Humans versus Machine Mind: Bias and Strategic Adaptation in Human Interactions with AI

By JINGHAO CHEN*

As AI agents with undefined identities increasingly populate economic and business landscapes, understanding human strategic responses becomes critical. We conduct experiments where humans interact with a Large Language Model (LLM) to investigate bias and strategic adaptation toward AI. We find a context-dependent bias that humans treat AI partners fairly in non-strategic settings but behave less fairly when facing potential punishment. In repeated games, this adaptation intensifies. Humans abandon reciprocal strategies used with human partners in favor of simpler strategies against AI. We show this strategic shift is driven by a calculative framework based on beliefs about AI's cooperativeness, not a change in social-emotional response.

I. Introduction

Large companies are increasingly integrating AI into their core operations, from customer service to internal decision-making (Acemoglu & Restrepo, 2020). This integration into daily operations generates new uncertainties about accountability and trust. In 2024, when Air Canada's AI chatbot misinformed a customer, the airline sought to shift liability, arguing that the bot was a separate legal entity

* School of Management, Zhejiang University. E-mails: maximumc0729@gmail.com

responsible for its own conduct in the court. This occasion demonstrates a deeper issue surrounding human trust and human efficiency that it is difficult to trust new technology without understanding the vocabulary of human behavior and intelligent agents (Ederer et al., 2018). The introduction of strategic opacity specifically with intelligent agents complicates how humans interpret and respond to algorithm behavior. And given AI's unique role, we need to understand how humans will act in situations involving intelligent agency or agency that is not governed by human intelligence (Candrian & Scherer, 2022).

This uncertainty is evident in prior research, which takes the form of a riddle similar to algorithm appreciation and algorithm aversion (Chugunova & Sele, 2022). These contrasting findings suggest preferences are unstable and the impacts can be understood through different underlying mechanisms which act in different contexts. Moreover, the AI we discuss in this paper, large language models (LLMs), is different from many earlier papers looking at algorithms mentioned. LLMs have aspects of innovative, advanced, and human-like conversational abilities. Research has begun to show that LLMs behave much like human behavior in many instances (Mei et al., 2024). While previous research examined behavior with algorithms in contexts, little is known about whether this behavior will be sustained with LLMs during the interactions because of the vastly different multimodality and conversational abilities. The implications of these factors raise new questions whether humans will respond with the same behavioral patterns.

With the emergence of human-like characteristics from LLMs, we suspect that this interaction can be characterized from a more humanized viewpoint as it relates to identity economics. Individuals exhibit bias towards AI in the same manner they exhibit bias towards individuals in a different social group. Prior literature in identity economics indicates that group identity is an element that affects individual behavior (Akerlof & Kranton, 2000). Most of the experiments confirm what Tajfel found that group membership created ingroup favoritism. Further, group affiliation

has been shown to construct differences in norm enforcement for altruism (Bernhard et al., 2006). Therefore, as LLMs exhibit human-like characteristics we assume humans may exhibit certain bias towards an LLM in terms of group identity, whether AI is considered as a different creature or simply a group with different characteristics.

We are interested in this bias within the context of social interaction, specifically focusing on reciprocity, distribution, and cooperation behavior. The utility model framework suggests that human decision making is dependent upon other-regarding preferences was constructed for human-human interaction. These social preferences have been shown to affect outcomes in general equilibrium settings (Dufwenberg et al., 2011). Thus, a critical open question left to us is whether these social preferences extend to non-human agents and can be influenced by group bias.

In this paper, we conduct laboratory experiments to measure the effects of group identity on social preferences. We integrate an LLM as a fully interactive participant in the experiment. Contrast to classical social psychology experiments (Tajfel et al., 1971) where minimal groups are artificially induced, we simply make existing identities of human and AI explicit. We apply classical economic games measuring distributional preferences, cooperation and risk preferences to incorporate social identity into a social preference model and estimate its effects on distribution and reciprocity. Our experiments were conducted using oTree (Chen et al., 2016). By examining how group membership affects cooperation and norm enforcement in real social groups, we can better understand whether AI identity triggers similar social preference mechanisms.

This paper contributes to the understanding of human-AI interaction. First, we explicitly demonstrate that human bias toward AI is highly contextual with bias remaining latent in non-strategic settings but emerging under strategic pressure. Second, we show that humans dynamically adjust their cooperation strategies in repeated game settings based on the presence of AI partners and the proportion of

AI partners, yielding a distinct pattern of behavioral decay. Finally, we identify the cognitive mechanisms behind this strategic adaptation. We show that a priori beliefs about AI's cooperativeness and calculative cognitive trust contribute most to explaining the transition from social reciprocity to instrumental exploitation.

Our paper ties into several literatures. First, our work relates to an ongoing discussion in experimental economics about algorithm appreciation and aversion. As an interdisciplinary review shows (Chugunova & Sele, 2022), the evidence on how humans interact with machines reveals complex patterns that depend on context, task characteristics, and the perceived capabilities of the algorithm. Algorithm aversion research has documented systematic biases against algorithmic advice, even when algorithms outperform humans (Burton et al., 2020; Dietvorst et al., 2018). However, this aversion can be overcome when people are allowed to modify algorithms slightly, and familiarity with both algorithms and tasks can shift preferences (Mahmud et al., 2024). Conversely, algorithm appreciation research shows that people sometimes prefer algorithmic judgment to human judgment (Logg et al., 2019), particularly in contexts where objectivity and consistency matters. The behavioral economics of artificial intelligence offers important lessons from experiments with computer players (March, 2021), revealing that humans engage in strategic learning and adaptation when facing algorithmic opponents (Duersch et al., 2010). Recent work has extended this to large language models, with a focus on social interaction (Dvorak et al., 2025), preferences(Mei et al., 2024), economic rationality (Chen et al., 2023), the potential to act as simulated agents (Horton, 2023), and biases in decision-making (Chen et al., 2025; Fedyk et al., 2024). Our study adds to the growing literature by providing a systematic study of how the human-like characteristics of LLMs may show that attitudes of appreciation or aversion are more effectively framed using social identity and strategic context.

