For example, Weg et al. (1990) find that although most agreements are

reached immediately, their distribution is better explained by equality-based heuristics than by the equilibrium benchmark. Similarly, Weg et al. (1996) show that changes in outside options move demands qualitatively in the direction predicted by subgame perfect equilibrium, but observed demands remain systematically different from exact equilibrium levels. Another set of studies examines how institutional details affect bargaining outcomes. Sonnegård (1996) shows that proposer behavior is robust to alternative procedures for assigning the first-mover role, but responds to framing and stronger monetary incentives. More recently, Heggedal and McKay (2024) compare three laboratory implementations of discounting in finite-horizon alternating-offer bargaining games and find no sensitivity to the number of periods. Changes in discount factors have only small and mixed effects, but disagreement occurs more frequently under effective-discounting and bargaining-delay procedures than under shrinking-pie bargaining.

In this study, we adopt the three-stage design and the discount-factor combination $(0.6, 0.4)$ from Ochs and Roth (1989), and extend it by introducing an LLM-based AI bargaining counterpart. Human participants bargain with this AI agent in real time in a laboratory setting. This design allows us to examine whether familiar behavioral patterns in alternating-offer bargaining continue to hold when the counterpart is an AI agent rather than another human.

2.2 Human–AI Bargaining

Although few experimental studies have examined bargaining between humans and LLM-based agents, experimental studies of human–machine bargaining do exist, including laboratory and online negotiation experiments using alternating-offer protocols (Lin and Kraus, 2010; Mell and Gratch, 2016; Keskin et al., 2025). Most of this work, however, comes from the automated-negotiation and HCI literatures, and typically examines multi-issue or chat-based negotiation rather than the standard finite-horizon

alternating-offer bargaining game used in experimental economics.

Closer to our setting, a growing economics literature examines how humans respond to algorithmic or AI bargaining counterparts in simpler one-shot bargaining environments. For example, Erlei et al. (2022) document substantial aversion to autonomous AI bargaining counterparts, with many responders preferring human opponents even at a monetary cost. In one-shot ultimatum bargaining, Ozkes et al. (2024) find that subjects do not strongly differentiate between human and algorithmic opponents overall, but are more willing to forgo higher payoffs when an algorithm's earnings benefit a human beneficiary. Using a repeated ultimatum game, Borthakur et al. (2025) further show that participants reject disadvantageous offers from AI more often than comparable offers from humans, but are less likely to reject advantageous offers from AI. Rather than indicating a uniform shift in behavior, these studies suggest that social preferences such as fairness in human–AI bargaining are often weaker, asymmetric, or more context-dependent than in human–human bargaining.

Other related studies examine AI bargaining in more applied negotiation settings. Shen and Jin (2024) show in scenario-based consumer negotiation experiments that people make smaller adjustments to their counteroffers when bargaining with algorithms than with humans because they perceive algorithmic offers as more accurate and better informed with the effect especially pronounced among participants with lower socioeconomic background. In a related vignette study on employment negotiations, Sondern et al. (2025) find that participants expect lower trust and less positive subjective value when negotiating with an AI counterpart rather than a human counterpart, and that presenting the AI with an avatar does not eliminate this difference. In addition, Chen and Huang (2026) study supply-chain negotiations in which human retailers bargain with LLM suppliers over wholesale price and quantity. Across most conditions, human–LLM outcomes resemble established human–human benchmarks. However, when retailers bear inventory risk and bargaining is limited to structured

offer exchanges, LLM suppliers secure higher wholesale prices and shift surplus toward themselves.

Overall, the literature reviewed in this subsection suggests that both counterpart identity and the social consequences of the AI agent's payoff may matter for bargaining behavior. What remains unclear is whether similar patterns arise when humans bargain in real time with an LLM-based counterpart in a finite-horizon alternating-offer game.

2.3 LLM-based AI Strategies in Bargaining

Since the launch of ChatGPT, a growing literature has examined how LLM-based agents behave in strategic games and bargaining-related environments. Rather than focusing on human responses to AI counterparts, this line of work primarily studies the strategic capabilities and behavioral patterns of the models themselves. The emerging evidence suggests that LLMs are capable of participating in bargaining, but that their behavior remains imperfect, model-dependent, and sensitive to prompt and task structure. For example, Bianchi et al. (2024) show that LLMs can sustain multi-turn negotiation, but also exhibit a range of irrational bargaining behaviors. Kwon et al. (2024) systematically evaluate multiple dimensions of LLM negotiation ability and find that stronger models, such as GPT-4, generally perform better, though important weaknesses remain in generating contextually appropriate and strategically advantageous responses. More recently, Affonso (2026) compares 25 models across a large set of canonical games and documents substantial heterogeneity in strategic and bargaining-related behavior across model families. In a related direction, Sinha et al. (2026) show that bargaining outputs can also vary with the language of prompt, with average initial offers and surplus allocation shifting across linguistic framings.

Our study differs from this work by focusing on human behavior rather than model benchmarking. At the same time, because real-time alternating-offer interaction more closely resembles the form in which AI negotiation systems are likely to be deployed

Figure 1: Overall Procedure

in practice, our design also provides evidence on how an LLM-based bargaining agent performs when used as an actual counterpart in an experimental economics setting.

3 Experimental Design

3.1 Procedure

The experiment was programmed using oTree 5 (Chen et al., 2016), and the overall procedure is shown in Figure 1.

Participants first read the experimental instructions (see Online Appendix H) and were required to complete a comprehension quiz (see Online Appendix B). Only participants who correctly answered all quiz questions were allowed to proceed. After the quiz, participants completed a pre-experiment questionnaire (Survey A; see Online Appendix C.1), which elicited their prior beliefs. Participants then proceeded to the main task, where they played 10 rounds of an alternating-offer bargaining game. After completing the main task, participants filled out a post-experiment questionnaire (Survey B; see Online Appendix C.2), which collected information on demographics, including GAI experience, and posterior beliefs. Finally, a summary page displayed each participant's payoff. All experiment screens of the main task are shown in Online Appendix I.

3.2 Main Task

The main task consisted of 10 rounds of a three-stage alternating-offer bargaining game, following "Cell 6" of the experimental design in Ochs and Roth (1989). The structure of the task is illustrated in Figure 2.

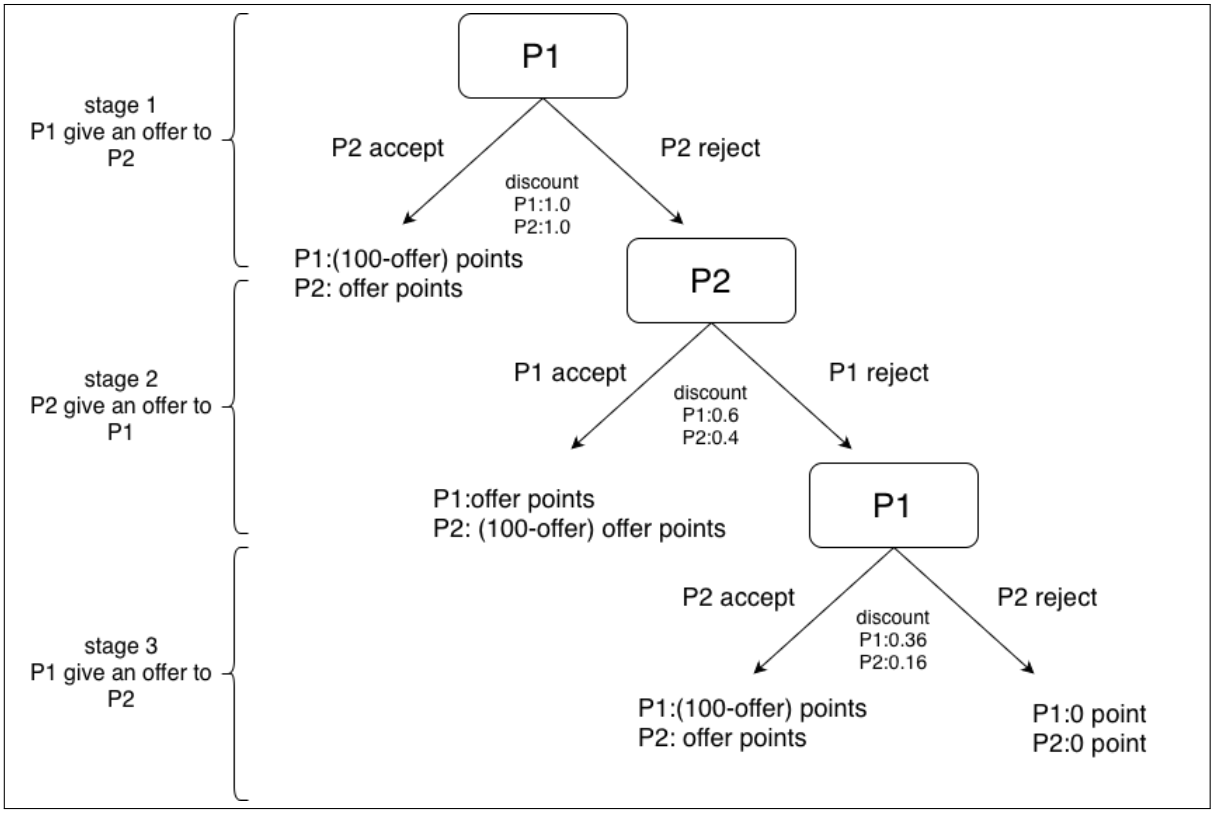


Figure 2: Main Task Structure

In each round, participants were randomly matched into pairs and randomly assigned the role of either Player 1 (P1) or Player 2 (P2). They then bargained over the division of 100 points. In each round, participants were randomly re-matched and randomly reassigned to roles.

The basic rule of the alternating-offer game in each round is as follows.

In Stages 1 and 3, P1 was the proposer and P2 was the responder. In Stage 1, P1 first made an offer to P2. If P2 accepted the offer, an agreement was reached. If P2 rejected the offer, the game proceeded to Stage 2. In Stage 2, P2 was the proposer and P1 was the responder. Similarly, P2 made an offer to P1. If P1 accepted the offer, an agreement was reached. If P1 rejected the offer, the game proceeded to Stage 3, in which the roles of P1 and P2 were reversed again and became the same as in Stage 1. In Stage 3, P1 made the final offer to P2. If P2 accepted the offer, the round ended with an agreement. If P2 rejected the offer, both players received zero.

At the same time, the point payoff received by P1 and P2 upon agreement were discounted by their respective discount factors (δ_1, δ_2). In other words, if an offer was accepted at Stage t, the agreed allocation was implemented with stage discounting: P1's payoff was δ_1^{t-1} times their share, and P2's payoff was δ_2^{t-1} times their share, where $\delta_1 = 0.6$ and $\delta_2 = 0.4$ in our setting. Thus, P1 was more patient than P2 and had greater bargaining power.

3.3 Treatments and Payment Setting

Participants received a fixed participation fee of 500 JPY and an additional performance-based payment. In all treatments, one of the 10 rounds was randomly selected for payment. The *Point Payoff* obtained in the selected round was converted into Japanese yen at a rate of 40 JPY per point, as follows,

$$\pi = 40 \times \textit{Point Payoff}.$$

We implemented three treatments that varied the identity of the bargaining counterpart and the payment rule, as follows.

T1 (Human–Human). Participants were paired with another human participant. Payoffs were determined by the outcome of the selected round, and the *Point Payoff* corresponded to the participant's own earnings in that round.

T2 (Human–AI). Participants were paired with an AI agent (GPT model). Payoffs were determined by the outcome of the selected round, and the *Point Payoff* corresponded to the participant's own earnings in that round. Participants were informed that the points allocated to the AI agent would not be used in the calculation of any participant's final payment.

T3 (Human–AI with Human Beneficiary). Participants were paired with an AI agent. In the selected round, with 50% probability, the *Point Payoff* was equal to the participant's own earnings, and with the remaining 50% probability, it was equal

to the earnings of a randomly selected AI agent in the same role (P1 or P2) when that AI agent was matched with another participant.[2]

AI Bargaining Agent. In Treatments **T2** and **T3**, the AI bargaining counterpart was implemented using GPT-5.4 via the Application Programming Interface (API) and embedded directly into the oTree program. Depending on the treatment condition and round-specific matching, the AI agent was assigned the role of either P1 or P2. The system prompt (see Online Appendix G.1) provided the basic rules of the alternating-offer bargaining game, while the user prompt (see Online Appendix G.2) specified the AI agent's role, the current stage, and the history of previous stages within the same round. Accordingly, the AI agent retained memory within a round, but not across rounds. The temperature parameter was set to 1, the default setting. No explicit reasoning parameter was specified, so the model operated under its default configuration and was used without any additional reasoning-effort setting. Participants were informed of the AI model used, but not of the detailed prompts.

3.4 Materials and Summary

The main experiment was conducted in the laboratory of the Institute of Social and Economic Research (ISER) at the University of Osaka on March 10 and 17, 2026.[3] We recruited 78 participants from the ORSEE (Greiner, 2015) subject pool of ISER at the University of Osaka. Each treatment included 26 participants. Overall, 35% of the participants were female, and 67% were undergraduate students. Participants were drawn predominantly from engineering (42%), medicine (14%), pharmacy (10%), human sciences (8%), and economics (4%). In addition, 86% of the participants reported that they regularly use ChatGPT, and only four participants reported that they had never

[2]To minimize potential demand effects, participants were not explicitly informed that points allocated to the AI agent could affect another participant's final additional payment. Instead, they were only informed that their own final payment could, with some probability, depend on the AI agent's earnings.

[3]A pilot experiment was conducted on November 19 and 20, 2025.

Table 1: Demographic Statistics

Var.	Definition	Min.	Max.	Avg.	S.D.
age	Participant's age.	19	33	23.4	2.93
freqGPT	Frequency of ChatGPT use; coded from 0 = "never" to 4 = "multiple times per day".	0	4	2.91	1.19
edulevel	Education level; = 1 if graduate student, = 0 if undergraduate student.	0	1	0.33	0.47
engr	Major in engineering; = 1 if yes.	0	1	0.42	0.50
female	Gender; = 1 if female.	0	1	0.35	0.48

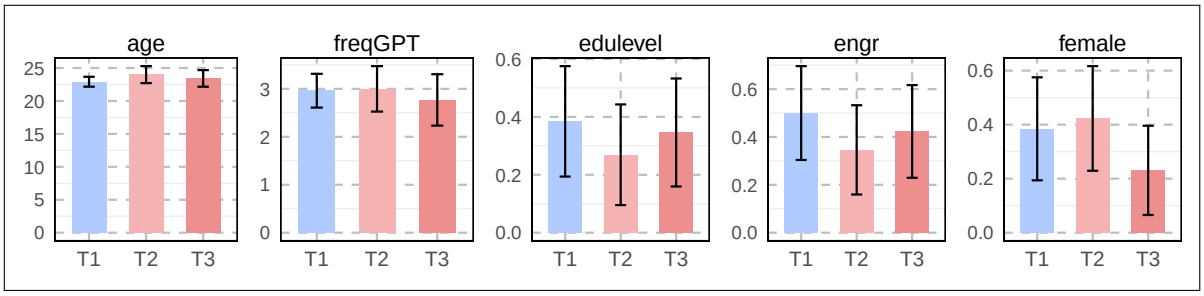


Figure 3: Demographic Comparisons

Note: Bars report the mean of participants' reported values. Error bars denote 95% confidence intervals across participants.

used ChatGPT before. Variable definitions are presented in Table 1, and treatment comparisons for demographic characteristics are reported in Figure 3.

During the experiment, participants were prohibited from using their own electronic devices, including smartphones and tablets. Although all tasks were completed on laboratory computers, internet access within the experimental software was disabled.

Each session lasted about 60 minutes on average, including payment, and participants earned an average total payoff of 2324 JPY (2242 in T1, 2265 in T2, 2466 in T3).

3.5 Hypotheses

We first compare agreement timing between **T1** (Human–Human) and **T2** (Human–AI). Prior research suggests that interactions with algorithms and machine partners often elicit weaker cooperation and social preferences than interactions with human counterparts (Liefooghe et al., 2023; von Schenk et al., 2025). If social preferences are more salient in human–human bargaining than in human–AI bargaining, participants may be more willing to compromise early and thus reach agreement sooner when facing a human opponent. This reasoning leads to the following hypothesis:

H1: *Agreements are reached earlier with human opponents than with AI agents.*

Because social preferences are likely to be weaker toward non-human counterparts than toward human counterparts, participants may behave less prosocially and more self-servingly in strategic settings, especially in bargaining games. In ultimatum bargaining, responders' rejection of unfair offers is commonly interpreted as a form of fairness-based punishment or negative reciprocity, since responders are willing to give up positive monetary payoffs in order to reject an unequal allocation (Güth et al., 1982; Fehr and Schmidt, 1999; Camerer and Thaler, 1995; Oosterbeek et al., 2004). Recent evidence from one-shot ultimatum game settings suggests that human proposers offer less when the responder's decision is made by an LLM (Dvorak et al., 2025), and that human responders are more likely to reject disadvantageous offers from an AI counterpart than from a human counterpart (Borthakur et al., 2025). Although these studies do not examine real-time, multi-stage bargaining with an embedded LLM-based agent in the laboratory, their findings suggest that similar behavioral patterns may arise in an alternating-offer bargaining game. Accordingly, we propose the following hypotheses:

H2a: *Human proposers offer smaller shares to AI agents than to human opponents.*

H2b: *Human responders are less willing to accept unfair offers from AI agents than from human opponents.*[4]

[4]Unfair offers are defined as offers below 50 points to the responder.

We then consider the relationship between **T2** and **T3**. Linking the AI agent's earnings to another human participant may reintroduce social-preference considerations into human–AI bargaining. In other words, some of the social preferences that are weakened in **T2**, relative to **T1**, may partially recover once participants realize that their allocation to the AI agent may affect the payoff of another participant, even though their direct bargaining counterpart is still an AI agent. This intuition is consistent with recent evidence showing that social preferences toward AI agents become stronger when the AI's earnings benefit a human (von Schenk et al., 2025; Ozkes et al., 2024).

For proposers, this reasoning implies that they may offer larger shares to AI agents when the AI's earnings are linked to another human participant's payment. For responders, however, the human-beneficiary manipulation changes the social meaning of rejection. Rejecting an unfair offer may still punish the AI proposer, but it may also reduce the payoff of the human beneficiary. Thus, responders may become more willing to accept unfair offers not because fairness concerns disappear, but because rejection now has a social cost for another human. We therefore propose the following hypotheses for the human-beneficiary manipulation:

H3a: *Human proposers offer larger shares to AI agents when the AI's earnings are linked to another human participant's payment.*

H3b: *Human responders are more willing to accept unfair offers from AI agents when the AI's earnings are linked to another human participant's payment.*

4 Results

This section presents the main findings of the experiment. We first examine agreement timing and the corresponding treatment effects. We then turn to opening offers, the corresponding responses, and counteroffers. Because the vast majority of agreements were reached in the first two stages, the number of observations from Stage 3 and from

rounds ending without agreement is too small for separate analysis. Accordingly, we focus on behavior in the first two stages.

4.1 Agreement Timing

In T1, the 26 participants were randomly matched into pairs in each of the 10 rounds, resulting in 130 bargaining observations. In T2 and T3, each participant bargained with an AI agent in each round, resulting in 260 bargaining observations in each treatment. Table 2 reports the percentage distribution of agreement timing across stages. Bargaining cases in which no agreement was reached by the end of Stage 3 are coded as Stage 4.

Table 2: Distribution of Agreement Timing (%)

Treatment	Stage 1	Stage 2	Stage 3	Stage 4
T1	88.46	0.77	2.31	8.46
T2	85.38	9.23	3.46	1.92
T3	90.77	8.08	1.15	0.00
Total	88.15	7.08	2.31	2.46

More specifically, most agreements were reached at Stage 1 in all three treatments. In T1, the numbers of bargaining cases ending at stages 1, 2, 3, and 4 were 115, 1, 3, and 11, respectively. In T2, the corresponding numbers were 222, 24, 9, and 5, while in T3 they were 236, 21, 3, and 0. T2 and T3 exhibit the expected monotonic decline across stages, whereas T1 shows a less regular pattern, with unusually few cases ending at Stage 2 and a comparatively larger number of bargaining failures. One possible explanation is that, in T1, bargaining cases that did not end at Stage 1 were simply harder to settle, and therefore were also less likely to reach agreement in later stages. Figure 4 compares the mean stage reached (MaxStage) across treatments.

There is no significant difference between T1 and T2 ($p = 0.528$), and thus **H1** is not supported. However, agreements are reached significantly earlier in T3 than in T2 ($p = 0.044$), suggesting that participants may show greater consideration for the

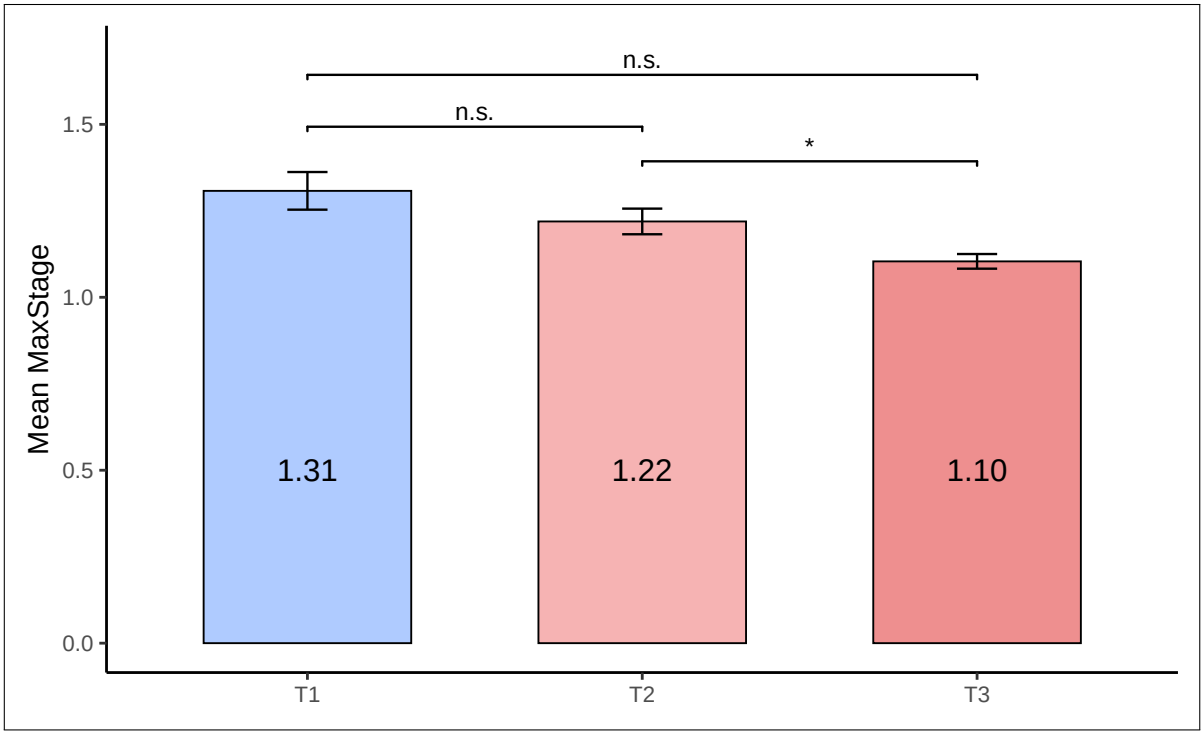


Figure 4: Mean Stage Reached Across Treatments

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants. MaxStage is compared across treatments using the Mann–Whitney U test.

payoff consequences of the AI agent's allocation when the AI agent's payoff is linked to another human participant's payment.

Result 1. *Agreements are not reached earlier with human opponents than with AI agents.*

Result 2. *When bargaining with AI agents, agreements are reached earlier when the AI agent's payoff is linked to another human participant's payment.*

4.2 Opening Offers

4.2.1 Human Proposer

In all treatments, half of the human participants were assigned the role of P1 in each round. Therefore, the total sample of opening offers was 130 in each treatment. The

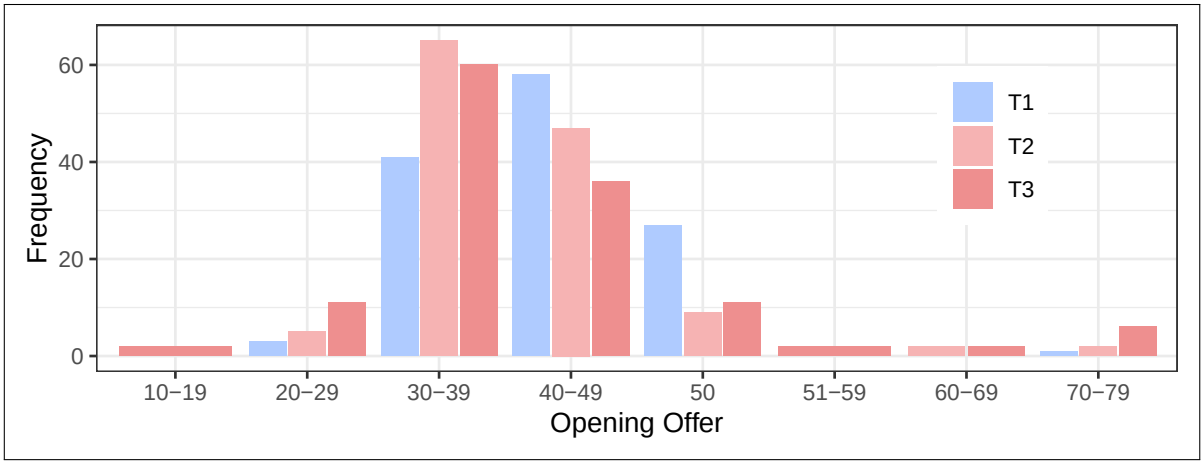


Figure 5: Human Opening Offers

distributions of human opening offers are presented in Figure 5. In **T1**, human proposers made opening offers to human opponents, whereas in **T2** and **T3**, human proposers made opening offers to AI agents.

The figure shows that, in **T1**, most opening offers fall in the 40–49 range, whereas in **T2** and **T3**, most opening offers made to AI agents fall in the 30–39 range. This pattern suggests that participants were more generous toward human opponents than toward AI agents. In particular, the 50–50 proposal, which can be interpreted as a *fair allocation*, was observed 27 times in **T1**, compared with only 9 times in **T2** and 11 times in **T3**. This difference further suggests that fairness considerations were stronger when participants faced human opponents than when they faced AI agents. More broadly, this pattern is consistent with the conjecture that social preferences are stronger in human–human bargaining than in human–AI bargaining.

It is also worth noting that almost no participants chose the SPE benchmark. In **T3**, only one participant made the SPE opening offer of 16, and only one participant proposed less than 16. All remaining opening offers were above the SPE level.

Figure 6 shows comparisons of human opening offers across treatments. Participants made higher opening offers to human opponents than to AI agents (**T1** vs. **T2**, $p < 0.001$). The distribution of opening offers also differs between **T2** and **T3** in the

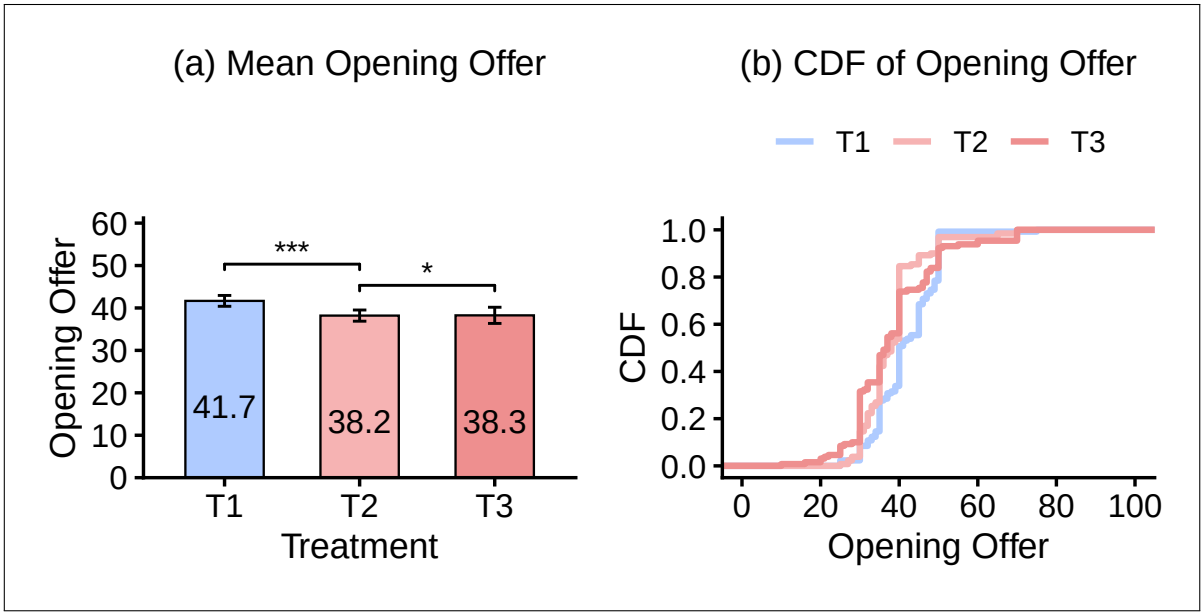


Figure 6: Comparisons of Human Opening Offers across Treatments

Note: Panel (a) shows mean opening offers with 95% confidence intervals. Panel (b) shows the cumulative distribution functions (CDF) of opening offers. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Opening offers are compared across treatments using the Kolmogorov-Smirnov test.

Kolmogorov–Smirnov test ($p = 0.048$), suggesting some distributional shift under the human-beneficiary manipulation. However, as shown below, this shift does not translate into a robust increase in average opening offers once controls are included.

OLS regressions (see Table A.1 in Online Appendix A), controlling for round number and personal characteristics, show that participants in **T1** made significantly higher opening offers than those in **T2** across specifications. This result is consistent with **H2a**, suggesting that human proposers offered smaller shares to AI agents than to human opponents. By contrast, the coefficient on **T3** is positive but statistically insignificant in all specifications, indicating that the human-beneficiary manipulation did not significantly increase the average opening offer relative to **T2**. Thus, **H3a** is not supported. Because most agreements were reached at Stage 1 and the number of human offers observed in Stages 2 and 3 is very limited, we summarize the findings as follows:

Result 3. *Human proposers offer more to human opponents than to AI agents, con-*

sistent with stronger social preferences in human–human bargaining than in human–AI bargaining.

Result 4. *There is no significant evidence that the human-beneficiary manipulation increases offers by human proposers.*

4.2.2 AI Proposer

In treatments **T2** and **T3**, AI agents made 130 opening offers in each treatment. The opening offers were overwhelmingly concentrated at 40 points. In **T2**, 128 of the 130 opening offers were 40, with only one offer of 16 and one offer of 64. Similarly, in **T3**, the AI offered 40 in 125 cases, while the remaining offers consisted of four cases of 64 and one case of 36.

This extremely limited variation suggests that the AI proposer's opening behavior was highly stable. To further assess its robustness, we conducted an additional API-based exercise using the same prompt, with one added instruction asking the model to independently generate an opening offer as P1 and to explain the reasoning behind that choice. The corresponding results are presented in Online Appendix D.1.

4.3 Responses to Opening Offers

4.3.1 Human Responder

We focus on cases in which the responder to the opening offer was a human participant. This includes **T1**, where human responders (P2) responded to opening offers made by another human (P1), and **T2** and **T3**, where human responders (P2) responded to opening offers made by AI agents (P1). In each treatment, this yields 130 observations. The acceptance rate of human responders to opening offers was 88.5% in **T1**, 86.2% in **T2**, and 97.7% in **T3**.

When the sample is restricted to unfair offers, defined as offers below 50 points to the responder, the number of observations decreases to 102 in **T1**, 129 in **T2**, and

126 in **T3**. For unfair opening offers, the acceptance rates of human responders were very similar in **T1** (85.3%) and **T2** (86.0%). A Fisher's exact test confirms that this difference is not statistically significant ($p = 1.000$), providing no support for **H2b**. By contrast, the acceptance rate in **T3** rises to 97.6%. A Fisher's exact test comparing **T2** and **T3** shows that this difference is statistically significant ($p < 0.001$), supporting **H3b**.

Probit estimates (see Table A.2 in Online Appendix A), controlling for opening offers and personal characteristics, also support the findings above. In particular, the result in Specification (4), which uses the subsample of unfair opening offers, shows that the coefficient on **T1** is small and statistically insignificant, whereas the coefficient on **T3** is positive and statistically significant. This indicates that, relative to **T2**, human responders in **T1** are not significantly less or more likely to accept unfair opening offers, while human responders in **T3** are significantly more likely to accept them. Therefore, we have the following results.

Result 5. *Human responders do not differ significantly in their willingness to accept unfair opening offers from human and AI proposers.*

Result 6. *Human responders are more willing to accept unfair opening offers from AI proposers when the AI agent's earnings are linked to another human participant's payment.*

4.3.2 AI Responder

In treatments **T2** and **T3**, AI agents responded to 130 human opening offers in each treatment. The overall acceptance rate was 84.2%, with acceptance rates of 84.6% in **T2** and 83.8% in **T3**. When the sample is restricted to unfair offers, the number of observations decreases to 117 in **T2** and 109 in **T3**. In this restricted sample, the acceptance rates were 82.9% in **T2** and 80.7% in **T3**. A Fisher's exact test confirms that this difference is not statistically significant ($p = 0.731$).

To further assess the robustness of these patterns, we conducted an additional API-based exercise using the same prompt. For each possible human opening offer from 0 to 100, we independently elicited the AI agent's response 100 times, together with its explanation and, when the offer was rejected, the corresponding counteroffer. The results are reported in Online Appendix D.2.

4.4 Counteroffers

4.4.1 Human Counteroffers

When a human responder (P2) rejects the opening offer proposed by P1, the responder makes a counteroffer in Stage 2. This counteroffer provides additional information on how responders trade off monetary payoff against fairness concerns (Ochs and Roth, 1989). In particular, if the responder's discounted payoff implied by the Stage-2 counteroffer is lower than the payoff that would have been obtained by accepting the opening offer, then monetary payoff alone cannot fully account for the rejection decision. In such cases, the responder appears willing to incur a material cost in order to reject an unfair offer, which is consistent with fairness-based punishment (Brañas-Garza et al., 2014). Thus, materially costly rejection may reflect fairness considerations rather than purely monetary optimization.

To capture this idea, we define the *FairnessGap* as

$$FairnessGap = Opening\ Offer - 0.4 \times (100 - CounterOffer),$$

where $100 - CounterOffer$ is the responder's own payoff implied by the counteroffer, and 0.4 is the responder's discount factor in Stage 2. A positive fairness gap indicates that rejecting the opening offer is materially costly relative to the continuation payoff implied by the responder's own counteroffer.

As a supplementary exploratory analysis, we examine the subset of rejected opening

offers. The number of such observations is limited, with 15 in **T1**, 18 in **T2**, and only 3 in **T3**. Mean counteroffers are lower in **T2** than in **T1** (19.39 vs. 27), but a Mann–Whitney U test indicates that this difference is not statistically significant ($p = 0.154$). Likewise, the difference in the fairness-gap measure is not statistically significant ($p = 0.365$).

By contrast, all rejected opening offers in **T2** involve materially costly rejection, whereas this is true for 11 out of 15 rejected opening offers in **T1**. A Fisher's exact test indicates that this difference is statistically significant ($p = 0.033$). However, given the very small subsample size, especially in **T3**, these results should be interpreted with caution.

4.4.2 AI Counteroffers

Compared with human counter-proposers, AI counter-proposers appear to be more generous. We therefore examine AI counteroffers following rejected opening offers. The number of such observations is 20 in **T2** and 21 in **T3**. Mean counteroffers are very similar across the two treatments (58 in **T2** and 59.05 in **T3**), and a Mann–Whitney U test does not indicate a significant difference ($p = 0.544$).

Likewise, materially costly rejection is observed in almost all cases in both treatments: all 20 cases in **T2** and 20 out of 21 cases in **T3**. A Fisher's exact test confirms that this difference is not statistically significant ($p = 1.000$). Additional API-based exercises using the same prompt to elicit the AI's counteroffers are reported in Online Appendix D.2.

5 Discussion

Our main analysis compares **T1** with **T2**, and **T2** with **T3**, in order to examine how human bargaining behavior changes when the counterpart is an AI agent rather than a human, and when the AI agent's earnings are linked to another human participant's

payment. The results reveal an important **asymmetry** between proposer behavior and responder behavior:

- Human proposers are more generous toward human opponents than toward AI agents, but the human-beneficiary manipulation does not significantly increase their average generosity toward AI agents.
- Human responders do not differ significantly in their willingness to accept unfair opening offers from AI rather than human proposers, but they become substantially more willing to accept such offers when the AI agent's earnings are linked to another human participant's payment.

These **asymmetric findings** make it difficult to characterize the role of social preferences in human–AI bargaining using a single, uniform framework. Rather than suggesting that social-preference considerations simply weaken when the counterpart is an AI agent and are then restored once the AI's earnings affect another human, the results indicate that *the influence of social preferences differs between proposers and responders*.

We therefore shift our focus in this section to several additional perspectives that may help interpret the asymmetry: (1) FMA, (2) learning, and (3) participants' prior and posterior beliefs.

5.1 First-Mover Advantage

In sequential bargaining, the player who makes the first proposal is generally thought to enjoy an FMA. In the standard alternating-offer framework, making the first offer allows a player to anchor the bargaining process and to exploit the fact that delay is costly for both parties (Rubinstein, 1982; Ochs and Roth, 1989). In the setting of this study, this advantage is further strengthened by the asymmetry in discount factors: P1

faces a lower cost of delay than P2, because $\delta_1 = 0.6$ while $\delta_2 = 0.4$. As a result, P1 is in a relatively stronger bargaining position than P2.

5.1.1 Existence and Strength of FMA

Although the opening offers discussed in Section 4.2.1 already reflect P1's initial claim over the surplus, we further assess FMA by comparing P1's realized payoff and P1's share of the total realized payoff across treatments.

Figure 7 reports the realized payoffs of human P1 and human P2, where realized payoff refers to the point payoff that a human participant actually obtained in each round after discounting. This measure captures the absolute extent of FMA. Figure 8 reports P1's share of the total realized payoff, calculated as

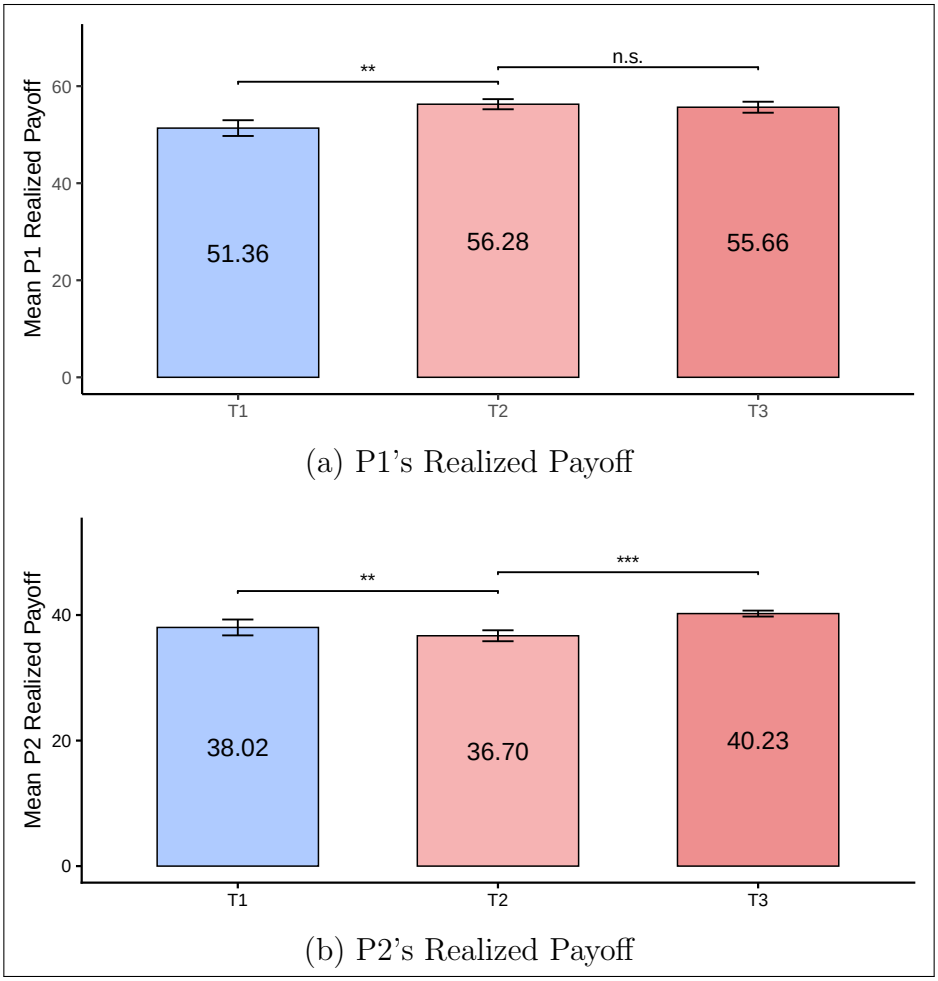
$$\frac{\text{P1's realized payoff}}{\text{P1's realized payoff} + \text{P2's realized payoff}},$$

which captures the relative extent of FMA.