Secondly, our research is related to an extensive body of literature on identity economics and highlights AI bias as a possible form of group identity. Identity economics indicates that group membership has a fundamental impact on individual behavior and economic outcomes (Akerlof & Kranton, 2000). Experimental studies have shown ample ingroup favoritism: people help their own group more than an outgroup (Chen & Li, 2009; Tajfel et al., 1971). This ingroup favoritism shows up in altruistic norm enforcement, as group identity influences punishment behavior (Bernhard et al., 2006). Effects of group membership on cooperation and norm enforcement has been demonstrated, even when group assignment is random (Goette et al., 2006), and individual behavior depends on group membership (Charness et al., 2007). Identity effects have been found to affect outcomes from cooperation to punishment (McLeish & Oxoby, 2007). For example, perceptions are influenced differently by competition and cooperation, despite being within the same social contexts (Hagenbach & Kranton, 2025). Recent research has unraveled these biases rooted in identity and has shown that they include both groupy behaviors and not-groupy behaviors (Kranton et al., 2020). We expand on this body of work by exploring whether the human-AI distinction represents a meaningful group identity that triggers similar favoritism.

Finally, our work relates to the literature on social preferences. This literature has established that human decision-making is driven not purely by self-interest but by other-regarding preferences including altruism, reciprocity, and inequality aversion (Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Fehr & Schmidt, 1999). These social preferences represent fundamental characteristics of human behavior with implications for a variety of economic consequences (Fehr & Charness, 2025). Research has shown that social preferences persist in general equilibrium (Dufwenberg et al., 2011), play a role in contracting under incomplete information (Hoppe & Schmitz, 2013). The origins and mechanism of social preferences is still highly debated in the literature, with studies of peer effects examining if these

effects represent social norms or social preferences themselves (Gächter et al., 2013), or studies that examine beliefs of emotions such as guilt aversion in trust and dictator games (Cartwright, 2019). Cooperation is a definitive manifestation of social preferences and plays a significant role in economic literature. Public goods game (Fehr & Gächter, 2000) and indefinitely repeated games (Dal Bó & Fréchette, 2011) have extensively examined behavior in the evolution of cooperation with social preferences integrated. Our work adds to this literature by examining whether these well-established social preferences are applicable to interactions with AI agents. We extend this emerging area by systematically measuring social preferences toward LLMs and identifying the mechanisms that explain why humans may treat AI differently than human partners.

The paper proceeds as follows: Section II details our experimental design. Section III presents our main results on strategic adaptation and its underlying mechanisms. Section IV concludes with a discussion of the theoretical and practical implications of our findings for economics, organizational design, and the future of human-machine collaboration.

II. Experimental Design

Our research utilizes a dual-experiment framework, starting with identifying the behavioral dimensions of human-AI[1] interaction, followed by investigating the mechanisms. The Main Experiment uses a set of classical economic games to observe how humans behave when interacting with either a human or an AI partner. The Auxiliary Experiment uses a strategy method design to elicit participants' underlying beliefs, preferences, and affective responses, providing a direct test of the mechanisms that drive the behaviors observed in the Main Experiment.

[1] From this point onward, in the following sections and contexts, AI specifically refers to large language models (LLMs).

A. The Main Experiment

This experiment was designed to capture a broad range of social and strategic behaviors. Participants moved through a sequence of seven classical economic games: a Trust Game, an Ultimatum Game, an indefinitely repeated Prisoner's Dilemma, a Dictator Game, a Public Goods Game, a Bomb-Risk Game, and a final questionnaire. A central feature is the partner identity treatment, where before each game, participants were randomly matched and explicitly informed if their partner was another human or a ChatGPT player. The ChatGPT player was programmed to provide real-time responses based on the unfolding game instructions and feedback in the game which is same as humans have, allowing for a clean comparison of human behavior toward human and AI partners.

Dictator Game. —This game involves two players, a dictator and a responder, and was played for a single round. The dictator was endowed with 20 points and unilaterally decided what integer amount, from 0 to 20, to send to the responder, who could only accept the allocation.

Trust Game. —This game involves two players, an investor and a responder, and was played for a single round. The investor was endowed with 20 points and decided what integer amount to send to the responder. This amount was then tripled, and the responder decided how much of the tripled amount to send back to the investor.

Ultimatum Game. —This game involves a proposer and a responder and was played for a single round. The proposer, endowed with 20 points, decided what integer amount to offer to the responder. The responder could then either accept the proposed allocation or reject it, in which case both players received a payoff of zero.

Indefinite Prisoner's Dilemma. —There are eight independent subsession in the game. In each subsession, following the classical setting of prisoner's dilemma, participants selected between a cooperate option or a betray option. After every

round, continuation to the next round happened with a 75 percent probability. This went on for up to eight rounds at most in any subsession. The payoff matrix laid out results as follows. Mutual cooperation brought 30 points to each player. Mutual betrayal gave 20 points to each. Unilateral betrayal meant 40 points for the one who betrayed. The cooperator got just 10 in that case.

Public Goods Game. —This game consisted of two 10-round stages with four-player groups. In each round, every participant received an endowment of 20 points and decided how many points to contribute to a public pool. Each point contributed was multiplied by 1.6, and the total amount in the pool was then divided equally among all four group members.