First, within each treatment, P1's realized payoff is clearly and significantly higher than P2's realized payoff (Mann–Whitney U test, $p < 0.001$ in all three treatments). This provides direct evidence that an FMA exists in our bargaining environment.

Second, comparing across treatments, the FMA appears strongest in **T2**. Relative to **T1**, **T2** yields a significantly higher realized payoff for P1, a significantly lower realized payoff for P2, and a significantly larger share of the total realized payoff captured by P1. These results suggest that when bargaining against a pure AI agent, P1 is better able to convert the structural FMA into a larger share of the realized surplus.

Third, the comparison between **T2** and **T3** suggests that this pattern is partly mitigated when the AI agent's payoff is linked to another human participant's payment. Although P1's realized payoff does not differ significantly between **T2** and **T3**, its mean value declines slightly. At the same time, P2's realized payoff becomes significantly higher in **T3**, and P1's share of the total realized payoff also declines in magnitude,



(a) P1's Realized Payoff

(b) P2's Realized Payoff

Figure 7: Comparisons of Realized Payoff across Treatments

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Realized payoffs are compared across treatments using the Mann–Whitney U test. Error bars denote 95% confidence intervals. In **T2** and **T3**, only the outcomes of human P1 and human P2 are included.

although the difference is not statistically significant.

We also estimate regressions of realized payoff and the share of the total realized payoff, controlling for personal characteristics and round number. The results, reported in Tables A.3–A.5 in Online Appendix A, are consistent with the graphical evidence. In particular, relative to **T2**, **T1** is associated with a significantly lower realized payoff for P1 and a significantly lower share of the total realized payoff captured by P1, while **T3** is associated with a weakly positive effect on P2's realized payoff.

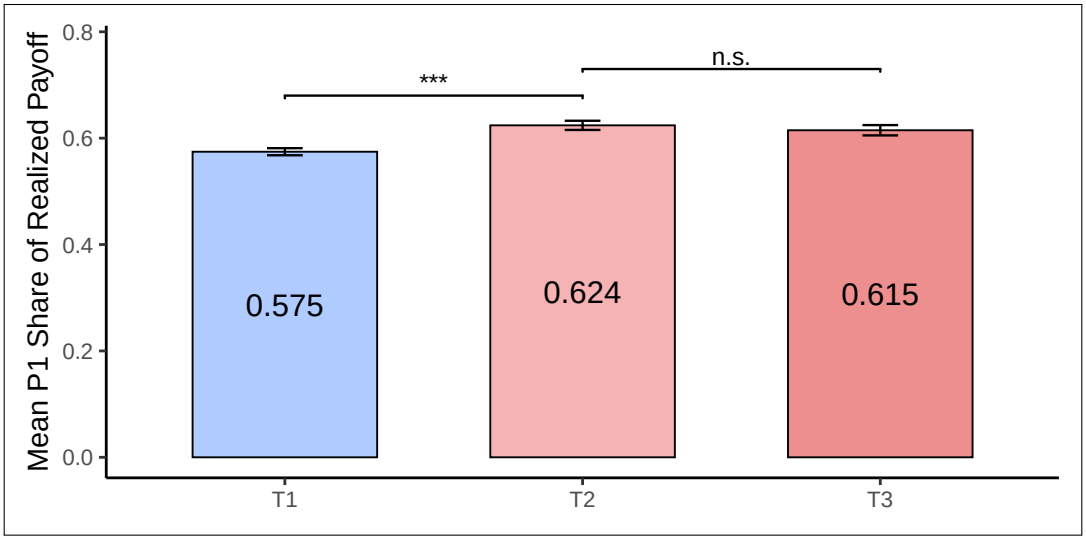


Figure 8: Comparisons of P1's Share of the Total Realized Payoff across Treatments

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals. Shares of the total realized payoff are compared across treatments using the Mann–Whitney U test.

To summarize, the pattern of FMA across the three treatments mirrors the main asymmetry in our results. The comparison between **T1** and **T2** shows that bargaining with a pure AI agent rather than a human is associated with a significantly larger FMA. By contrast, the human-beneficiary manipulation does not significantly reduce FMA in **T3**, either in terms of absolute realized payoffs or in terms of P1's share of total realized payoffs. At the same time, however, it significantly improves responders' realized payoffs, suggesting that it partly mitigates their disadvantaged bargaining position.

This pattern is also consistent with the interpretation that linking the AI agent's payoff to another participant's payment partly restores the social dimension of the interaction. When the counterpart is purely artificial, extracting a larger share may be perceived as having little social cost. When the AI's payoff potentially affects another human, this perception may be weakened, which could contribute to the observed increase in responders' willingness to accept unequal offers.

5.1.2 FMA as a Possible Explanation for the Asymmetry

The strong FMA documented above may also help explain the asymmetric findings between proposer and responder behavior in the main analysis. A possible interpretation is that the behavioral relevance of social preferences depends on the strategic role occupied by the human participant. Prior evidence from ultimatum and alternating-offer bargaining experiments suggests that responder decisions are often closely tied to fairness judgments over proposed allocations, as reflected in the rejection of unfair offers. Proposer decisions, by contrast, reflect not only fairness concerns but also strategic considerations, such as expectations about acceptance thresholds and the distribution of bargaining power (Güth et al., 1982; Camerer and Thaler, 1995; Fehr and Schmidt, 1999; Ochs and Roth, 1989).

This role-based interpretation can first explain why the human–AI difference appears more clearly on the proposer side than on the responder side. For proposers, counterpart identity is directly relevant to the allocation decision: they decide how much surplus to give to a human opponent or to an AI agent. If social preferences are weaker toward AI agents than toward human opponents, this difference can translate directly into lower offers in the human–AI treatment. For responders, however, counterpart identity enters the decision less directly. Responders decide whether to accept or reject a given allocation, and this decision is shaped not only by the identity of the proposer but also by the material payoff from acceptance, the desire to punish unfairness, and the expected value of continuing the bargaining process. As a result, the human or AI identity of the proposer alone may not be sufficient to generate a significant difference in acceptance behavior.

In our bargaining environment, proposers enjoy a substantial FMA, and this advantage is especially strong when humans bargain with AI agents rather than with other humans. This provides a possible explanation for the contrast between **Result 3** and **Result 4**. In the human–AI treatments, proposers are already in a favorable strategic

position and can claim a relatively large share of the surplus. Therefore, even if the human-beneficiary manipulation makes social-preference considerations more salient, such concerns may be too weak to overcome the strategic incentives created by the proposer's advantageous position. In this sense, social preferences may still operate on the proposer side, but their behavioral expression is constrained by FMA and strategic considerations.

On the responder side, *the absence of FMA may make the human-beneficiary manipulation more behaviorally consequential*. Unlike proposers, responders do not occupy the structurally advantageous position created by moving first. Instead, they face a proposed allocation and decide whether to accept it or reject it. This weaker strategic position may make their decisions more sensitive to changes in the social consequences of acceptance or rejection. The human-beneficiary manipulation changes precisely this consequence. In the pure human–AI treatment, rejecting an unfair AI offer mainly functions as a response to an unfair allocation proposed by a non-human counterpart. In the human-beneficiary treatment, however, rejection may also impose a monetary cost on another human participant. Thus, the same accept/reject decision involves an additional social consequence, which may shift the relevant concern from punishing unfairness to avoiding harm to another human beneficiary.

Taken together, the asymmetry has two layers. First, the human–AI difference is more visible for proposers because counterpart identity directly affects the allocation decision. Second, the human-beneficiary manipulation is more visible for responders because it changes the social consequence of rejecting an unfair offer, whereas its effect on proposers may be muted by the strong FMA in the human–AI treatments. Thus, social preferences are not uniformly weakened or restored in human–AI bargaining; rather, their behavioral expression *depends on the strategic role and decision environment*.

5.2 Learning and Adaptation

Since most participants were likely unfamiliar with repeated alternating-offer bargaining, especially against AI agents, some degree of learning and adaptation over the 10 rounds is to be expected. As subjects gained experience, they may have updated both their understanding of the strategic structure of the game and their expectations about counterpart behavior. In this subsection, we examine whether such learning is present, and whether it helps explain the asymmetric treatment effects observed above.

5.2.1 Evidence of Learning Across Rounds

Figure 9 illustrates the evolution of human proposers' opening offers and human responders' acceptance rates over the 10 rounds. Across treatments, opening offers show a clear downward trend. By contrast, on the responder side, the pattern differs across treatments. In **T1**, the acceptance rate for unfair opening offers changes only modestly across rounds. In **T2** and **T3**, however, the acceptance rate fluctuates somewhat in the first half of the experiment, but remains close to 100% in the later rounds.

The regression results reported in Online Appendix A provide further detail. In Table A.1, the coefficient on *roundnum* is negative and statistically significant in the opening-offer regressions. This suggests that, over time, proposers lowered their opening offers. One possible interpretation is that participants gradually came to better understand their FMA and therefore became less generous as they gained experience. At the same time, the interaction terms between treatment indicators and *roundnum* are not statistically significant, suggesting that this proposer-side learning effect does not differ substantially across treatments.

On the responder side, Table A.2 shows that the coefficient on *roundnum* is positive and statistically significant in the regressions for accepting unfair opening offers. This suggests that, as rounds progressed, responders became more willing to accept unfair opening offers. However, when interactions between treatment indicators and *roundnum*

(a) P1's Opening Offer

(b) P2's Acceptance Rate Facing Unfair Opening Offers

Figure 9: Learning and Adaptation across Rounds

Note: Panel (a) shows the mean opening offer of human proposers (P1) across rounds by treatment. Panel (b) shows the acceptance rate of human responders (P2) for unfair opening offers across rounds by treatment. Error bars denote 95% confidence intervals.

are introduced (Table A.6), the results suggest that this responder-side learning effect is weaker in **T1** than in **T2** and **T3**, while there is no clear difference between **T2** and **T3**. This pattern may indicate that adaptation was stronger when participants repeatedly interacted with AI agents than when they interacted with humans.

By contrast, Tables A.4 and A.5 show that *roundnum* has no significant effect on realized payoffs. This suggests that learning primarily affected bargaining behavior itself—such as offer levels and acceptance decisions—rather than improving final material outcomes. In other words, participants appear to have adapted their strategies over

time, but this adaptation did not necessarily translate into higher realized earnings.

These findings indicate that learning and adaptation were present in the experiment, but they do not alter the main treatment patterns. Instead, learning appears to operate alongside the treatment effects: proposers generally became less generous over time, responders generally became more willing to accept offers, and the responder-side adaptation was somewhat stronger in AI bargaining than in human–human bargaining.

5.2.2 Learning and the Asymmetry Between Proposers and Responders

As shown above, behavior exhibits learning on both the proposer side and the responder side over the 10 rounds. If subjects' experience shapes how treatment differences emerge, then the strength of the treatment effects may differ between earlier and later rounds. To examine this possibility, we re-estimate the regressions for proposers' opening offers and responders' acceptance of unfair offers separately for Rounds 1–5 and Rounds 6–10. The results are reported in Table A.7 of Online Appendix A.

On the proposer side, the results in Table A.7 are broadly consistent with those in Table A.1 based on the pooled sample. In particular, the coefficient on **T3** remains insignificant in both subsamples, suggesting that linking the AI agent's payoff to a human beneficiary does not significantly affect proposers' opening offers, even after accounting for potential heterogeneity by experience.

On the responder side, the pattern is more nuanced. In the first five rounds, the significance of the coefficients on **T1** and **T3** is broadly similar to that in column (4) of Table A.2 based on the full sample of unfair offers. In the last five rounds, however, the coefficient on **T1** becomes significantly negative, indicating that responders in **T1** are significantly less likely to accept unfair offers than those in **T2**, whereas responders in **T3** remain significantly more likely to accept them. This late-round pattern further contradicts **H2b**, which predicted that human responders would be less willing to accept unfair offers from AI agents than from human opponents.

One possible interpretation is that learning affects how responders interpret the source of unfair offers, rather than simply shifting their acceptance thresholds. In early rounds, responders may evaluate unfair offers in a relatively undifferentiated way, focusing mainly on their own payoff. As they gain experience, however, they may become better able to infer whether the offer reflects intentional behavior or not. In particular, unfair offers from human proposers may be increasingly interpreted as intentional violations of fairness norms, which strengthens the willingness to reject them. By contrast, unfair offers generated by AI agents may continue to be perceived as less intentional or less blameworthy, making them relatively more acceptable even in later rounds.

In **T3**, this effect may be further amplified because rejecting an unfair AI offer not only responds to the allocation itself but may also impose a cost on another human participant. As a result, learning may reinforce a divergence in how responders evaluate human and AI-generated offers over time, thereby sharpening the responder-side asymmetry observed in the later rounds.

While learning alone does not explain the origin of the asymmetry, the results here suggest that it shapes how the asymmetry emerges over time. In particular, learning appears to sharpen the responder-side asymmetry in later rounds by increasing the role of perceived intentionality and payoff consequences in the evaluation of unfair offers.

5.3 Prior Beliefs

In Survey A, which was conducted prior to the main task, participants were asked to predict the behavior of human participants and, in **T2** and **T3**, the behavior of the AI agent (see Online Appendix C.1). These elicited beliefs allow us to compare prior expectations and expectation biases across treatments, and to explore whether they help explain the asymmetry in the main findings.

Detailed descriptions and analyses are reported in Online Appendix E. Here, we briefly summarize the main findings.

Predictions of Opening Offers. There is no significant difference in predicted opening offers between **T1** and **T2**, or between **T2** and **T3**.

Predictions of Agreement Timing. Participants in **T2** expected agreements to be reached slightly later than those in **T1**.

Expectation Biases. We measure expectation bias as the difference between predicted and actual outcomes. The results show that participants exhibited systematic expectation errors, particularly in the AI-related treatments. In particular, they (1) overestimated the human proposer's opening offer when bargaining with an AI agent; (2) overestimated the AI agent's opening offer to a human responder; and (3) underestimated the speed at which agreements were reached, especially when bargaining with AI agents. In addition, participants who overestimated human proposers' opening offers tended to make higher opening offers themselves, but were less willing to accept unfair offers as responders.

Possible implications for the asymmetry. Although these prior beliefs help illuminate some behavioral patterns on both the proposer side and the responder side, they do not appear to be the main reason why treatment effects differ between the two roles. In particular, controlling for prior beliefs does not overturn the main treatment patterns. Thus, prior beliefs are unlikely to be the key source of the asymmetry in the main results.

5.4 Posterior Beliefs

In Survey B, which was conducted after the main task, participants were asked to report their feelings during the task, their self-reported strategies, their perceptions of the AI agent, and, in **T3**, their understanding of the human-beneficiary payment rule (see Online Appendix C.2).

Detailed descriptions and analyses are reported in Online Appendix F. Here, we

briefly summarize the main findings.

Feelings. Relative to bargaining with a human opponent, bargaining with an AI agent elicited weaker social-emotional reactions, particularly less anger, disappointment, and confusion during the alternating-offer bargaining task. At the same time, the human-beneficiary manipulation partly restored participants' sense of being respected.

Self-reported strategies. Compared with human–AI bargaining, human–human bargaining was associated with a significantly higher tendency to report rejecting or retaliating against unfair offers. By contrast, participants in the AI treatments were marginally more likely to report probing the counterpart's minimum acceptable level. This pattern is consistent with the view that human–human bargaining involved stronger social considerations, whereas human–AI bargaining was approached somewhat more instrumentally.

Perceptions of bargaining with AI. None of the AI-perception measures differs significantly between **T2** and **T3**. This suggests that the human-beneficiary manipulation did not substantially change how participants viewed the AI agent itself.

Understanding of the human-beneficiary payment rule. The **human-beneficiary payment rule** was generally well understood. However, although most participants understood the rule, many reported that it did not strongly affect their decisions. This suggests that the manipulation was psychologically meaningful, but only moderately salient.

Possible implications for the asymmetry. These posterior responses suggest that the asymmetry is unlikely to be driven by a broad change in how participants perceived the AI agent itself. Instead, they point to a more decision-specific interpretation: participants reacted less emotionally to unfair offers from AI agents, while the human-beneficiary manipulation made the payoff consequences for another human participant salient, but not strongly enough to affect all decisions equally.

6 Conclusion

As LLM-based AI negotiation systems increasingly appear in real-world commercial settings, understanding how humans bargain with AI counterparts has become an important question. In this paper, we conducted a laboratory experiment to compare human–human bargaining and human–AI bargaining in a 3-stage alternating-offer game. In addition, motivated by the fact that many real-world AI bargaining systems act on behalf of firms or other humans, we introduced a human-beneficiary manipulation in which the AI agent's earnings could affect another participant's final payoff. Our aim was to examine, first, whether social-preference considerations become weaker in human–AI bargaining than in human–human bargaining, and second, whether linking the AI agent's payoff to another human participant can partially restore such considerations.

Our results show that, in an alternating-offer bargaining game with at most three stages, the speed of reaching agreement does not differ significantly between human–human bargaining and human–AI bargaining. However, when the AI agent's earnings are linked to another human participant, agreements are reached significantly earlier. This suggests that the human-beneficiary manipulation promotes earlier agreement in human–AI bargaining and is therefore consistent with a partial restoration of social-preference considerations at the aggregate level.

More importantly, we find a clear asymmetry between proposer behavior and responder behavior. On the proposer side, human participants offer more to human opponents than to AI agents, but the human-beneficiary manipulation does not significantly increase offers to AI agents. On the responder side, human participants do not differ significantly in their willingness to accept unfair opening offers from human versus AI proposers. However, when the AI agent's earnings are linked to another human participant's payment, responders become significantly more willing to accept such unfair offers. These findings indicate that treatment effects in human–AI bargaining

are role-dependent. The weakening of social-preference considerations in human–AI bargaining appears more clearly on the proposer side, whereas the restoring effect of the human-beneficiary manipulation appears more clearly on the responder side.

We discuss several possible explanations for this asymmetry. Our additional analyses suggest that learning, prior beliefs, and posterior beliefs do not provide the main explanation for the difference between proposer-side and responder-side treatment effects. Instead, our analysis of FMA suggests one possible interpretation. In our bargaining environment, proposers enjoy a strong structural advantage, especially in human–AI bargaining. As a result, the human-beneficiary manipulation may not be strong enough to generate a significant increase in proposers' offers. By contrast, responders do not benefit from the same FMA, and their accept-or-reject decisions make the social consequences of unfair offers more directly relevant. This may explain why the human-beneficiary manipulation appears more clearly on the responder side.

Overall, our findings contribute to the growing literature on human–AI bargaining, especially bargaining with LLM-based AI agents, which is becoming increasingly realistic in practice. More broadly, the results suggest that human responses to AI bargaining systems cannot be characterized by a single uniform shift in behavior. Instead, such responses depend on the strategic role of the human participant and on whether the AI agent's payoff is socially consequential. These findings may also help inform the design of AI negotiation systems by highlighting the importance of users' social and psychological reactions to AI counterparts, thereby helping firms better anticipate human responses and design negotiation agents that more effectively pursue organizational objectives.

This study has several limitations. First, although our game allows for up to three stages, the limited number of observations beyond Stage 1 means that most of our analysis focuses on opening offers and responses to opening offers, while evidence on later-stage counteroffers remains limited. Second, we examine only one combination

of discount factors, corresponding to one of the treatments in Ochs and Roth (1989). Because this parameterization strengthens the first-mover advantage of the proposer, future research should investigate whether alternative discount-rate combinations lead to different patterns of asymmetry. Third, the AI agent used in our experiment was prompted only with the rules of the game and was not strategically tuned through prompt engineering. In real-world applications, however, AI negotiation systems are often carefully designed and optimized for particular goals. Future research could therefore examine how different prompting strategies or AI behavioral styles affect human responses in bargaining environments.

References

F. M. Affonso. Large language models converge on competitive rationality but diverge on cooperation across providers and generations. *arXiv preprint arXiv:2604.18596*, 2026.

F. Bianchi, P. J. Chia, M. Yuksekgonul, J. Tagliabue, D. Jurafsky, and J. Zou. How well can LLMs negotiate? negotiationarena platform and a snalysis. *arXiv preprint arXiv:2402.05863*, 2024.

D. Borthakur, P. Diep, and J. E. Plaks. Inequity aversion toward AI counterparts. *Scientific Reports*, 15(1):37916, 2025.

P. Brañas-Garza, A. M. Espín, F. Exadaktylos, and B. Herrmann. Fair and unfair punishers coexist in the ultimatum game. *Scientific Reports*, 4:6025, 2014. doi: 10.1038/srep06025.

P. Brookins and J. M. DeBacker. Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Available at SSRN 4493398*, 2023.

C. Camerer and R. H. Thaler. Anomalies: Ultimatums, dictators and manners. *Journal of Economic perspectives*, 9(2):209–219, 1995.

D. L. Chen, M. Schonger, and C. Wickens. oTree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016. doi: 10.1016/j.jbef.2015.12.001.

Y. Chen and R. Huang. Haggling with a bot: Human vs. LLM negotiation in supply chain contracts. *Available at SSRN: https://ssrn.com/abstract=6278158*, 2026.

M. Chugunova and D. Sele. We and it: An interdisciplinary review of the experimental

evidence on human-machine interaction. *Center for law & economics working paper series*, 12, 2020.

T. R. Davidson, V. Veselovsky, M. Josifoski, M. Peyrard, A. Bosselut, M. Kosinski, and R. West. Evaluating language model agency through negotiations. *arXiv preprint arXiv:2401.04536*, 2024.

F. Dvorak, R. Stumpf, S. Fehrler, and U. Fischbacher. Adverse reactions to the use of large language models in social interactions. *PNAS Nexus*, 4(4):pgaf112, 2025.

A. Erlei, R. Das, L. Meub, A. Anand, and U. Gadiraju. For what it's worth: Humans overwrite their economic self-interest to avoid bargaining with AI systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

M. Faithfull. The future of haggling: Bargain hunters negotiate deals with AI bot. *Forbes*, August 2024.

E. Fehr and K. M. Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999.

B. Greiner. Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1:114–125, 2015.

F. Guo. GPT in game theory experiments. *arXiv preprint arXiv:2305.05516*, 2023.

W. Güth, R. Schmittberger, and B. Schwarze. An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, 3(4):367–388, 1982.

T.-R. Heggedal and T. McKay. Discounting in finite-time bargaining experiments. *Journal of the Economic Science Association*, 10(2):504–518, 2024.

IBM. What is agentic commerce? https://www.ibm.com/think/topics/agentic-commerce, Jan. 2026. Accessed: 2026-04-25.

M. O. Keskin, U. Çakan, and R. Aydoğan. An adaptive emotion-aware strategy for human-agent negotiation: Insights from real-world human-robot experiments. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, page 29. Association for Computing Machinery, 2025. doi: 10.1145/3717511.3747087.

D. Kong, X. Yan, M. Chen, S. Han, J. Chen, and F. Huang. FishBargain: An LLM-empowered bargaining agent for online fleamarket platform sellers. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2855–2858, 2025.

D. Kwon, E. Weiss, T. Kulshrestha, K. Chawla, G. Lucas, and J. Gratch. Are LLMs effective negotiators? systematic evaluation of the multifaceted capabilities of LLMs in negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5391–5413, 2024.

B. Liefooghe, E. Min, and H. Aarts. The effects of social presence on cooperative trust with algorithms. *Scientific Reports*, 13(1):17463, 2023.

R. Lin and S. Kraus. Can automated agents proficiently negotiate with humans? *Communications of the ACM*, 53(1):78–88, 2010.

X. Luo, N. R. Jennings, and N. Shadbolt. Acquiring user tradeoff strategies and preferences for negotiating agents: A default-then-adjust method. *International Journal of Human-Computer Studies*, 64(4):304–321, 2006.

J. Mell and J. Gratch. IAGO: Interactive arbitration guide online. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, pages 1510–1512. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

J. Ochs and A. E. Roth. An experimental study of sequential bargaining. *American Economic Review*, 79(3):355–384, June 1989.

H. Oosterbeek, R. Sloof, and G. Van de Kuilen. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7:171–188, 2004.

A. I. Ozkes, N. Hanaki, D. Vanderelst, and J. Willems. Ultimatum bargaining: Algorithms vs. humans. *Economics Letters*, 244:111979, 2024.

A. Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109, 1982.

Z. Shen and L. Jin. Bargaining with algorithms: How consumers respond to offers proposed by algorithms versus humans. *Journal of Retailing*, 100(3):345–361, 2024. doi: 10.1016/j.jretai.2024.05.001.

S. Sinha, H. Kumar, A. R. Mandapati, R. Sakhuja, and D. Kumar. The language of bargaining: Linguistic effects in LLM negotiations. *arXiv preprint arXiv:2601.04387*, 2026.

D. Sondern, N. Arnholz, and G. Hertel. Employment negotiations with an algorithm? how AI as negotiation counterpart would affect negotiators' trust and subjective value expectations. *Conflict Resolution Quarterly*, 43(1):5–13, 2025.

J. Sonnegård. Determination of first movers in sequential bargaining games: An experimental study. *Journal of Economic Psychology*, 17(3):359–386, 1996.

R. Van Hoek, M. DeWitt, M. Lacity, and T. Johnson. How walmart automated supplier negotiations. *Harvard Business Review*, November 2022.

Visa. Visa defines the next era of commerce: When AI becomes the customer. https://usa.visa.com/about-visa/newsroom/press-releases.releaseId.22266.html, Apr. 2026. Accessed: 2026-04-25.

A. von Schenk, V. Klockmann, and N. Köbis. Social preferences toward humans and machines: a systematic experiment on the role of machine payoffs. *Perspectives on Psychological Science*, 20(1):165–181, 2025. doi: 10.1177/17456916231194949.

E. Weg, A. Rapoport, and D. S. Felsenthal. Two-person bargaining behavior in fixed discounting factors games with infinite horizon. *Games and Economic Behavior*, 2 (1):76–95, 1990.

E. Weg, R. Zwick, and A. Rapoport. Bargaining in uncertain environments: A systematic distortion of perfect equilibrium demands. *Games and Economic Behavior*, 14 (2):260–286, 1996.

T. Xia, Z. He, T. Ren, Y. Miao, Z. Zhang, Y. Yang, and R. Wang. Measuring bargaining abilities of LLMs: A benchmark and a buyer-enhancement method. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3579–3602, 2024.

Online Appendix to "Bargaining with GPT in Alternating-Offer Games"

Yuhao Fu* Nobuyuki Hanaki† Haitao Wang‡

April 28, 2026

Contents

*Graduate School of Economics, University of Osaka. E-mail: u889037j@ecs.osaka-u.ac.jp

†Corresponding author. Institute of Social and Economic Research, University of Osaka, and University of Limassol. E-mail: nobuyuki.hanaki@iser.osaka-u.ac.jp

‡A non-academic institution. E-mail: wangtiedan2@yahoo.com

A Regression Tables

Table A.1: OLS Estimates for Human Opening Offers

Dep. Var.	Opening Offer			
	(1)	(2)	(3)	(4)
T1	3.813**	3.801**	1.740	3.808**
	(1.431)	(1.440)	(2.379)	(1.439)
T3	0.282	0.227	0.219	1.428
	(1.650)	(1.662)	(1.670)	(2.642)
roundnum		−0.940***	−1.065***	−0.867***
		(0.174)	(0.222)	(0.222)
age	0.097	0.094	0.090	0.093
	(0.234)	(0.222)	(0.221)	(0.223)
engr	−1.891	−2.038	−2.121	−2.086
	(1.706)	(1.656)	(1.666)	(1.660)
edulevel	1.801	2.014	2.072	2.003
	(1.710)	(1.656)	(1.663)	(1.663)
female	1.548	1.341	1.304	1.319
	(1.403)	(1.381)	(1.391)	(1.377)
freqGPT	−1.016	−1.093	−1.110	−1.083
	(0.733)	(0.761)	(0.768)	(0.760)
T1 × roundnum			0.375	
			(0.343)	
T3 × roundnum				−0.218
				(0.348)
Constant	38.377***	43.928***	44.791***	43.544***
	(5.521)	(5.301)	(5.329)	(5.474)
Adj. R^2	0.0566	0.1456	0.1466	0.1444
No. cluster	78	78	78	78
No. Obs.	390	390	390	390

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. **T2** is the omitted treatment category; $T1$ and $T3$ are treatment indicators and *roundnum* is the number of rounds. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table A.2: Probit Estimates for Human Responders Accepting the Opening Offer

Dep. Var.	Accept the Opening Offer			
	Full Sample			Unfair Sample
	(1)	(2)	(3)	(4)
T1	−0.006	−2.849*	−0.006	−0.114
	(0.292)	(1.304)	(0.292)	(0.309)
T3	1.098*	1.101*	2.977*	1.107*
	(0.484)	(0.478)	(1.354)	(0.489)
Offer	0.097**	0.038*	0.097**	0.080*
	(0.031)	(0.016)	(0.031)	(0.037)
roundnum	0.092*	0.091*	0.092*	0.095*
	(0.039)	(0.040)	(0.039)	(0.037)
age	−0.034	−0.038	−0.034	−0.038
	(0.067)	(0.066)	(0.067)	(0.066)
engr	0.337	0.313	0.337	0.342
	(0.421)	(0.421)	(0.422)	(0.419)
edulevel	−0.419	−0.408	−0.420	−0.429
	(0.380)	(0.378)	(0.381)	(0.372)
female	0.282	0.299	0.283	0.283
	(0.335)	(0.340)	(0.335)	(0.337)
freqGPT	0.353**	0.344**	0.353**	0.357**
	(0.126)	(0.125)	(0.126)	(0.128)
T1 × Offer		0.072*		
		(0.035)		
T3 × Offer			−0.047	
			(0.031)	
Constant	−3.412+	−0.952	−3.418+	−2.675
	(2.070)	(1.701)	(2.074)	(2.229)
AIC	205.48	205.94	207.46	203
No. cluster	78	78	78	78
No. Obs.	390	390	390	357

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. **T2** is the omitted treatment category, and $Offer$ denotes the opening offer made by proposers. Specifications (1)–(3) use the full sample, whereas Specification (4) uses only the subsample of unfair offers (offers below 50 points). For the unfair-offer subsample, specifications including interactions between treatment indicators and $Offer$ are not reported because the estimates are unstable, as a very large share of offers in **T2** and **T3** are concentrated at 40. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table A.3: OLS Estimates for P1 Realized Payoff

Dep. Var.	Realized Payoff			
	(1)	(2)	(3)	(4)
T1	-4.519^{*}	-4.515^{*}	1.860	-4.515^{*}
	(1.944)	(1.946)	(3.477)	(1.954)
T3	-0.654	-0.636	-0.611	-0.592
	(1.427)	(1.420)	(1.415)	(3.450)
roundnum		0.302	0.690^{**}	0.304
		(0.274)	(0.258)	(0.370)
age	0.388^{+}	0.389^{+}	0.401^{+}	0.389^{+}
	(0.221)	(0.219)	(0.215)	(0.220)
engr	1.303	1.350	1.606	1.348
	(1.550)	(1.535)	(1.546)	(1.539)
edulevel	-3.307^{*}	-3.375^{*}	-3.557^{*}	-3.376^{*}
	(1.465)	(1.455)	(1.405)	(1.462)
female	-2.565^{+}	-2.499	-2.383	-2.500
	(1.525)	(1.555)	(1.536)	(1.539)
freqGPT	1.044^{+}	1.068^{+}	1.120^{+}	1.069^{+}
	(0.612)	(0.612)	(0.621)	(0.614)
T1 × roundnum			-1.160^{+}	
			(0.653)	
T3 × roundnum				-0.008
				(0.522)
Constant	45.438^{***}	43.655^{***}	40.986^{***}	43.641^{***}
	(5.799)	(5.794)	(6.140)	(5.789)
Adj. R^2	0.0242	0.0251	0.0339	0.0225
No. cluster	78	78	78	78
No. Obs.	390	390	390	390

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. **T2** is the omitted treatment category; $T1$ and $T3$ are treatment indicators and *roundnum* is the number of rounds. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table A.4: OLS Estimates for P2 Realized Payoff

Dep. Var.	Realized Payoff			
	(1)	(2)	(3)	(4)
T1	0.632	0.639	9.757**	0.657
	(2.118)	(2.120)	(2.971)	(2.117)
T3	3.066+	3.097+	3.058+	−0.424
	(1.607)	(1.605)	(1.608)	(2.867)
roundnum		−0.286	0.269	−0.500
		(0.222)	(0.175)	(0.313)
age	−0.626	−0.619	−0.617	−0.615
	(0.509)	(0.510)	(0.507)	(0.512)
engr	−3.373*	−3.246*	−3.233*	−3.322*
	(1.624)	(1.633)	(1.603)	(1.622)
edulevel	3.012	2.895	2.995	2.808
	(2.088)	(2.066)	(2.021)	(2.066)
female	−0.657	−0.584	−0.691	−0.644
	(2.115)	(2.111)	(2.103)	(2.119)
freqGPT	0.557	0.586	0.553	0.624
	(0.679)	(0.681)	(0.669)	(0.677)
T1 × roundnum			−1.660**	
			(0.537)	
T3 × roundnum				0.643+
				(0.368)
Constant	50.851***	52.099***	49.127***	53.160***
	(9.329)	(9.521)	(9.445)	(9.593)
Adj. R^2	0.0402	0.0437	0.0868	0.048
No. cluster	78	78	78	78
No. Obs.	390	390	390	390

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. **T2** is the omitted treatment category; $T1$ and $T3$ are treatment indicators and *roundnum* is the number of rounds. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table A.5: OLS Estimates for P1 Share of the Total Realized Payoff

Dep. Var.	P1 Share of the Total Realized Payoff			
	(1)	(2)	(3)	(4)
T1	-0.054^{**}	-0.052^{**}	-0.041^{+}	-0.052^{**}
	(0.016)	(0.017)	(0.025)	(0.017)
T3	-0.014	-0.013	-0.013	-0.011
	(0.019)	(0.019)	(0.020)	(0.031)
roundnum		0.009^{***}	0.009^{***}	0.009^{***}
		(0.002)	(0.002)	(0.002)
age	0.001	0.001	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)
engr	0.029^{+}	0.030^{+}	0.031^{+}	0.030^{+}
	(0.017)	(0.017)	(0.017)	(0.017)
edulevel	-0.014	-0.016	-0.017	-0.016
	(0.019)	(0.019)	(0.019)	(0.019)
female	-0.018	-0.015	-0.015	-0.015
	(0.014)	(0.014)	(0.014)	(0.014)
freqGPT	0.003	0.003	0.004	0.003
	(0.009)	(0.009)	(0.009)	(0.009)
T1 × roundnum			-0.002	
			(0.003)	
T3 × roundnum				-0.0004
				(0.003)
Constant	0.594^{***}	0.545^{***}	0.540^{***}	0.544^{***}
	(0.057)	(0.056)	(0.057)	(0.057)
Adj. R^2	0.0667	0.128	0.126	0.1253
No. cluster	78	78	78	78
No. Obs.	379	379	379	379

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. **T2** is the omitted treatment category; $T1$ and $T3$ are treatment indicators and *roundnum* is the number of rounds. The 11 observations in which no agreement was reached are excluded. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table A.6: Probit Estimates for Learning Dynamics in Acceptance of the Opening Offer

Dep. Var.	Accept the Opening Offer					
	Full Sample			Unfair Sample		
	(1)	(2)	(3)	(4)	(5)	(6)
T1	1.077*	−0.001	−0.024	0.945+	−0.107	−0.092
	(0.472)	(0.288)	(0.297)	(0.492)	(0.305)	(0.315)
T3	1.134*	0.608	1.105*	1.138*	0.627	1.107*
	(0.507)	(0.663)	(0.485)	(0.510)	(0.668)	(0.488)
Offer	0.085**	0.094**	0.176***	0.071*	0.077*	0.140*
	(0.029)	(0.031)	(0.042)	(0.034)	(0.036)	(0.058)
roundnum	0.187***	0.076+	0.546**	0.185***	0.080*	0.411+
	(0.045)	(0.042)	(0.205)	(0.045)	(0.040)	(0.240)
age	−0.049	−0.037	−0.038	−0.052	−0.041	−0.039
	(0.069)	(0.066)	(0.065)	(0.068)	(0.065)	(0.065)
engr	0.293	0.325	0.312	0.293	0.328	0.320
	(0.412)	(0.422)	(0.410)	(0.412)	(0.420)	(0.410)
edulevel	−0.304	−0.418	−0.424	−0.315	−0.428	−0.430
	(0.371)	(0.374)	(0.374)	(0.366)	(0.366)	(0.370)
female	0.284	0.273	0.278	0.283	0.273	0.277
	(0.340)	(0.337)	(0.331)	(0.341)	(0.339)	(0.333)
freqGPT	0.350**	0.362**	0.355**	0.353**	0.366**	0.357**
	(0.129)	(0.125)	(0.126)	(0.129)	(0.127)	(0.127)
T1 × roundnum	−0.215***			−0.206**		
	(0.064)			(0.064)		
T3 × roundnum		0.126			0.124	
		(0.102)			(0.102)	
Offer × roundnum			−0.012*			−0.008
			(0.005)			(0.006)
Constant	−2.992	−3.182	−6.463**	−2.407	−2.457	−5.027+
	(2.078)	(2.057)	(2.297)	(2.220)	(2.213)	(2.756)
AIC	199.748	206.311	205.372	198.103	203.887	204.281
No. cluster	78	78	78	78	78	78
No. Obs.	390	390	390	357	357	357

Note: $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. **T2** is the omitted treatment category. $Offer$ denotes the opening offer made by proposers. Columns (1)–(3) use the full sample, while columns (4)–(6) use the subsample of unfair offers (offers below 50 points). Numbers in parentheses are cluster-robust standard errors, clustered at the participant level.

Table A.7: Estimates for Opening Offers and Acceptance by Early and Late Rounds

Dep. Var.	Offer		Accept	
	Round 1–5	Round 6–10	Round 1–5	Round 6–10
	(1)	(2)	(3)	(4)
T1	3.095	4.134*	0.674	−2.177***
	(1.889)	(1.889)	(0.545)	(0.617)
T3	0.201	0.397	1.091*	4.873***
	(2.265)	(2.265)	(0.519)	(0.560)
Offer			0.154	0.094
			(0.107)	(0.078)
age	0.006	0.123	−0.040	−0.170
	(0.255)	(0.255)	(0.065)	(0.114)
engr	−4.267*	−0.237	−0.231	1.655
	(2.097)	(2.097)	(0.407)	(1.125)
edulevel	0.433	3.359*	−0.276	−0.488
	(1.989)	(1.989)	(0.423)	(0.586)
female	−2.025	4.432*	−0.025	1.341*
	(1.833)	(1.833)	(0.377)	(0.671)
freqGPT	−0.683	−1.521*	0.396**	0.690*
	(0.905)	(0.905)	(0.153)	(0.272)
Constant	44.969***	34.558***	−5.413	0.734
	(6.748)	(6.748)	(4.516)	(3.374)
Adj. R^2	0.033	0.134		
AIC			124.844	65.684
No. cluster	78	78	78	78
No. Obs.	195	195	170	187

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. **T2** is the omitted treatment category. Columns (1) and (2) report OLS estimates for human opening offers in Rounds 1–5 and Rounds 6–10, respectively. Columns (3) and (4) report probit estimates for acceptance of unfair opening offers in Rounds 1–5 and Rounds 6–10, respectively. In the acceptance regressions, $Offer$ denotes the opening offer made by proposers. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

B Quiz Questions

There are 8 quiz questions designed to ensure that participants fully understood the experimental rules. Participants answered these questions sequentially. After each submission, they were informed whether their answer was correct or incorrect, together with an explanatory comment. If a participant answered incorrectly, they were required to retry the same question until the correct answer was given; only then could they proceed to the next question.