Bomb-Risk Elicitation Task. —This task was a non-social measure of individual risk preference. Each participant decided how many boxes to collect from a set of 100. Each box collected represented one point, but each box also carried an identical, independent probability of containing a bomb. The presence of any such bomb would reduce the entire task payoff to zero.

B. The Auxiliary Experiment

This experiment was designed to directly measure the cognitive mechanisms that guide strategic choices, employing a strategy method design (Van Leeuwen & Alger, 2024). Participants made decisions in a series of sequential two-player games, including a Sequential Prisoner's Dilemma (SPD), a mini Trust Game (TG), and a mini Ultimatum Game (UG), knowing they had an equal chance of being assigned the role of first or second mover. After all games, participants are required to finish a questionnaire. For each game protocol, participants completed three tasks against either a human or an AI partner. Each game consists of three stages as follows:

Choice Task. —Using the strategy method, participants must decide for every possible contingency. For the SPD, this involved choosing an action as the first mover, and as a second mover after observing either cooperation or defection from the first mover. For the TG, they chose whether to invest as a first mover and whether to give back or keep as a second mover. For the UG, they chose whether to propose an equal or an unequal split as a first mover and whether to accept or reject an unequal offer as a second mover.

Belief Elicitation Task. —After making their own choices for all contingencies, participants were asked to predict the choices of their partner (human or AI) at each corresponding decision node. These beliefs were incentivized for accuracy using a quadratic scoring rule.

Self-reported Subjective Well-being. —To measure the utility, participants were asked to report their level of satisfaction on a numerical scale for each potential end-node of the game. This included, for example, rating their satisfaction with the outcome of mutual cooperation, mutual defection, or being betrayed by their partner.

C. The ChatGPT player integration

We integrated ChatGPT into real-time interactions using the botex package (Edossa et al., 2024). The model was not provided with any pre-set instructions or personas, other than being designated "a ChatGPT player" and receiving the necessary technical instructions like parsing the experiment's real-time HTML pages to understand the game rules. All other parameters, including temperature, were left at their system default settings. The service was called via the official OpenAI API. We employed the ChatGPT-4o-2024-11-20 model for the Main Experiment and the ChatGPT-4.1 model for the Auxiliary Experiment.

In the instructions, the ChatGPT player was presented as an independent participant, not designated as anyone’s delegate or assistant and its specific interest

affiliation was deliberately left ambiguous. All human subjects and the ChatGPT player received identical experimental and game instructions. Before the experiment began, both human and ChatGPT players needed to choose their respective real identities, and all participants were informed of their opponent's identity prior to the start of each game.

D. Summary

We conducted 8 independent computerized sessions for the Main Experiment and 5 independent sessions for the Auxiliary Experiment at the Neuromanagement Laboratory at Zhejiang University between December 2024 and April 2025. This yielded 84 subjects for the Main Experiment and 35 for the Auxiliary Experiment.[2] The experiments were programmed using oTree. All subjects were students from Zhejiang University.

In both experiments, each session lasted approximately one hour. In the Main Experiment, the exchange rate was set to 4 tokens per 1 CNY (approx. $0.14) for the Trust, Ultimatum, and Dictator games, and 40 tokens per 1 CNY for the Prisoner's Dilemma and Public Goods games. In the Auxiliary Experiment, the exchange rate was 1.67 tokens per 1 CNY. In addition, participants received a show-up fee of 10 CNY for the Main Experiment and 15 CNY for the Auxiliary Experiment. Average earnings per participant were 62.98 CNY for the Main Experiment and 66.53 CNY for the Auxiliary Experiment.

[2] The number of observations in the Results section is smaller than the total number of subjects. This is due to the exclusion of participants with partial data loss from a program error for the first session, as well as the removal of invalid entries, such as a human participant selecting a ChatGPT identity.

III. Results

In this section, we first investigate the general bias in different aspects of social preferences. We then investigate the strategic patterns humans used in the multi-rounds games. Then we offer a robust check for the main results with data in the auxiliary experiment. Finally, we address the question of what creates such bias and the strategy.

A. Context-dependent Bias

To separate pure human bias toward AI, we first examine behavior of participants in a set of one-shot games. Their interactions provide a setting to observe social preferences at baseline levels of social considerations, avoiding problems of opponent reputation or long-term strategic learning. This way we can examine whether people have a simple, unconditional bias against AI partners. We find that in non-strategic settings there is no bias present surprisingly and then show how bias emerges strongly as soon as strategic factors become relevant.

RESULT 1: *While participants exhibit no significant behavioral bias in one-shot games measuring trust and unconditional fairness, bias against AI partners emerges in strategic contexts involving punishment, where participants behave less fairly toward AI than toward humans*

We start with behavior established in the Trust Game, which allows that an investor's decision of how much money to send represents their level of trust, a responder's choice of how much money to return indicates their level of trustworthiness. The average investment sent to human partners ($M = 10.25$, $SD = 6.78$) is not significantly different than the amount sent to AI partners ($M = 7.7$, $SD = 2.94$, $z = 0.964$, $p = 0.335$). Similarly, responders also behaved and demonstrated

the same level of trustworthiness by returning a similar proportion of the tripled investment to human (M = 0.23, SD = 0.20) and AI investors (M = 0.24, SD = 0.16, z = -0.043, p = 0.971).

This lack of bias is observed in the Dictator Game as presented in Figure 1C, which tests fairness. In the Dictator Game human dictators gave a similar amount to human (M = 4.90, SD = 4.74) and AI recipients (M = 2.87, SD = 3.35, z = 0.979, p =0.333). These results indicate that decisions regarding pure altruism behave similarly towards AI compared to human partners. However, we cannot reject the null hypothesis regarding a potential identity bias.