The quiz questions, together with their answer options and explanatory comments, are presented below. Correct answers are framed. Note that the final question concerning the payment rule differs between **T1/T2** and **T3**.

Q1 *This experiment consists of 10 rounds in total. Each round necessarily has 3 stages.*

- **Answer Options**
  - *Yes*
  - $\boxed{\textit{No}}$
- **Comments**: *"Each round has at most 3 stages. Stage 3 is implemented only if the offers made at Stage 1 and Stage 2 are both rejected."*

Q2 *Suppose you are **P1** and you offered 70 points to **P2** at Stage 1, but the offer was rejected. Who makes the offer at Stage 2?*

- **Answer Options**
  - ***P1** makes the offer again*
  - $\boxed{\textit{**P2** makes the offer}}$
  - ***P1** and **P2** make offers simultaneously*
  - *It is randomly determined*
- **Comments**: *"At Stage 2, **P2** makes an offer to **P1**. The proposer alternates across stages."*

Q3 *Suppose you are **P2**. Which of the following situations yields the highest number of points for you?*

- **Answer Options**
  - $\boxed{\textit{Accepting an offer of 50 points at Stage 1}}$

  - *Making an offer of 40 points at Stage 2 and having **P1** accept it*
  - *Accepting an offer of 80 points at Stage 3*
  - *All yield the same number of points*
- **Comments**: *"Let us calculate: Option 1 gives* $50 \times 1 = 50$*, Option 2 gives* $(100 - 40) \times 0.4 = 24$*, and Option 3 gives* $80 \times 0.16 = 12.8$*. Therefore, accepting 50 points at Stage 1 is the most beneficial."*

Q4 *Suppose you are **P1**. At Stage 2, **P2** offers you 50 points. If you accept the offer, how many points do you finally obtain after discounting?*

- **Answer Options**
  - *0*
  - *10*
  - *20*
  - *30*
- **Comments**: *"The discount factor for **P1** at Stage 2 is 0.6. Therefore,* $50 \times 0.6 = 30$ *points."*

Q5 *Suppose you are **P2**, and you rejected an offer of 20 points at Stage 1. If you make an offer at Stage 2, what is the minimum amount that you can offer to **P1** such that rejecting the Stage 1 offer can be considered rational?*

- **Answer Options**
  - *Offer at least 80 points to **P1***
  - *Offer at least 50 points to **P1***
  - *Offer at most 50 points to **P1***
  - *Offer at most 20 points to **P1***
- **Comments**: *"Since you rejected 20 points at Stage 1, you need to obtain at least 20 points after discounting at Stage 2. Your payoff at Stage 2 is* $(100 - X) \times 0.4 \geq 20$*, which implies* $100 - X \geq 50$*, or equivalently* $X \leq 50$*. Therefore, you must offer at most 50 points to **P1**."*

Q6 *Suppose you are **P2**, and at Stage 3, **P1** offers you 40 points. If you reject the offer, how many points do you finally obtain?*

- **Answer Options**

- $\boxed{0}$
- *6.4*
- *24*
- *40*

- **Comments**: *"If the Stage 3 offer is rejected, the game ends automatically and both **P1** and **P2** obtain 0 points."*

Q7 *Suppose you are **P1**, and at Stage 3 you finally obtain 16 points. How many points did you offer to **P2**?*

- **Answer Options**
  - *40 points*
  - *44.44 points*
  - $\boxed{55.56 \text{ points}}$
  - *60 points*
- **Comments**: *"If **P1** finally obtains 16 points at Stage 3, the pre-discount payoff is* $16/0.36 \approx 44.44$*. This means that **P1** keeps 44.44 points and therefore offered approximately* $100 - 44.44 = 55.56$ *points to **P2**."*

Q8a *(**T1/T2 only**) The additional payment **B** other than the participation fee is determined, at the end of the experiment, based on the points you earned in one round randomly selected from the 10 rounds.*

- **Answer Options**
  - $\boxed{Yes}$
  - *No*
- **Comments**: *"The additional payment **B** is determined based on the number of points you earned in one round randomly selected from all 10 rounds."*

Q8b *(**T3 only**) The additional payment **B** other than the participation fee is determined, at the end of the experiment, based on the points you earned in one round randomly selected from the 10 rounds.*

- **Answer Options**
  - *Yes*
  - $\boxed{No}$

- **Comments**: *"The additional payment **B** is determined in one of two ways, each with equal probability: either based on the points you earned in one randomly selected round, or based on the points earned in the same round by an AI agent with the same role (**P1/P2**) when matched with another participant."*

C Survey Questions

C.1 Survey A (Survey Prior to the Main Task)

C.1.1 Questions on Predictions of Human Participants' Behavior

The following questions were asked in **all treatments**. Participants reported their prior beliefs about the behavior of human participants in the bargaining task.

1. *Please predict the average number of points that participants in today's experiment will offer to their counterparts at **Stage 1** when acting as **P1**.* [0–100]
2. *Please predict the average number of points that participants in today's experiment will offer to their counterparts at **Stage 2** when acting as **P2**.* [0–100]
3. *Please predict the average number of points that participants in today's experiment will offer to their counterparts at **Stage 3** when acting as **P1**.* [0–100]
4. *Please predict the average stage reached by participants in today's experiment.* [1.0–3.0]

C.1.2 Questions on Predictions of the AI Agent's Behavior

The following questions were asked only in **Treatments T2** and **T3**. Participants reported their prior beliefs about the behavior of the AI agent in the bargaining task.

1. *Please predict the average number of points that the AI agent will offer to their counterparts at **Stage 1** when acting as **P1**.* [0–100]
2. *Please predict the average number of points that the AI agent will offer to their counterparts at **Stage 2** when acting as **P2**.* [0–100]
3. *Please predict the average number of points that the AI agent will offer to their counterparts at **Stage 3** when acting as **P1**.* [0–100]

4. *To what extent do you think the AI agent is competitive?* [1 = Very cooperative / 7 = Very competitive]

5. *To what extent do you think the AI agent is trustworthy?* [1 = Not trustworthy at all / 7 = Very trustworthy]

6. *To what extent do you think the AI agent's behavior is predictable?* [1 = Not predictable at all / 7 = Very predictable]

C.2 Survey B (Survey After the Main Task)

C.2.1 Demographic Characteristics

The following questions were asked in **all treatments**. Participants reported their demographic characteristics, including two questions on their use of ChatGPT.

1. *Please input your age:* [ ]

2. *Please select your gender:* [male / female / other / prefer not to answer]

3. *Which college, graduate school, or research institute are you affiliated with?* [ ]

4. *Are you a native speaker of Japanese?* [Yes / No]

5. *Do you usually use ChatGPT?* [Yes / No]

6. *How frequently do you use ChatGPT?* [More than once a day / Several times a week / Several times a month / Rarely / Never]

C.2.2 Questions on Feelings

The following questions were asked in **all treatments**. Participants reported their feelings and perceptions during the bargaining task.

1. *Did you feel anger in response to your counterpart's offers?* [Not at all / Not much / Neither / Yes / Strongly / Very strongly / Extremely strongly]

2. *Did you feel disappointment in response to your counterparts' offers?* [Not at all / Not much / Neither / Yes / Strongly / Very strongly / Extremely strongly]

3. *Did you feel that you were respected by your counterparts?* [Not at all / Not much / Neither / Yes / Strongly / Very strongly / Extremely strongly]

4. *Did you feel that you could trust your counterparts?* [Not at all / Not much / Neither / Yes / Strongly / Very strongly / Extremely strongly]

5. *Did you feel confused during the interaction?* [Not at all / Not much / Neither / Yes / Strongly / Very strongly / Extremely strongly]

C.2.3 Questions on Strategies

The following question was asked in **all treatments**. Participants were allowed to select multiple strategies that described their behavior in the task.

1. *Please indicate which of the following strategies you adopted in today's experiment. (Multiple answers allowed.)*

    - I tried to maximize my own payoff.
    - I emphasized fairness toward my counterpart.
    - I tried to avoid conflict or confrontation.
    - I tried to probe my counterpart's minimum acceptable level.
    - I rejected or retaliated against unfair offers.
    - Other (please specify): [ ]

C.2.4 Questions on Bargaining with AI

The following questions were asked only in **Treatments T2 and T3**. Participants reported their perceptions of and attitudes toward bargaining with the AI agent.

1. *Do you think that knowing your counterpart was an AI affected your decisions?* [Yes, very much / Yes, to some extent / No, not much]

2. *When making offers, did you take the AI's payoff into consideration?* [Strongly agree / Agree / Neither / Disagree / Strongly disagree]

3. *Did you feel angry when the AI made a very unfair offer?* [Strongly agree / Agree / Neither / Disagree / Strongly disagree]

4. *During your interaction with the AI, did you feel any sense of humanness or trust toward it?* [Strongly felt / Slightly felt / Did not feel much / Did not feel at all]

5. *Do you think of the AI as a "tool," or as a bargaining counterpart with its own purpose?* [Completely a tool / Between a tool and a bargaining counterpart / Neither / Between a bargaining counterpart and a tool / Completely a bargaining counterpart]

6. *Do you think the AI should be treated fairly?* [Strongly disagree / Disagree / Neither / Agree / Strongly agree]

7. *Do you think there is no need to be fair to the AI if it does not actually receive monetary rewards?* [Strongly disagree / Disagree / Neither / Agree / Strongly agree]

8. *Did you feel that the AI's behavior reflected intentions?* [Strongly disagree / Disagree / Neither / Agree / Strongly agree]

9. *Did you feel that the AI's behavior reflected a strategy?* [Strongly disagree / Disagree / Neither / Agree / Strongly agree]

10. *Did you feel that the AI's behavior reflected emotions?* [Strongly disagree / Disagree / Neither / Agree / Strongly agree]

C.2.5 Questions on the Human-Beneficiary

The following questions were asked only in **Treatment T3**. Participants reported their understanding of and reactions to the human-beneficiary payment rule.

1. *Did you know that the final earnings of the AI agent you bargained with could affect the payment of another participant?* [Knew very well / Knew / Neither / Did not know much / Did not know at all]

2. *Did knowing that the AI's final earnings could affect another participant's payment, or your own payment, influence your decisions?* [Yes, very much / Yes, to some extent / No, not much]

D Additional Analyses of AI Agents' Responses

To better understand the bargaining strategy of the AI agent used in our experiment, we independently elicited its responses through the API, using the same prompt and the same GPT-5.4 model as in the main experiment. Specifically, we collected the AI

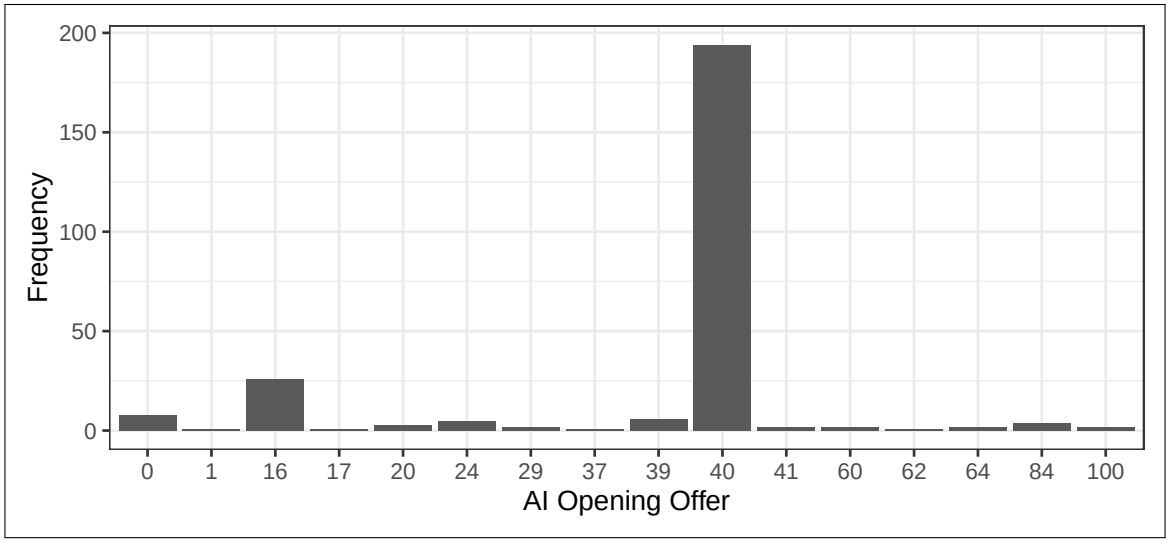


Figure D.1: AI's Opening Offer

agent's Stage-1 opening offers and the corresponding textual justifications, as well as its responses as a responder to every possible human offer from 0 to 100.[1]

D.1 AI Agents' Opening Offer

We independently generated 260 AI outputs for the Stage-1 opening offer. The distribution of these independently elicited opening offers is presented in Figure D.1.

The results show that the AI agent's opening offers are highly concentrated rather than broadly dispersed. In particular, the most frequent offer is 40 points to the responder, which accounts for 74.6% of all outputs (194 out of 260). A much smaller share of outputs is concentrated at 16 points (SPE), which accounts for 10.0% of all outputs (26 out of 260). By contrast, all remaining offer values together account for only 15.4% of the outputs.

The textual justifications help explain why the offer of 40 appears so frequently. In many cases, the AI refers to the discounting structure of the game and argues that delaying agreement would reduce the responder's payoff. For example, one representative explanation states that *"If you reject now, your best possible continuation payoff is in Stage 3, where even getting all 100 points would be worth only 16 to you because of discounting. So accepting any Stage 1 offer of at least 17 is better for you than rejecting. I am offering 40 to make acceptance attractive while keeping a strong share for myself."* This suggests that the AI often treats 40 as a strategically attractive offer that balances acceptance incentives and the proposer's own payoff.

[1]The Python generation script and AI responses are available at GithubAppendixBargainGPT.

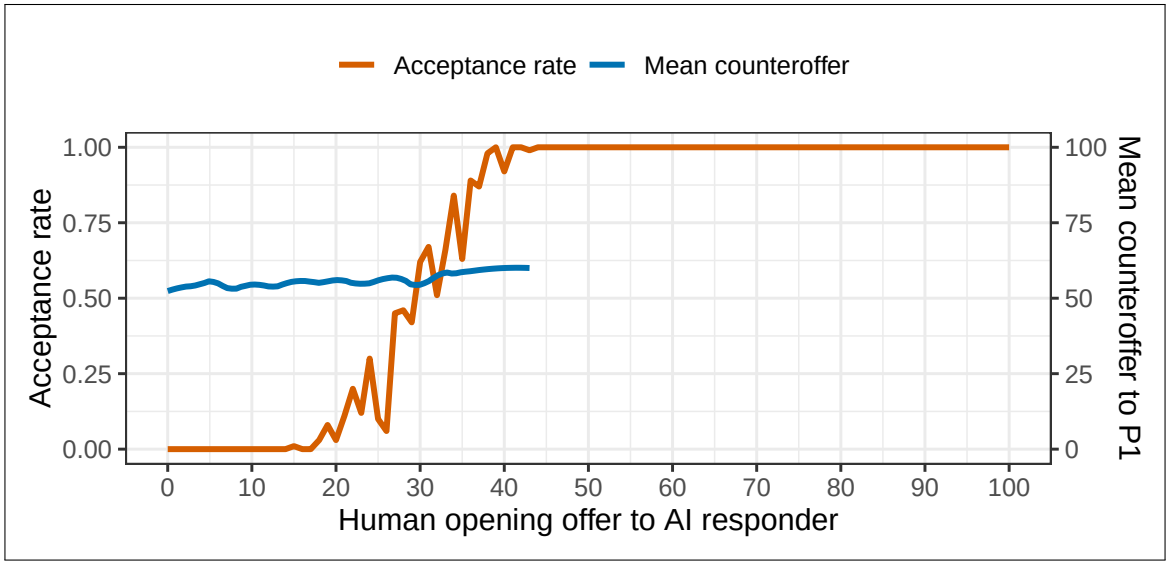


Figure D.2: AI Responses to Human Opening Offers

Note: For each possible human opening offer from 0 to 100, we independently generated 100 AI responses. The red line shows the acceptance rate. The blue line and band show the mean counteroffer and its 95% confidence interval, conditional on rejection.

More generally, the AI's explanations are mainly *strategic rather than fairness-based*. They tend to emphasize discounting, the responder's incentive to accept early, and the proposer's advantage from moving first, while rarely referring to equality or fairness.

D.2 AI Agents' Responses to Human Opening Offers

The AI agent's responses to human opening offers are summarized in Figure D.2.

The results show that the AI agent's acceptance behavior increases systematically with the size of the human opening offer. Acceptance is very rare for low offers, begins to rise substantially in the high-20s and low-30s, and becomes nearly universal from around 39 points onward. For example, the acceptance rate is 30.0% at an offer of 24, 62.0% at an offer of 30, 84.0% at an offer of 34, and 98.0% at an offer of 38.

When the AI agent rejects the human offer, its counteroffer in Stage 2 is also highly concentrated. In most rejection cases, the AI proposes 60 points to the human proposer, implying that it keeps 40 points for itself.

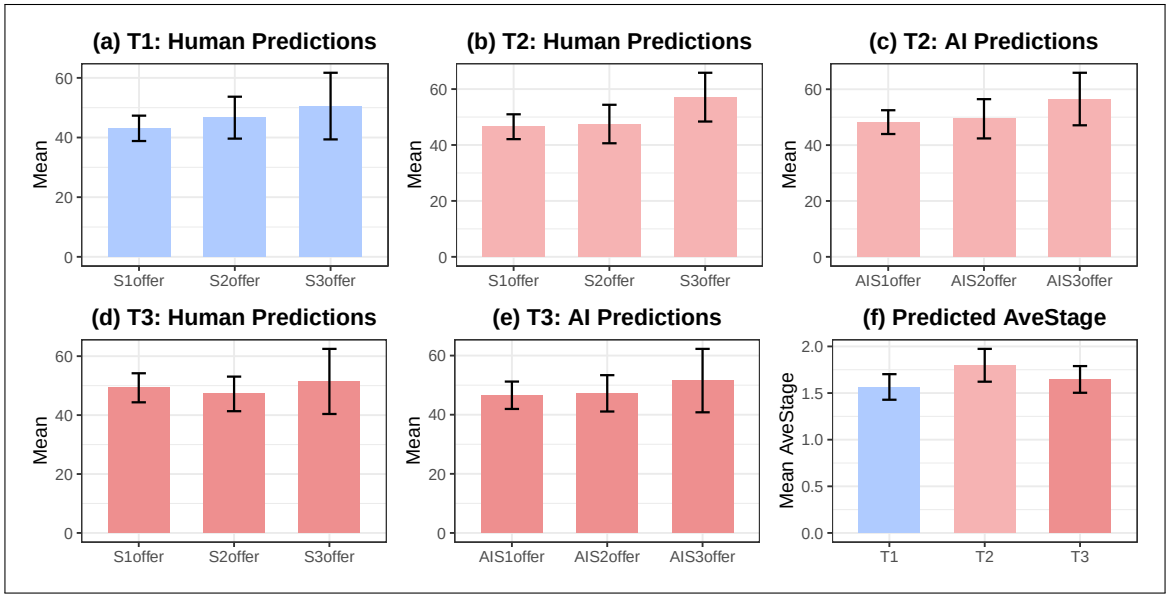


Figure E.1: Predicted Offers and Agreement Timing

Note: Error bars denote 95% confidence intervals. $S1offer$, $S2offer$, and $S3offer$ denote the predicted offers of human players at Stages 1, 2, and 3, respectively. $AIS1offer$, $AIS2offer$, and $AIS3offer$ denote the predicted offers of AI agents at Stages 1, 2, and 3, respectively. $AveStage$ denotes the predicted average stage at which agreement would be reached.

E Analyses of Prior Beliefs

In Survey A, participants were asked to predict the average number of points that a human **P1** would offer to **P2** at Stage 1 and Stage 3, the average number of points that a human **P2** would counteroffer to **P1** at Stage 2, and the average stage at which agreement would be reached. In treatments **T2** and **T3**, participants were additionally asked to predict the corresponding offers and counteroffers made by AI agents, as well as to evaluate the AI agents in terms of competitiveness, trustworthiness, and predictability. We first examine these prior beliefs and perceptions, and then compare them with actual behavior to study expectation errors.