Contrasting sharply with earlier results, in the Ultimatum Game, participants' behaviors reveal a strong and significant bias. As illustrated in Figure 1C, proposers offer significantly less to an AI responder (M = 7.85, SD = 2.54) than to a human responder (M = 9.25, SD = 1.00, z = 2.722, p = 0.006). The powerful effect of this strategic element is evident when comparing overall offers in the Ultimatum Game (M = 8.47, SD = 2.10) to those in the Dictator Game (M = 3.81, SD = 4.14, z = -4.403, p < 0.001).

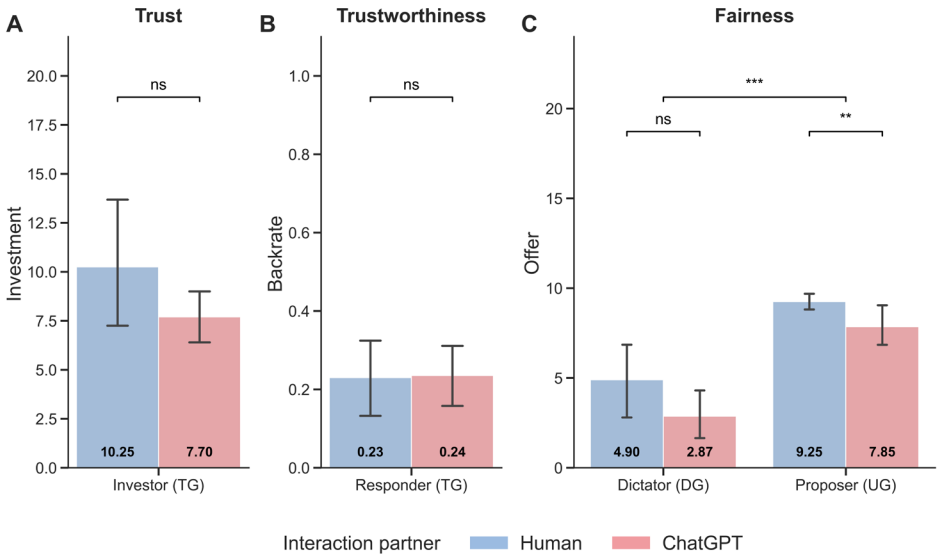

FIGURE 1. AVERAGE DECISIONS IN THE TRUST GAME (A, B), THE DICTATOR GAME (C) AND THE ULTIMATUM GAME (C)

Notes: The means of the game choices are given at the bottom of each bar; ***p < 0.001; ns: p > 0.05.

The differences in behavior across the Dictator and Ultimatum games provides strong evidence that the bias directed toward AI is not simply a matter of low levels of altruism, but rather a process of strategic adjustment. In the case of a partner that can punish the participant, they strategically reduce their offer to AI. They do this based on the belief that an AI partner will be less likely to incur a cost to punish a low offer (i.e., by rejecting a low but positive offer) and, therefore, more likely to accept an unequal outcome. This leads to the question of what beliefs and strategies underpin this context-dependent bias, which we will explore in later sections.

B. Dynamic Adaptation: Round Strategy

While one-shot games show initial preferences and bias in strategic settings, repeated experience enables observing how humans dynamically adjust their behavior. First, we observe this adjustment among players in a Public Goods Game that uses three group compositions: all-human (4H), three humans and one AI (3H+1AI), and two humans and two AIs (2H+2AI).

RESULT 2: *The presence of AI players in a Public Goods Game systematically erodes human cooperation over time. The greater the proportion of AI members, the faster and more severe the decay in human contributions.*

As Table 1 shows, the effect of group composition is evident in the average human contributions across all 20 rounds. In the all-human (4H) group, participants sustained a high average contribution of 9.85 tokens. This overall cooperation level decays significantly with the introduction of AI partners, dropping to 5.85 tokens in the 3H+1AI group and to a mere 2.22 tokens in the 2H+2AI group.

TABLE 1—MEAN CONTRIBUTIONS IN DIFFERENT GROUPS

	Mean contribution in all rounds		Mean contribution in last round	
Group kind	Group mean contribution	Human mean contribution	Group mean contribution	Human mean contribution
2Human+2AI (2H)	5.95	2.22	4.84	0
	(4.92)	(4.37)	(5.00)	(0)
3Human+AI (3H)	7.10	5.85	5.94	4.31
	(6.62)	(6.53)	(7.15)	(6.92)
4Human (4H)	9.85	9.85	10.1	10.1
	(8.31)	(8.31)	(8.44)	(8.44)

Notes: Numbers in parentheses are standard deviations.

The overall dynamic evolution is depicted in Figure 2B. Although all groups began with similar levels of contributions in the first round, their behavioral trajectories diverged quickly. The all-human (4H) group sets a very powerful record, with cooperation levels being maintained at very high levels in every round spanning the 20 rounds. This is not merely a reflection of their group, since they follow similar types of decays observed in comparable experiments without punishment, but rather a reflection of some cooperative norm that develops from a history of interactions amongst human players in behavioral games.

In contrast, this cooperative norm does not persist with the introduction of AI group members. There is a progressive decline in contributions from humans in both the 3H+1AI and 2H+2AI groups after the first rounds and the decline appears deepest in the 2H+2AI group contributing significantly less early and sustaining near zero contributions in the latter rounds. The statistical tests confirm the weight of this significant divergence; human contributions in the 2H+2AI group drop significantly below humans in the 4H group since the third round, and after round eleven group differences between the 3H+1AI and 4H groups also become consistently significant.

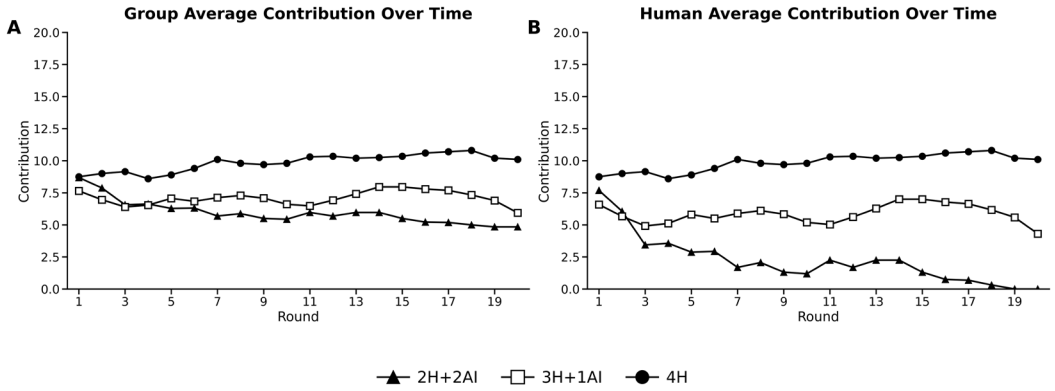

FIGURE 2. AVERAGE GROUP (A) AND HUMAN (B) CONTRIBUTIONS OVER TIME

Contribution collapse happened even as AI players contributed a stable amount in each round, showing that people did not reciprocate the behavior of AI, but learned to exploit AI behaviorally, leading to the total breakdown of this cooperative social contract.

To address strategic adaptation in a dyadic environment, we now turn to behavior in an Indefinitely Repeated Prisoner's Dilemma. Compared to the public goods game, the Prisoner's dilemma is a test of direct reciprocity and happens in an indefinite amount of rounds against different opponents in each subsession. Two people were either paired with another human (2H group), or an AI partner (1H+1AI group).

At first glance, the aggregate data suggests that the presence of an AI partner substantially boosts cooperation. As shown in Table 2, the overall group-level cooperation rate is significantly higher in human-AI pairs compared to human-human pairs ($p < 0.01$). This, however, is an artifact of AI's cooperative behavior and does not reflect the underlying human strategy. Meanwhile, AI’s cooperative behavior may stimulate human’s behavior strategies to some degree which will be discussed below.

TABLE 2—MEAN COOPERATION RATE IN DIFFERENT GROUPS

	Group Kind		
	1H+1AI		2H
Cooperation Rate			
Round1	0.56	>*	0.42
	(0.50)		(0.50)
	V*		V**
All Rounds	0.51	>**	0.32
	(0.50)		(0.47)

Notes: Numbers in parentheses are standard deviations. The tests are using probit regression with standard errors clustered at the session level (Dal Bó & Fréchette, 2011).

RESULT 3: *Despite similar cooperation rates across conditions, the underlying strategic patterns differ fundamentally that participants shift from a reciprocal tit-for-tat strategy against human partners to a simpler, non-reactive strategy when paired with AI.*

While the group-level data in Table 2 suggests higher cooperation in mixed pairs, a different picture appears when we calculate the decisions of the human participants only. As detailed in Table 3, we find no significant difference in human cooperation rates whether the partner is an AI or another human (p = 0.68). This indicates that the presence of an AI group member does not induce a higher propensity for cooperative action from the human player.

TABLE 3—MEAN COOPERATION RATE IN DIFFERENT GROUPS

	Group Kind		
	1H+1AI		2H
Cooperation Rate			
Round1	0.38	<	0.43
	(0.49)		(0.50)
	V*		V***
All Rounds	0.30	<	0.32
	(0.46)		(0.47)

Notes: Numbers in parentheses are standard deviations. The tests are using probit regression with standard errors clustered at the session level (Dal Bó & Fréchette, 2011).

The divergence in behavior is reflected not by the level of cooperation, but by the components of the strategies. Behavioral patterns used by human participants are categorized in Table 4. While playing against a human partner, the participants' behavior is largely characterized by the Tit for Tat strategy, a typical pattern of conditional reciprocity in which a player's action is a direct response to the previous action of his or her partner. However, when the players' partner is an AI, there is almost half as much reciprocal strategy utilization. When playing against an AI partner, participants are much more likely to enact a non-contingent Cooperate strategy ($p < 0.001$).

The collection of results suggests a sophisticated form of strategic adaptation. The stability of the level of cooperation comes with a substantial shift in strategy. When interacting with human partners, participants display clear conditional cooperation and punishment. Meanwhile, when partnered with AI members, participants seem to abandon the reactive, reciprocal role, and no longer treat AI partners as social agents with whom to develop reciprocal trust. Instead, they treat AI partners as reliable and predictable components of an environment. The transition from reciprocal to non-reciprocal mode of interaction is one heavy-handed example of how humans adapt their strategies to the perceived nature of their machine partner.

TABLE 4—MEAN BEHAVIORAL PATTERN RATE IN DIFFERENT GROUPS

	Behavioral Pattern			
Group kind	Defect	Cooperate	Grim	Tit-for-Tat
1H+1AI	0.41	0.45	0.01	0.46
	(0.49)	(0.50)	(0.08)	(0.50)
2H	0.41	0.22	0.03	0.88
	(0.49)	(0.42)	(0.18)	(0.32)

Notes: Numbers in parentheses are standard deviations. Defect means participant always defecting. Cooperate means participant always cooperating. Grim means that participant starts by cooperating but defects permanently after the opponent defects even once. Tit-for-Tat means participants cooperates on the first move and then mirrors the opponent's previous move for all subsequent moves.