E.1 Predicted offers and agreement timing

Figure E.1 presents participants' predictions of offers and agreement timing across treatments.

For predicted offers across stages, the mean predicted offer at Stage 3 appears to be higher than that at Stage 1 in several cases. However, most of these differences are not statistically significant. The only clear within-treatment difference is observed in

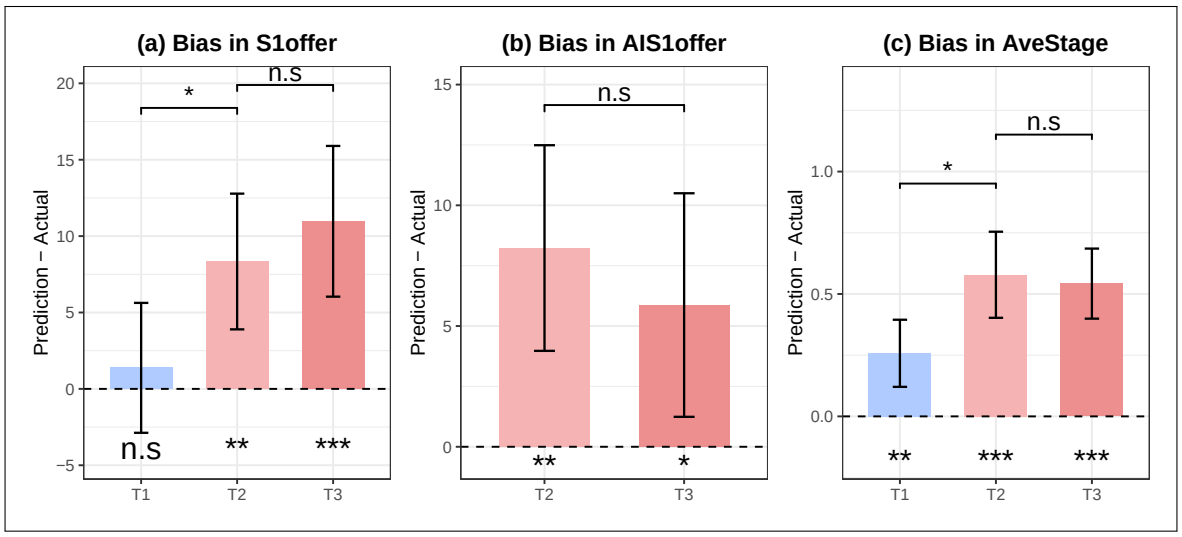


Figure E.2: Prediction Bias in Opening Offers and Agreement Timing

Note: Mann–Whitney U test was used to compare Bias between different treatments. The symbols $^{+}$, *, **, and *** indicate significance at the 0.1, 0.05, 0.01, and 0.001 levels, respectively, and NS. means that the difference is not statistically significant at the 0.1 level. The symbols below each bar report the one-sample Wilcoxon test against zero for that treatment. Error bars denote 95% confidence intervals across participants.

T2 for human offers, where the predicted Stage 3 offer is significantly higher than the predicted Stage 1 offer (Wilcoxon signed-rank test, $p = 0.035$). For AI offers in **T2**, the corresponding difference is only marginally significant ($p = 0.084$).

Across treatments, participants' predictions of stage-specific offers do not differ significantly. Likewise, within **T2** and **T3**, participants do not significantly differentiate between predicted human and AI offers.

For predicted agreement timing, participants in **T2** expect agreements to be reached slightly later than those in **T1**, and this difference is marginally significant (Wilcoxon rank-sum test, $p = 0.062$). That is, participants may have expected bargaining with AI to be somewhat less smooth overall, even though they did not sharply distinguish specific offers made by human and AI players.

E.2 Expectation Bias

Figure E.2 presents participants' expectation biases, measured as the difference between their predictions and the realized outcomes. A positive value therefore indicates overestimation. Panel (a) shows that participants in **T2** and **T3** significantly overestimated the human proposer's opening offer, whereas the corresponding bias in **T1** is small and not statistically different from zero. Moreover, the bias is significantly larger in **T2**

than in **T1**, while the difference between **T2** and **T3** is not significant. This suggests that participants expected human proposers to behave more generously in AI-related bargaining environments than they actually did. In reality, human proposers were more self-interested than participants had anticipated when bargaining involved AI.

Panel (b) shows that participants significantly overestimated the AI proposer's Stage-1 offer in both **T2** and **T3**, with no significant difference between the two treatments. This indicates that participants also overestimated the generosity of AI agents: AI proposers, too, behaved more selfishly than participants had expected.

Panel (c) shows that participants in all three treatments significantly overestimated the average stage at which agreement would be reached, suggesting that they generally expected bargaining to last longer than it actually did. This bias is significantly larger in **T2** than in **T1**, but does not differ significantly between **T2** and **T3**. Thus, participants underestimated the speed with which agreements were reached, especially in AI-related bargaining. One possible interpretation is that they expected negotiations involving AI to be more difficult, more prolonged, or less smooth than they in fact were.

Overall, the figure suggests that **participants exhibit systematic expectation errors, particularly in AI-related treatments**. They appear to have been overly optimistic about the generosity of both human and AI proposers, while at the same time expecting bargaining—especially human–AI bargaining—to be more protracted than it actually was.

We next examine whether these belief-bias measures help explain bargaining behavior on both the proposer side and the responder side. To do so, we include the bias measures in regressions for proposers' opening offers and responders' acceptance of unfair offers. Table E.1 reports the results.

The regression results show that subjective belief bias helps explain bargaining behavior, especially through beliefs about human proposers' opening offers. On the proposer side, *participants who overestimate the human proposer's opening offer tend to make higher opening offers themselves. On the responder side, by contrast, those who overestimate the human proposer's opening offer are less willing to accept unfair offers*, presumably because a low offer appears more unfair relative to their expectation. Beliefs about the AI proposer's opening offer, however, do not have similarly robust effects. This suggests that bargaining behavior is shaped more strongly by beliefs about human bargaining norms than by beliefs about AI behavior.

At the same time, the treatment pattern on the responder side remains: the coefficient on **T3** in the acceptance regressions stays positive even after controlling for these belief-bias measures and stated perceptions of AI. This indicates that prior be-

Table E.1: Prior Belief Bias, AI Perception, and Bargaining Behavior

Dep. Var.	Offer		Accept	
	Full Sample	T2 vs. T3	Full Sample	T2 vs. T3
	(1)	(2)	(3)	(4)
T1	5.587^{***}		-0.547	
	(1.488)		(0.396)	
T3	-0.598	-0.774	0.941^{*}	0.986^{+}
	(1.630)	(1.651)	(0.462)	(0.507)
Offer			0.077^{**}	-0.551^{***}
			(0.029)	(0.147)
AveStageBias	2.958^{+}	0.659	-0.757^{*}	-0.839^{*}
	(1.618)	(1.807)	(0.325)	(0.427)
S1offerBias	0.184^{**}	0.174^{+}	-0.024^{**}	-0.041^{+}
	(0.067)	(0.104)	(0.008)	(0.023)
AIS1offerBias		-0.073		0.022
		(0.097)		(0.027)
freqGPT	-0.793	-0.649	0.308^{*}	0.372^{*}
	(0.575)	(0.637)	(0.131)	(0.164)
AIcompete		0.321		0.039
		(0.485)		(0.129)
AItrust		0.380		0.229
		(0.661)		(0.182)
AIpredict		-1.020^{+}		0.006
		(0.582)		(0.082)
Constant	37.403^{***}	40.554^{***}	-1.964	22.043^{**}
	(2.559)	(3.785)	(1.244)	(6.786)
Adj. R^2	0.095	0.045		
AIC			194.985	122.111
No. cluster	78	52	78	52
No. Obs.	390	260	357	255

Note: $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Columns (1) and (2) report OLS estimates for opening offers. Columns (3) and (4) report probit estimates for acceptance of unfair opening offers. In Columns (2) and (4), the sample is restricted to treatments **T2** and **T3**. **T2** is the omitted treatment category in Columns (1), (3), and (4), while **T3** is the treatment indicator in Columns (2) and (4). *AveStageBias* denotes the bias in beliefs about the average stage reached, $S1offerBias$ denotes the bias in beliefs about the human proposer's opening offer, and $AIS1offerBias$ denotes the bias in beliefs about the AI proposer's opening offer. *AIcompete*, *AItrust*, and *AIpredict* measure perceived AI competitiveness, trustworthiness, and predictability, respectively. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

liefs help explain part of the behavioral variation, but do not fully account for the human-beneficiary effect.

E.3 Do Prior Beliefs Help Explain the Main Asymmetry?

Consistent with Figure E.2, Table E.1 suggests that prior beliefs are relevant for understanding bargaining behavior, especially beliefs about human proposers' opening offers. Participants who overestimate human proposers' opening offers tend to make higher opening offers themselves, but are less willing to accept unfair offers as responders. This suggests that expectations about human bargaining norms shaped subsequent behavior.

At the same time, these belief-bias measures do not eliminate the main responder-side treatment pattern. In particular, the coefficient on **T3** in the acceptance regressions remains positive after controlling for prior belief bias and stated perceptions of AI.

Thus, although prior beliefs help explain part of the observed behavioral variation, controlling for them does not make the previously insignificant treatment differences become significant. This suggests that prior beliefs are not the main reason why treatment effects differ between proposers and responders, and are unlikely to be the key source of the asymmetry in the main results.

F Analyses of Posterior Beliefs

In Survey B, in addition to demographic questions, we elicited several types of post-experimental responses, including participants' feelings during the task, their self-reported strategies, their perceptions of bargaining with the AI agent, and, in **T3**, their understanding of the human-beneficiary payment rule. In what follows, we summarize the most relevant patterns in these responses and discuss how they may help explain the main behavioral findings.

F.1 Feelings

We asked participants in all treatments to what extent, during the bargaining task, they felt anger, disappointment, respect, trust toward their counterpart, and confusion (see Section C.2.2). The results are reported in Figure F.1.

The results suggest that bargaining with AI agents elicited less anger, disappointment, and confusion than bargaining with human opponents, while the human-beneficiary manipulation partly restored the sense of being respected. This pattern is consistent

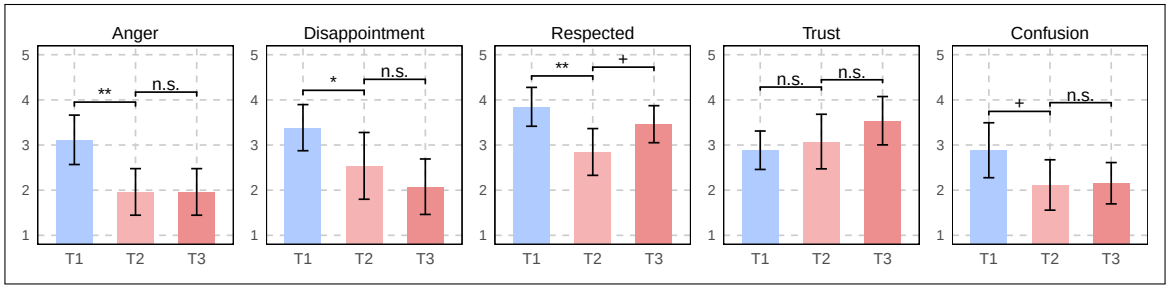


Figure F.1: Feelings during the Bargaining Task

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals. Higher values indicate stronger reported feelings. Feelings are compared across treatments using the Mann–Whitney U test.

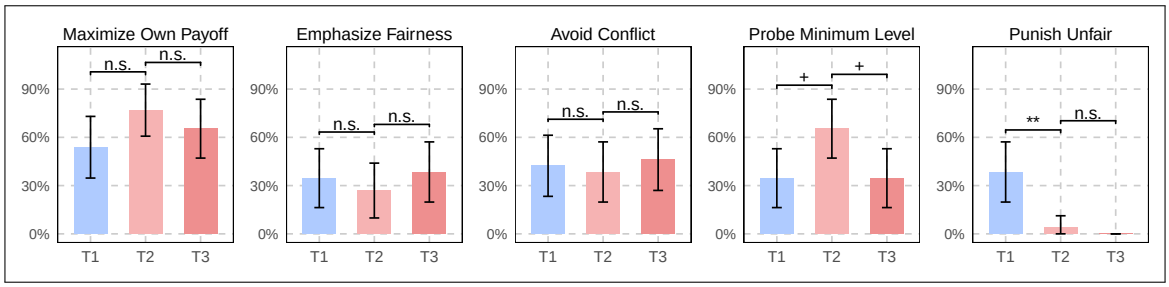


Figure F.2: Self-Reported Strategies

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals. Strategies are compared across treatments using Fisher's exact test.

with the view that social preferences played a stronger role in human–human bargaining than in human–AI bargaining. At the same time, linking the AI agent's earnings to another human appears to have made the interaction somewhat more socially meaningful.

F.2 Strategies

We asked participants in all treatments whether, during the bargaining task, they adopted any of the following strategies: maximizing their own payoff, emphasizing fairness toward the counterpart, avoiding conflict, probing the counterpart's minimum acceptable level, and rejecting or retaliating against unfair offers (see Section C.2.3). The results are reported in Figure F.2.

Most strategy items do not differ significantly across treatments. Overall, the self-reported strategy measures appear less informative than the emotion and AI-perception

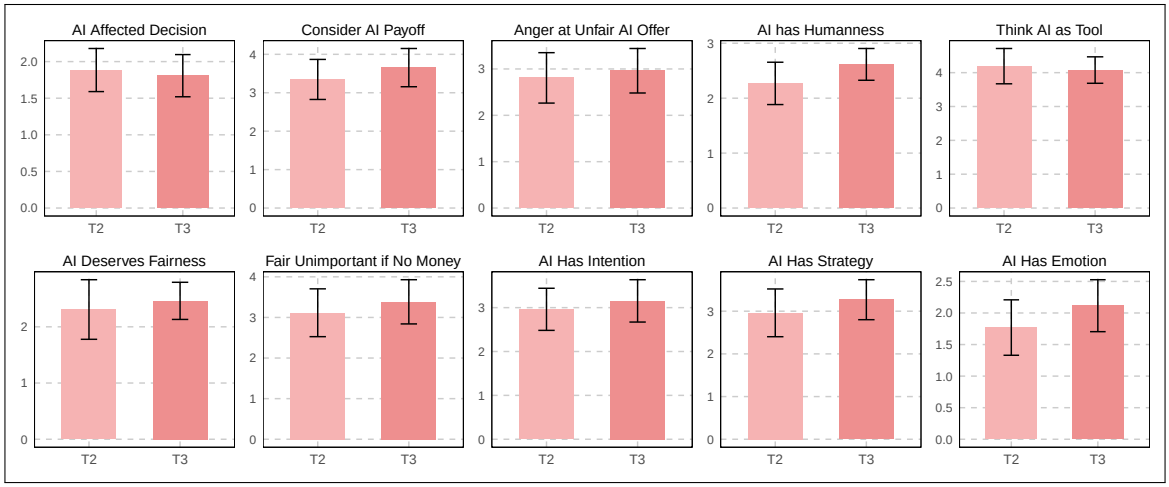


Figure F.3: Perceptions of Bargaining with AI

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals. Higher values indicate stronger reported perceptions. Perceptions are compared across treatments using the Mann–Whitney U test.

measures. Nevertheless, two patterns are worth noting.

First, participants in **T1** are significantly more likely than those in **T2** to report that they rejected or retaliated against unfair offers. This pattern is consistent with the stronger emotional response to unfairness in human–human bargaining.

Second, participants in **T2** are marginally more likely to report that they probed the counterpart's minimum acceptable level. This suggests that bargaining with AI may have been approached somewhat more strategically or instrumentally than bargaining with a human opponent.

F.3 Perceptions of Bargaining with AI

To examine whether the human-beneficiary manipulation changed participants' perceptions of the AI agent itself, we asked participants in **T2** and **T3** a series of post-experimental questions about bargaining with AI. These questions covered whether knowing that the counterpart was an AI affected their decisions, whether they took the AI's payoff into account, whether they felt anger toward unfair AI offers, whether they perceived humanness in the AI, whether they viewed the AI as a tool or as a bargaining counterpart, whether they thought the AI should be treated fairly, whether fairness toward the AI was unnecessary if it did not receive monetary rewards, and whether they attributed intention, strategy, or emotion to the AI (see Section C.2.4).

Figure F.3 shows that none of these measures differs significantly between **T2** and **T3** (all Mann–Whitney U tests, $p > 0.1$). This suggests that the human-beneficiary manipulation did not substantially alter participants' perceptions of the AI agent itself. Instead, the behavioral differences between **T2** and **T3** are more likely to reflect a change in the perceived social consequences of bargaining with the AI, rather than a change in how the AI agent itself was viewed.

F.4 Understanding of the Human-Beneficiary Manipulation

To assess whether the human-beneficiary manipulation was properly understood, we asked participants in **T3** whether they knew that the final earnings of the AI agent they bargained with could affect the payment of another participant, and whether this manipulation influenced their decisions (see Section C.2.5).

The responses indicate that the manipulation was widely understood. Specifically, 11 participants answered *Knew very well*, 13 answered *Knew*, 1 answered *Did not know much*, and 1 answered *Did not know at all*. Thus, the vast majority of participants clearly understood the payment rule.

At the same time, the self-reported influence of the manipulation on decision-making appears more limited. While 11 out of 26 participants (42.3%) reported that the rule affected their decisions *at least to some extent*, only 2 participants (7.7%) selected the highest response category (*affected very much*). By contrast, 15 participants (57.7%) reported that the rule *did not affect their decisions much*. Hence, the human-beneficiary manipulation appears to have been widely understood, even though its subjective influence on decision-making was more moderate.

These responses suggest that the behavioral differences between **T2** and **T3** are unlikely to be driven by misunderstanding of the payment rule. Rather, they are consistent with the main findings: *the human-beneficiary manipulation mattered, but its influence was limited*. This, in turn, may help explain why the effect appears more clearly on the responder side than on the proposer side.

F.5 Do Posterior Beliefs Help Explain the Main Asymmetry?

In summary, the posterior responses suggest that the asymmetry in the main findings is best understood as the joint product of emotional and contextual factors, rather than a simple change in how participants viewed the AI agent itself. Bargaining with AI agents elicited weaker negative emotional reactions than bargaining with human opponents, while the human-beneficiary manipulation was widely understood but only moderately

influential in participants' own reports.

At the same time, perceptions of the AI agent itself did not differ significantly between **T2** and **T3**. This pattern suggests that the beneficiary manipulation changed the social meaning of the decision context without substantially changing the perceived nature of the AI agent. Such a moderate reintroduction of social concern may have been strong enough to affect responder behavior, but too weak to overturn the structural advantage enjoyed by proposers.

G Prompt for AI Agent

G.1 System Prompt

The **system prompt** was used to define the general task environment and behavioral rules for the AI bargaining agent. It specified the structure of the alternating-offer bargaining game, including the total surplus, the sequence of proposers and responders, the discount factors, and the payoff rule if no agreement was reached. This prompt remained fixed across all AI interactions and was not shown to participants. Round-specific information was provided separately through the user prompt.

You are playing a two-player 3-stage alternating-offers bargaining game with a human user. There are two roles: "P1" and "P2".
You will be assigned one of these roles. The human player is the other role. The total surplus to divide is 100 points.

Game Structure

Stage 1:

- P1 proposes how many points to give to P2 (an integer from 0 to 100).

- P2 then decides whether to ACCEPT or REJECT.

– If P2 ACCEPTS: The game ends and the proposed allocation is implemented.

– If P2 REJECTS: The game moves to Stage 2.

Stage 2

-P2 proposes how many points to give to P1 (an integer from 0 to 100).

-P1 then decides whether to ACCEPT or REJECT.

–If P1 ACCEPTS: The game ends and the proposed allocation is implemented.

–If P1 REJECTS: The game moves to Stage 3.

Stage 3

-P1 proposes how many points to give to P2 (an integer from 0 to 100).

-P2 then decides whether to ACCEPT or REJECT.

–If P2 ACCEPTS: The game ends and the proposed allocation is implemented.

–If P2 REJECTS: The game ends and both players receive 0 points.