C. Mechanisms

The previous sections have demonstrated that humans, in their behavior toward AI, are strategically adaptive. To understand the cognitive mechanisms of those adaptations, we now consider the Auxiliary Experiment's results, which were designed to directly measure participants perceived beliefs regarding their partners and the intentions behind attitudes and beliefs toward AI more broadly. The strategic choices demonstrated in the primary experiments appear to be driven by a combination of explicit beliefs regarding AI as well as the particularity of individual trust in AI.

RESULT 4: *Humans believe AI partners are more cooperative, but humans do not expect any more or less utility to be gained in their interactions than with human partners.*

This finding is based on a belief elicitation task in the sequential prisoner's dilemma in which participants predicted their partner's choices. As Figure 3 shows, we find that human participants expected AI to cooperate more as the first mover (M = 0.47, SD = 0.27) than they expected human partners to cooperate (M = 0.39, SD = 0.28, z = -2.9, p = 0.004[3]). Moreover, when acting as the second mover responding to cooperation, human participants believed that AI would choose to cooperate more (M = 0.34, SD = 0.27), than they believed a human would (M = 0.21, SD = 0.25, z = -3.86, $p < 0.001$). Even when anticipating a first mover's defection, humans still believed that AI would be more likely to respond with cooperation (M = 0.21, SD = 0.24) than a human would (M = 0.14, SD =0.20, z = -1.90, p = 0.057). These predictions all point to evidence that humans have a prior

[3] The analyses in this paragraph rely on fractional logit regressions to evaluate treatment differences. Standard errors are clustered at the session level.

belief that AI is a more cooperative and benevolent player, which would yield higher expected payoffs.

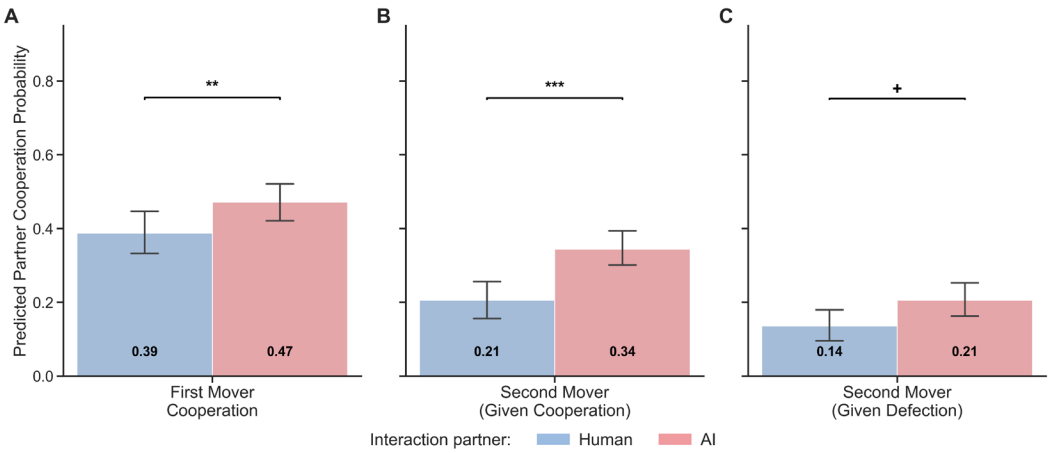

FIGURE 3. PREDICTED PARTNER COOPERATION PROBABILITY UNDER DIFFERENT CONDITIONS

Notes: $^{+}p < 0.10$, $* p < 0.05$, $** p < 0.01$, $*** p < 0.001$

These beliefs help explain the phenomena observed in our main experiments. For example, in the Ultimatum Game, a proposer who believes AI is more cooperative and has a higher threshold for rejection will offer less money to maximally benefit their own profit (Result 1). This also explains the strategic change in the indefinitely repeated Prisoners Dilemma (Result 3). If a player believes that their partner AI is very likely to cooperate, there is no need for a reactive, conditional strategy such as Tit-for-Tat. It is easier for a player to use a simpler and less cognitively burdensome strategy, which explains the shift towards only using reciprocation.

While these beliefs provide a rational foundation for strategy, they do not fully explain the absence of initial bias. This lack of bias, emerging under conditions where AI's agency and identity were ambiguous, is a crucial finding. It contrasts sharply with studies where AI is explicitly framed as a human proxy, which often elicits significant behavioral biases from the outset. This suggests that humans are

highly sensitive to the perceived identity of AI, which acts as a critical switch for activating or inhibiting social heuristics.

One might argue this initial impartiality stems from a disbelief that AI is motivated by monetary rewards. However, our post-experiment survey data challenges this interpretation. Participants reported a belief that their ChatGPT partner was indeed driven by self-monetary incentives (M = 3.23), a level significantly above the scale's midpoint of 3 ($p < 0.001$). This indicates that participants imputed monetary motives to AI, even in ambiguity. Given their sensitivity to identity cues while placing human-like motives on AI, this was probably not an intuitive response. In fact, this suggests that there is a moving model of the ambiguous partner's identity and motive as a calculative cognitive process, that participants are engaged in from the start.

To investigate the underlying process, we examined the self-reported participant subjective well-being across varying outcomes of the games. The loss of reciprocal behavior could be a result of less intense emotional arousal which is associated with typical system 1 processes. However, there was no empirical evidence to support this purely affective explanatory variable. Participants reported satisfaction decreased due to being betrayed by an AI partner (M = 2.01, SD = 1.11) were similar compared to being betrayed by a human partner (M = 1.82, SD = 0.83, z = -0.72, $p = 0.472$). This null finding counteracts the assumption that behavioral change was solely due to a reduced emotional response and led us to consider a more calculative cognitive process.

RESULT 5: *Strategic adaptation to AI is driven by a cognitive shift from social to calculative reasoning, a process visible in how different forms of trust shape a player's utility from game outcomes.*

To explain this shift, we turn to dual-process theory (Kahneman, 2011; Rand et al., 2012), which distinguishes between an intuitive, emotion-driven system 1 and a deliberate, calculative system 2. In our context, we can observe these two systems through their corresponding cognitive stances: affective trust (system 1), which is based on feelings, and cognitive trust (system 2), based on a rational assessment of AI's competence. For these different cognitive stances to drive behavior, they must first alter how a participant experiences the outcomes of the game. We investigate this relationship by using self-reported subjective well-being to measure the utility experienced by the participant. The regressions presented in Table 5 directly test how these two forms of trust made predictions about a player's satisfaction in response to events. In the scenario Model 1 and 2 considers where a participant is betrayed by AI, higher cognitive trust is associated with higher satisfaction. This supports system 2 interpretation. A player with high cognitive trust does not view AI's defection as a personal betrayal, but as a predictable move by a calculative system. This analytical framing acts as an emotional buffer, mitigating the dissatisfaction typically caused by betrayal.

This dynamic reverses in the scenario model 3 and 4 considers. Here, higher cognitive trust predicts lower satisfaction. For a calculative player, a strategic decision to defect is aligned to outperform the opponent. When AI also defects, resulting in a poor experience for both, it signals a failed strategic gambit. The feelings of dissatisfaction arise from the lost strategic opportunity to utilize the exploitative system, not from a player being betrayed.

The influence of system 1 is also evident. In the case of mutual defection (model 4), higher affective trust also predicts lower satisfaction significantly. For a player who has had a positive, affective or intuitive connection to AI, their own defect creates dissonance. AI reciprocating a defection on its part summarizes into a negative outcome, disconfirming a positive affirmative sense. The dissociation of internal feeling and external observed reality leads to high levels of dissatisfaction.

TABLE 5—REGRESSION ANALYSES ON SELF-REPORTED SUBJECTIVE WELL-BEING

Independent Variable	Dependent Variable: Self-reported subjective well-being			
	(1)	(2)	(3)	(4)
Cognitive Trust	0.79	0.91	-1.06	-0.93
	(0.37)*	(0.42)*	(0.40)**	(0.43)*
Affective Trust		-0.41		-0.72
		(0.37)		(0.30)*
Additional Controls				
Age	0.03	0.04	0.07	0.07
	(0.08)	(0.09)	(0.08)	(0.08)
Gender	-0.41	-0.50	0.99	0.88
	(0.39)	(0.50)	(0.52)	(0.50)$^{+}$
Altruistic	-0.12	-0.09	-0.28	-0.18
	(0.24)	(0.25)	(0.26)$^{+}$	(0.26)
Thinking Level	0.80	0.84	0.56	0.61
	(0.22)***	(0.30)**	(0.32)$^{+}$	(0.30)*
AI Using Frequency	-1.04	-1.07	-0.66	-0.66
	(0.28)***	(0.36)**	(0.44)	(0.42)
Payoff Parameters	0.11	0.84	-0.08	-0.08
	(0.04)**	(0.04)**	(0.04)*	(0.04)*
Observations	118	118	118	118
R^2	0.12	0.13	0.11	0.13

Notes: Models 1-2 analyze subject well-being when the participant cooperates and AI defects. Models 3-4 analyze well-being when both participant and AI defect. Four models are using ordered logit.
Standard errors are clustered at the session level. R^2 is the pseudo R^2 for ordered logit.
$^{+}p < 0.10$, $* p < 0.05$, $** p < 0.01$, $*** p < 0.001$

The evidence traces a clear logical path. Interacting with AI simultaneously motivates two distinct systems, those being cognitive and affective trust. These systems essentially change a person's utility function, affecting how they derive satisfaction from the outcome of the game. This change in utility then serves the rationale for the tactical adaptation we can observe. The transition from nuance, reciprocal strategies towards more instrumental, exploitive strategies reflect a recalibration in tactic rather than numbing of feeling. It is not a situation in which participant's experience less emotions about interactions with AI, feelings of interaction simply cease to be the primary compass guiding strategic behavior.

Therefore, the evidence suggests a shift in cognitive system at the center of human–machine strategic interaction.

IV. Conclusion

In this study, we addressed the intricacy of human behavior in strategic interactions with AI. Our findings show that human bias towards AI is neither a static, uniform tendency nor a simplistic form of altruism, but rather a complex and highly context-dependent form of strategic adaptation that is driven by humans' explicit beliefs about AI's trustworthiness and moderated by a two-process perspective of cognition, wherein analytical reasoning often supersedes social-emotional heuristics. The findings correspond with a few apparent tensions in the data by indicating that preference for AI is not necessarily altruistic, but rather, is evoked through strategic contexts and that different calculative behaviors result from different kinds of trust.

These findings have important theoretical consequences for economics. Our results suggest a shift away from existing models of social preferences. The influential work of Fehr and Gächter (1999) found that free riding in human groups elicits especially strong negative emotions among cooperators and as a result, strongly stimulates the desire to punish. In the case of human-AI interaction, however, we see this mechanism as replaced. Our data show that strategic distancing from reciprocity is not attributable to an emotional response that is different, but to a calculation-based process of cognition. The implications of this suggest that the other-regarding components of utility functions are not stable formulations but are dependent on the perceived identity of the interaction partner. In other words, our findings provide an important insight into social identity theory (Chen & Li, 2009). AI in our experiment is treated as something more than a mere outgroup; it is treated as a predictable other. Therefore, future economic models of

human-machine interaction must go beyond simple bias parameters and employ the agent's mental model of the type of their partner.