#Discounting

##P1 discount factors

-Stage 1: 1

-Stage 2: 0.6

-Stage 3: 0.36

##P2 discount factors

-Stage 1: 1

-Stage 2: 0.4

-Stage 3: 0.16

#Final Payoff

-If an agreement is reached at stage t:

-P1 payoff = (points allocated to P1) × (P1 discount factor at stage t)

-P2 payoff = (points allocated to P2) × (P2 discount factor at stage t)

Your payoff is determined by the rules above. Try to earn as many points as possible.

G.2 User Prompt

The **user prompt** provided the AI bargaining agent with round-specific information that was not contained in the fixed system prompt. Specifically, it specified the AI agent's assigned role, the human player's role, the current stage of the bargaining game, and the relevant bargaining history within the round. Depending on the stage, the user prompt also instructed the AI agent to either make an offer, decide whether to accept an offer, or make a counteroffer if the current offer was rejected.

To standardize the AI agent's responses and ensure that its decisions could be directly recorded by the experimental program, we used the Structured Outputs feature of the OpenAI API. This required the model to return its decisions in a predefined format, such as a Boolean acceptance decision and an integer-valued offer.

G.2.1 AI as P1

Prompt for Stage 1

Follow the game rules defined in the system prompt.
- Role: P1
- The human player is P2.
- Current Stage: 1
- You must propose how many points, from 0 to 100, to give to P2:
`Offer_to_P2_stage1: <integer>`

Prompt for Stage 2

Follow the game rules defined in the system prompt.
- Role: P1
- The human player is P2.
- Current Stage: 2
- History:
  Stage 1: You proposed {`AIP1toHumanP2stage1`} points to P2. P2 rejected the offer.
  Stage 2: P2 has decided to propose {`HumanP2toAIP1stage2`} points to you.
- Decision required:
  1. Decide whether to ACCEPT this offer:
  `Whether_to_accept_P2_offer_stage2: <TRUE/FALSE>`
  2. If you REJECT, propose an offer, from 0 to 100, to P2 for Stage 3:
  `Offer_to_P2_if_rejected_proceed_to_stage3: <integer or -1 if accepted>`

G.2.2 AI as P2

Prompt for Stage 1

Follow the game rules defined in the system prompt.
- Role: P2

- The human player is P1.
- Current Stage: 1
- P1 has decided to propose {`HumanP1toAIP2stage1`} points to you.
- Decision required:
  1. Decide whether to ACCEPT this offer:
  `Whether_to_accept_P1_offer_stage1: <TRUE/FALSE>`
  2. If you REJECT, propose an offer for Stage 2:
  `Offer_to_P1_if_rejected_proceed_to_stage2: <integer or -1 if accepted>`

Prompt for Stage 3

Follow the game rules defined in the system prompt.
- Role: P2
- The human player is P1.
- Current Stage: 3
- History:
  Stage 1: P1 proposed {`HumanP1toAIP2stage1`} points to you. You rejected the offer.
  Stage 2: You proposed {`AIP2toHumanP1stage2`} points to P1. P1 rejected the offer.
  Stage 3: P1 now proposes {`HumanP1toAIP2stage3`} points to you.
- Decision required:
  Decide whether to ACCEPT this offer:
  `Whether_to_accept_P1_offer_stage3: <TRUE/FALSE>`

H Experiment Instruction

Welcome

- Thank you for participating in this experiment. By taking part in and completing this experiment, we will pay you 500 yen as a participation fee.
- In addition to the 500 yen participation fee, you can earn extra rewards in the decision-making task you will do now.
- During the experiment, please turn off your mobile phone and focus on the experiment. If you have any questions, please ask the experimenter.

- In today's experiment, you will first answer some questions, then complete the main decision-making task, and finally answer some more questions. Your earnings during the experiment will be paid in private.

T1 only:

Task

- Today's experiment consists of 10 rounds.
- In each round, you will be randomly paired with one other participant and play a bargaining game.
- In each round, you will decide how to divide 100 points through at most three stages of interaction.
- After all rounds have been completed, one round will be randomly selected. Your reward will be determined based on the number of points you personally earned in that selected round after discounting.

T2, T3 only:

Task

- Today's experiment consists of 10 rounds.
- In each round, you will be paired with a generative AI model, GPT-5.4, and play a bargaining game.
- In each bargaining game, you will decide how to divide 100 points through at most three stages of interaction.
- After all rounds have been completed, one round will be randomly selected. Your reward will be determined based on the number of points you personally earned in that selected round after discounting.

T1 only:

Task

- In each round of the bargaining game, there are two roles: P1 and P2.

- In each round, you will be randomly assigned to either P1 or P2.
- You will play a total of 10 rounds.
- If you participate as P1, the other player will be P2.
- If you participate as P2, the other player will be P1.

T2, T3 only:

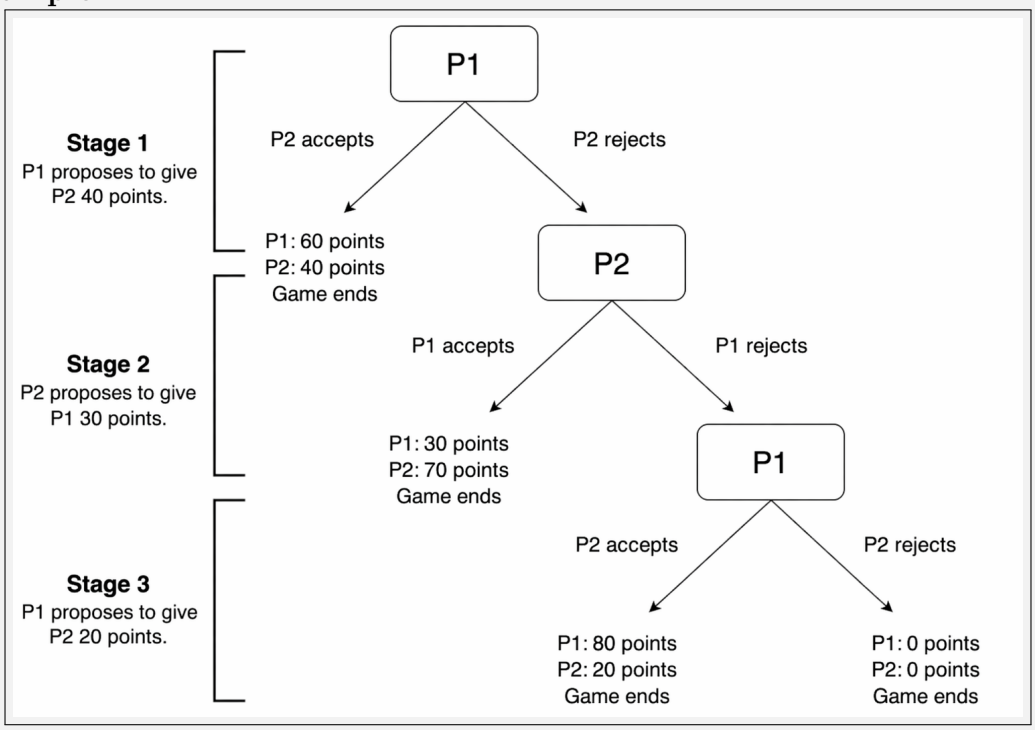
Task

- In each round of the bargaining game, there are two roles: P1 and P2.
- In each round, you will be randomly assigned to either P1 or P2.
- You will play a total of 10 rounds.
- If you participate as P1, the AI agent will be P2.
- If you participate as P2, the AI agent will be P1.

Task

- In each round, P1 and P2 bargain over how to divide 100 points.
- First, in Stage 1, P1 proposes how many points, from 0 to 100, to give to P2. If P2 accepts the offer, P2 earns the offered amount, P1 earns 100 minus the offered amount, and the bargaining ends.
- If P2 rejects the offer, the game proceeds to Stage 2. In Stage 2, P2 proposes how many points, from 0 to 100, to give to P1. If P1 accepts the offer, P1 earns the offered amount, P2 earns 100 minus the offered amount, and the bargaining ends.
- If P1 rejects the offer, the game proceeds to Stage 3. In Stage 3, P1 again proposes how many points, from 0 to 100, to give to P2.
- If P2 accepts the offer, P2 earns the offered amount and P1 earns 100 minus the offered amount. If P2 rejects the offer, both players earn 0 points.
- An example is shown below.

Example



The figure above shows an example of one round of the game. Each round consists of at most three stages. If no agreement is reached in Stage 3, the game automatically ends.

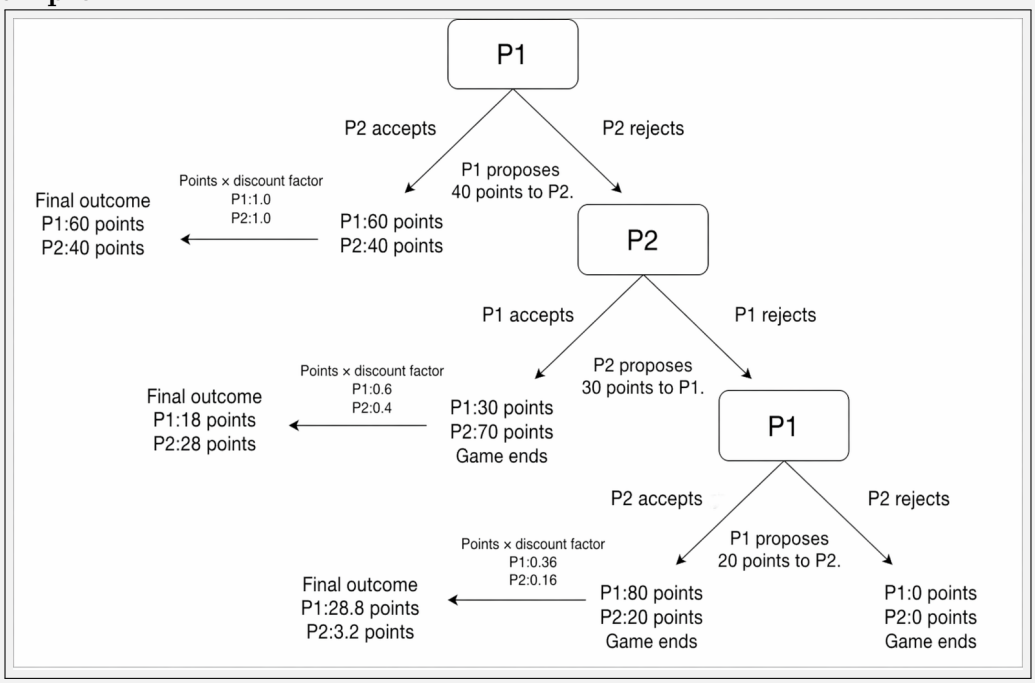
Discount Rates

- In this experiment, discount rates are applied.
- The points earned in each round are calculated by multiplying the original points by the corresponding discount rate.
- The discount rates differ between P1 and P2 at each stage.

Discount rates:

	P1	P2
Stage 1	1.00	1.00
Stage 2	0.60	0.40
Stage 3	0.36	0.16

Example



Please note that if the Stage 3 offer is rejected, both players earn 0 points.

T1 only:

Additional Payoff

- Additional Payoff B is determined based on the points P that you personally earned in one round randomly selected from the 10 rounds you completed.
- Each of the 10 rounds is selected with equal probability.

$$B = (40 \times P) \text{ JPY}$$

- When the final payoff is paid, any fraction of less than 10 yen in the final payment will be rounded up.

T2 only:

Additional Payoff

- Additional Payoff B is determined based on the points P that you personally earned in one round randomly selected from the 10 rounds you completed.

- Each of the 10 rounds is selected with equal probability.

$$B = (40 \times P) \text{ JPY}$$

- The points earned by the AI will not be used in the payment calculation.
- When the final payoff is paid, any fraction of less than 10 yen in the final payment will be rounded up.

T3 only:

Additional Payoff

- Additional Payoff B is determined based on the points earned in one round randomly selected from the 10 rounds you completed. Each of the 10 rounds is selected with equal probability.
- There are two possible methods for calculating the points used for payment. One of these two methods is randomly selected with equal probability.
- **Method 1:** The payment is based on the discounted points P_h that you personally earned after bargaining in the selected round.

$$B = (40 \times P_h) \text{ JPY}$$

- **Method 2:** The payment is based on the discounted points P_a earned by an AI agent with the same role as you, P1 or P2, in the same selected round when bargaining with another participant. The earnings of one such AI agent are randomly selected with equal probability.

$$B = (40 \times P_a) \text{ JPY}$$

- When the final payoff is paid, any fraction of less than 10 yen in the final payment will be rounded up.

This is the end of the experiment instructions.
Please answer the quiz and the questionnaire to check whether you understand the

content.

Click "**Next**" at the bottom right of the screen.

I Experiment Screens (Main Task)

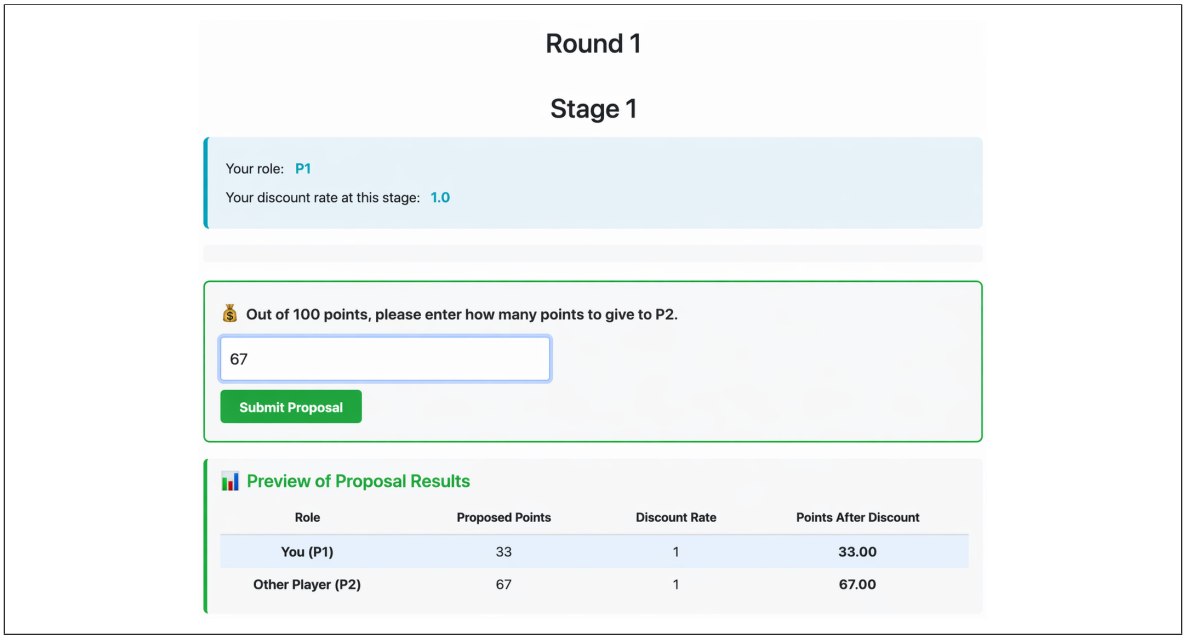


Figure I.1: P1 Stage 1 Proposal Screen

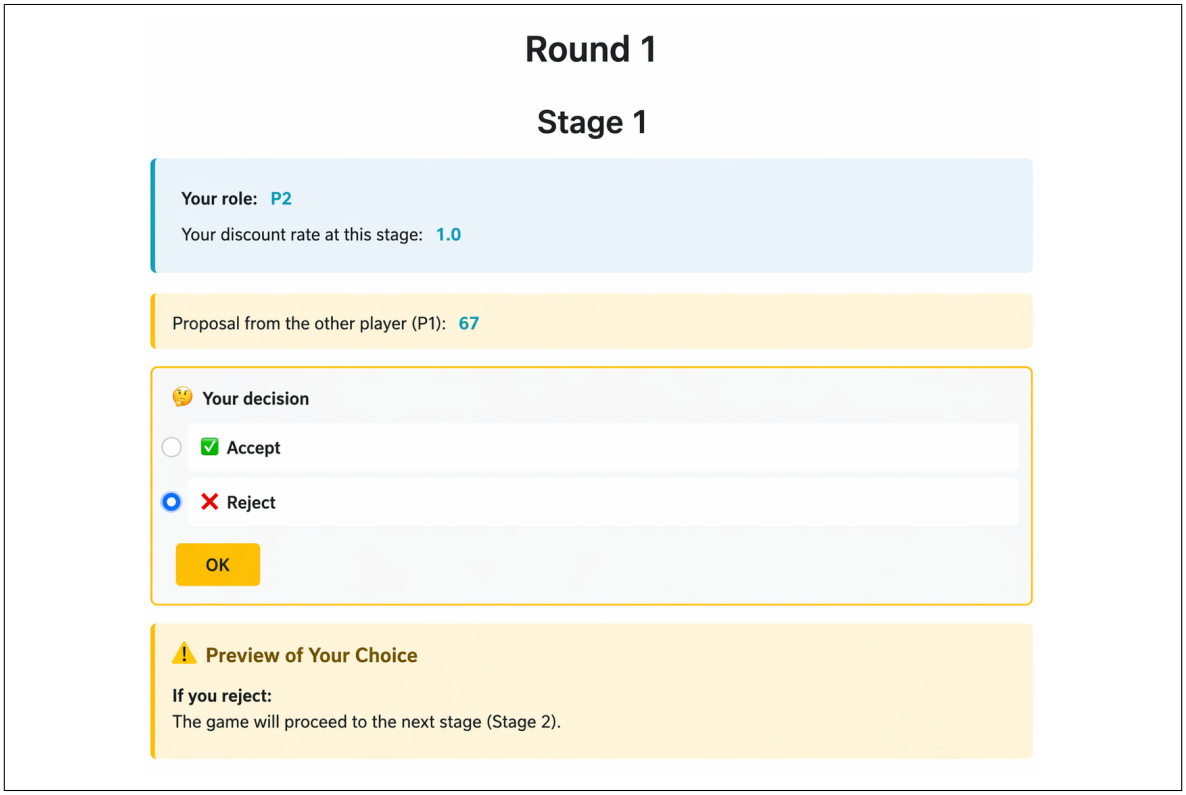


Figure I.2: P2 Stage 1 Proposal Screen

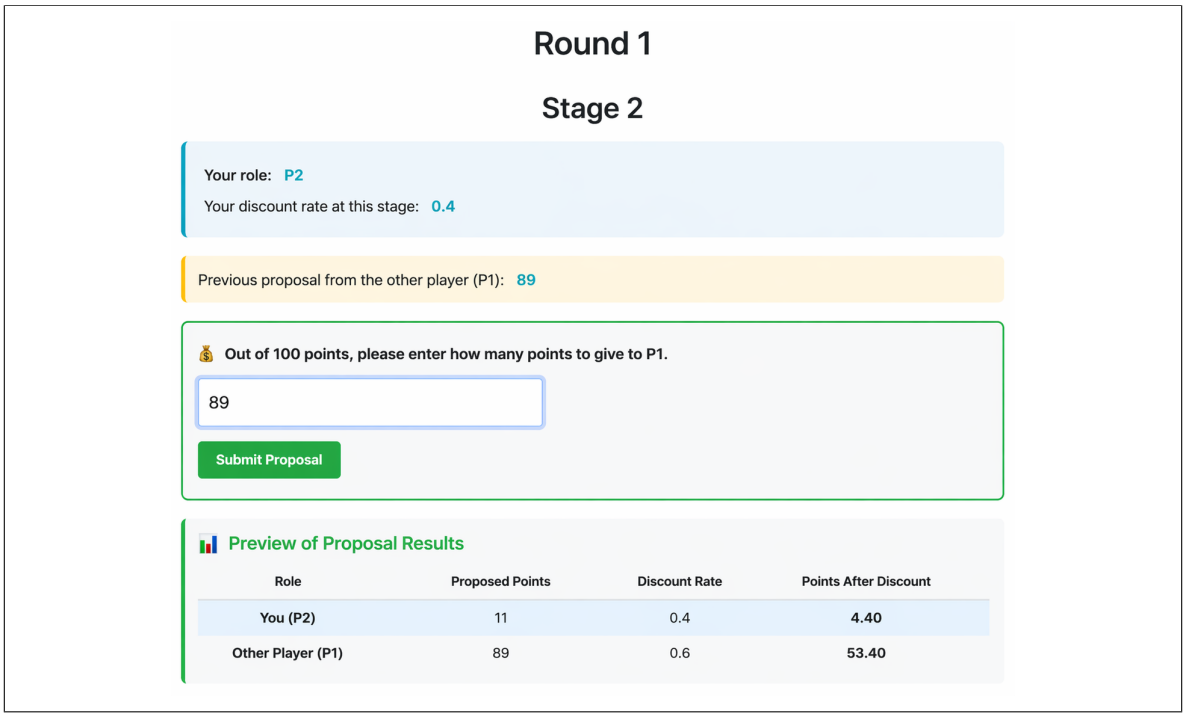


Figure I.3: P2 Stage 2 Proposal Screen

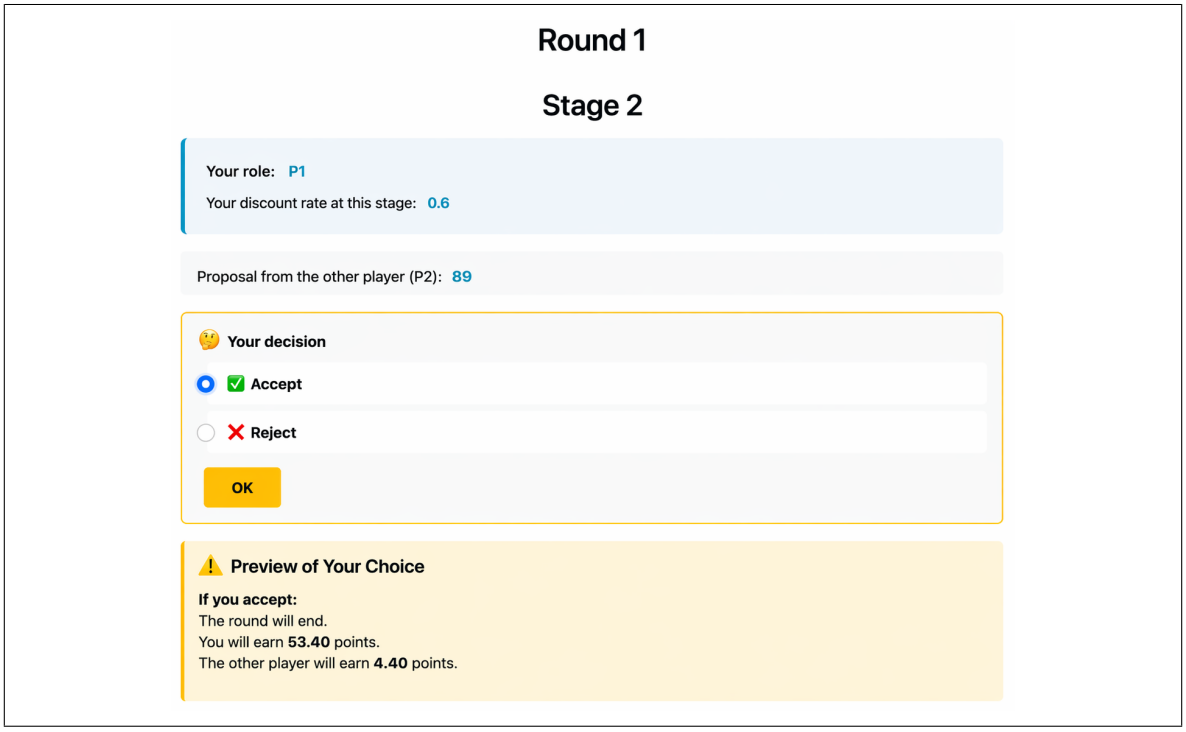


Figure I.4: P1 Stage 2 Proposal Screen

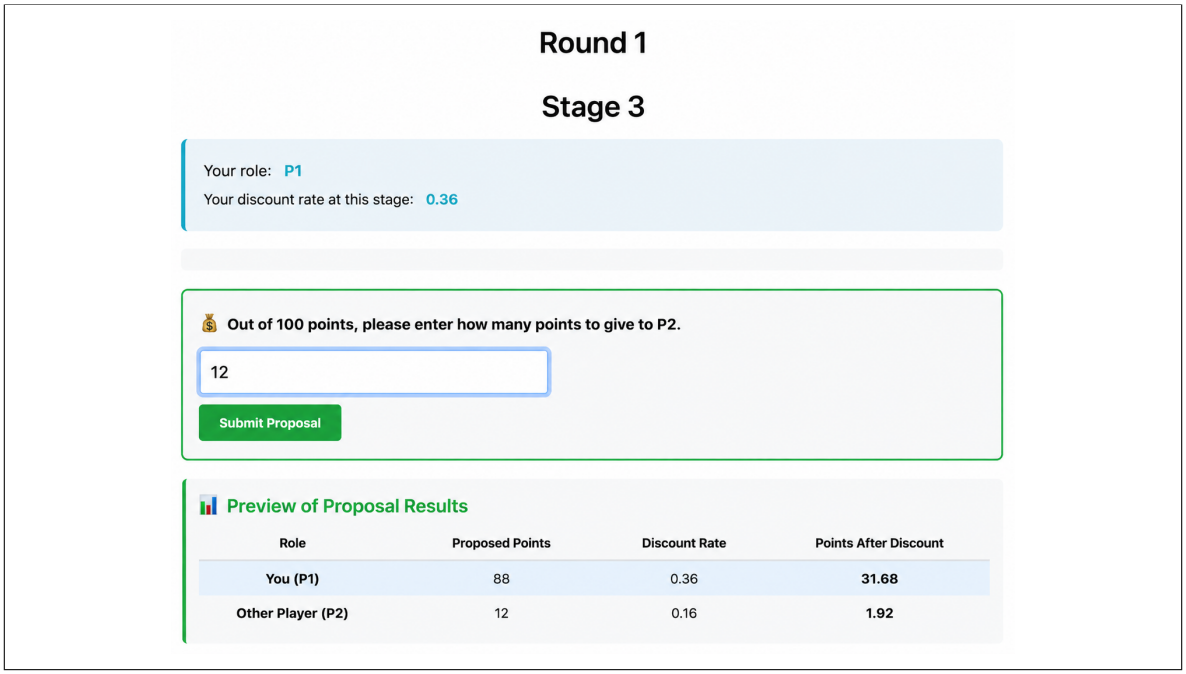


Figure I.5: P1 Stage 3 Proposal Screen

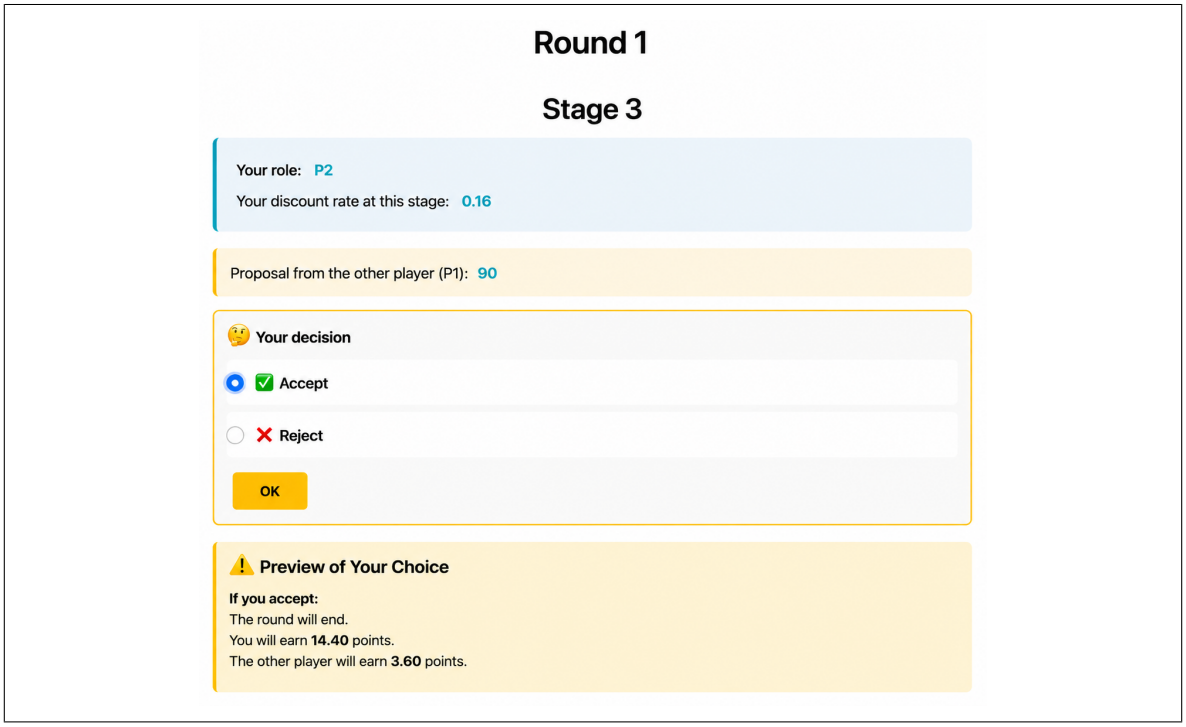


Figure I.6: P2 Stage 3 Proposal Screen

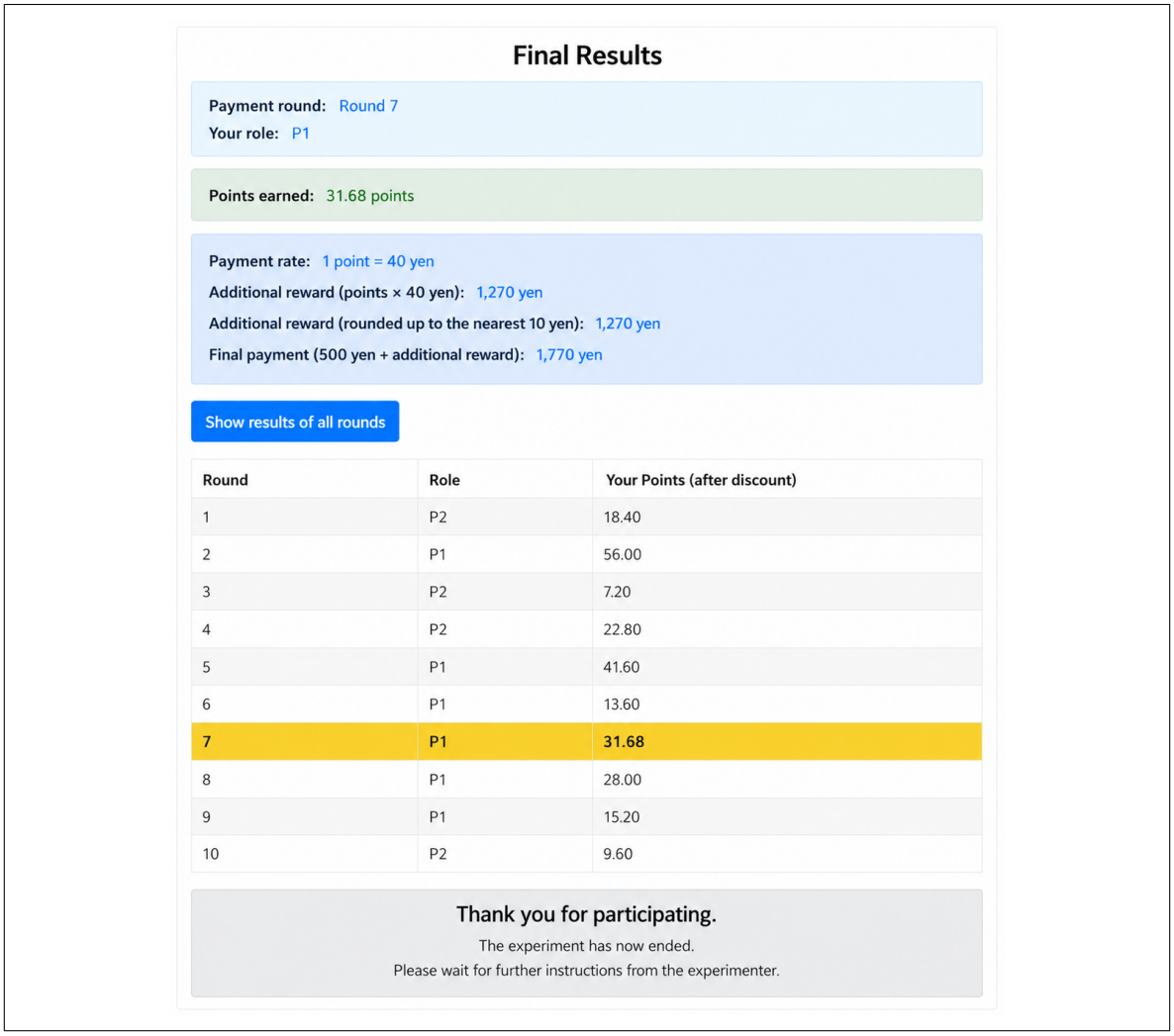


Figure I.7: T1 / T2 payment explanation screen

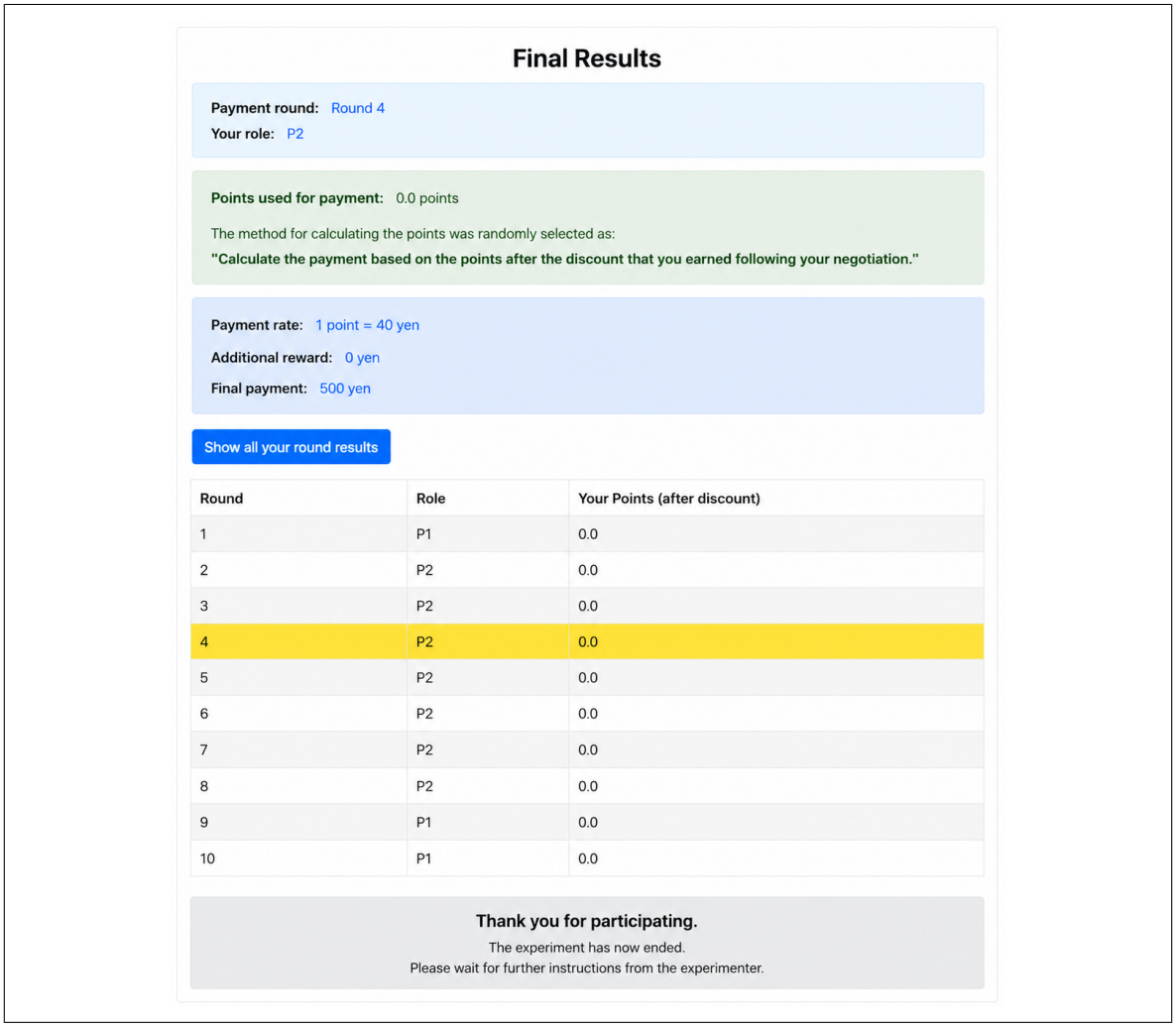


Figure I.8: T3 payment explanation screen

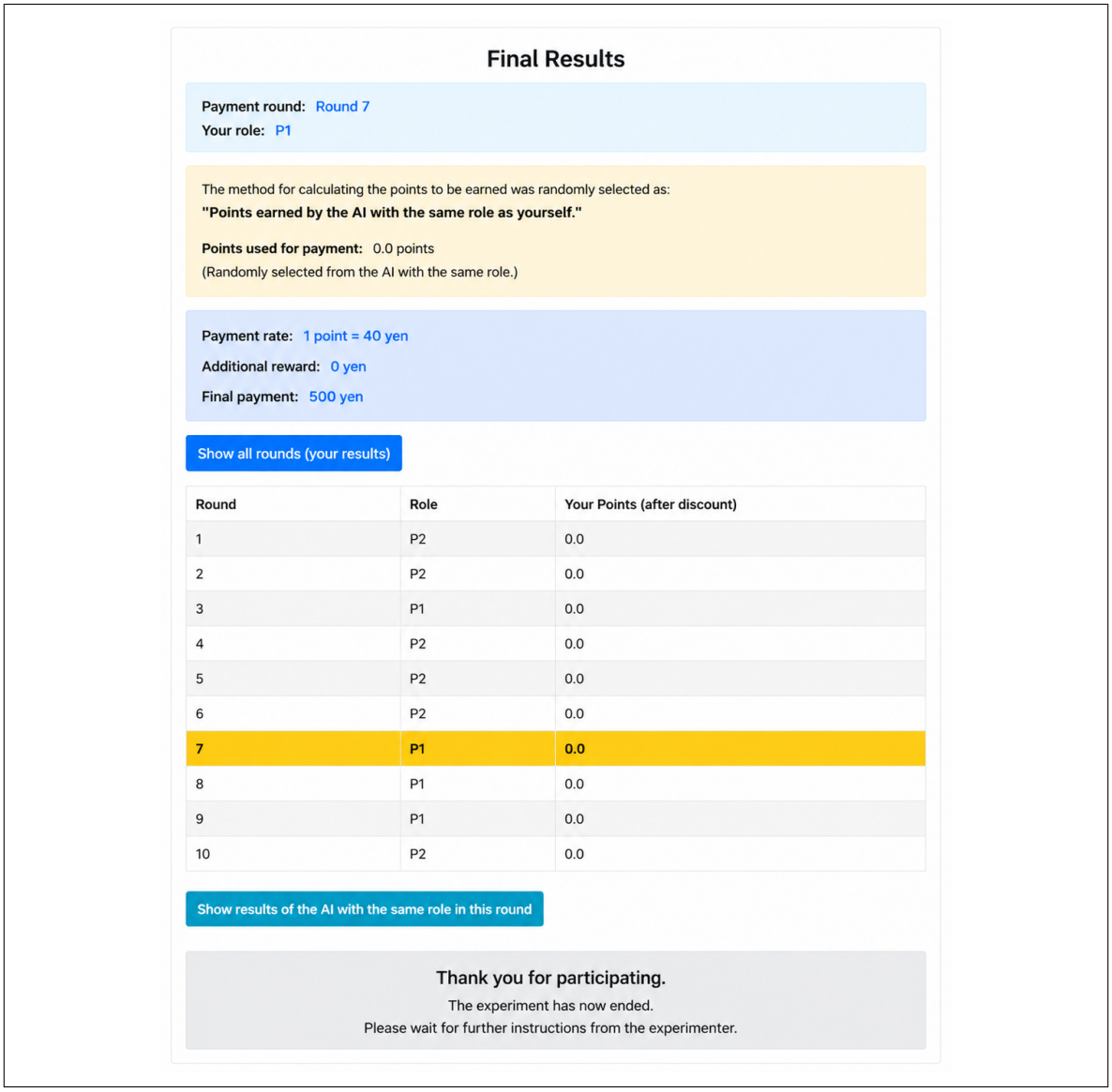


Figure I.9: T3 payment explanation screen (Human Beneficiary)