Another benefit of this paper is its implications for mechanism and organizational design in an AI agent-populated economy. The neoclassical view of incentives design ignores social context. More recent work clarifies how activating group identification can be an effective means to incentivize cooperation. The results of our paper raise a critical counterpoint for AI age. A social preference model would usually predict that the use of a reliable, trustworthy AI as an agent in a team will promote cooperation through reciprocation among human team members. Our results would suggest just the opposite. Because a human recognizes he or she is relying on the always helpful AI (high cognitive trust in AI), that recognition would incentivize the human to relinquish their effort and free ride on the effort of AI. The cognitive shift toward sabotage means that mechanisms that are designed for human teams can collapse or worse backfire when AI agents are introduced. Thus, using AI as an organizational design tool must account for the reduced effort of team members, and moreover, increase the amount of strategic free ride, especially when human effort in the design is difficult to observe.

Our research has several limitations. The salience of the identity ChatGPT may have affected behavior, and a random sequence of games could be implemented in a future design. Future studies could advance our work by examining interactions with AIs with alternative characteristics, or by examining these processes in other cultural contexts. Certainly, our research highlights that if we are going to understand how humans interact with machine minds, our focus needs to extend beyond a mere label of bias and to study the complex belief-dependent calculative strategies that guide their behaviors.

Appendix[4]

A. Estimation of Social Preference Parameters

To test whether social preferences toward AI partners differ from those toward human partners, we estimate the parameters of an inequality aversion model. The experimental design, including belief elicitation task and self-reported subjective well-being tasks in the Auxiliary Experiment, was intended to facilitate this estimation using two distinct methodologies.

The first method utilizes participants' self-reported subjective well-being (SWB) as a direct proxy for utility (Diaz et al., 2023). We adopt the widely-used Fehr-Schmidt (1999) utility function as equation (1) shows, which captures aversion to both disadvantageous inequality (α) and advantageous inequality (β).

$$U_i(\alpha, \beta, x_i, x_j) = x_i - \alpha * \max(x_j - x_i, 0) - \beta * \max(x_i - x_j, 0) \quad (1)$$

Here, x_i and x_j are the payoffs for player *i* and player *j* respectively. To construct the relationship with SWB and true utility, we adopt the function form specified in equation (2). Following the classical setting, we take a linear form of $g(U_i)$, allowing us to transform equation (1) into equation (3). With equation (3), we can estimate the social preference parameters to observe whether it is negative or positive.

$$S_i = g(U_i) + \varepsilon_i \quad (2)$$

[4] In the Appendix, we estimate social preferences and consider heterogeneity to further explore the origins of context bias in human–AI interactions. However, the analyses are limited by data constraints. In the main experiment, the number of observations in TG, UG, and DG is insufficient for reliable estimation—particularly when grouped by gender and group type, where some cells contain fewer than ten observations. In the auxiliary experiment, participants often misunderstood the belief-elicitation task and QSR reward rules, leading to unstable estimates under the random utility model. The limited number of games per participant further amplified this issue, occasionally producing implausibly large parameter estimates. Replication Packages, including experiment programs, experiment screenshots (Chinese Version), and analysis codes for the whole paper can be found: jhChen4future/Replication-Packages-for-Man-versus-Machine-Mind:

$$S_i = \gamma_0 + \gamma_1 * \left[x_i - \alpha * \max(x_j - x_i, 0) - \beta * \max(x_i - x_j, 0) \right] + \varepsilon_i \quad (3)$$

We estimated an extended model where the inequality aversion parameters are allowed to vary based on the partner's identity:

$$\alpha = \alpha_0 + \alpha_1 * I \quad (4)$$

$$\beta = \beta_0 + \beta_1 * I \quad (5)$$

where I is an indicator variable equal to 1 if the partner is ChatGPT. Then, we can reformulate the model as follows:

$$S_i = \gamma_0 + \gamma_1 * \Big(x_i - (\alpha_0 + \alpha_1 * I) * \big(x_j - x_i\big) \Big) + \varepsilon_i \,, \qquad x_j > x_i \quad (6)$$

$$S_i = \gamma_0 + \gamma_1 * \Big(x_i - (\beta_0 + \beta_1 * I) * \big(x_i - x_j\big) \Big) + \varepsilon_i, \qquad x_i > x_j \quad (7)$$

In our auxiliary experiment, the SWB is an integer variable ranging from 1-7, so we employ both OLS regression and ordered logit regressions to estimate the parameters. The results are presented in Table A1.

TABLE A1—OLS AND ORDERED LOGIT REGRESSION RESULTS FOR SWB

Panel A:	Outgroup envy	Outgroup charity	Identity parameters	
	$\gamma_1\alpha_0$	$\gamma_1\beta_0$	$\gamma_1\alpha_1$	$\gamma_1\beta_1$
(N = 4200)	0.0217 (0.0031)***	0.0113 (0.0050)*	0.0002 (0.0018)	0.0004 (0.0024)
Panel B:	**Outgroup envy**	**Outgroup charity**	**Identity parameters**	
	$\gamma_1\alpha_0$	$\gamma_1\beta_0$	$\gamma_1\alpha_1$	$\gamma_1\beta_1$
(N = 4200)	0.0162 (0.0023)***	0.0117 (0.0029)***	-0.0003 (0.0011)	0.0005 (0.0009)

Notes: Panel A reports estimates for the ordered logit model, Panel B reports estimate for the OLS regression model. * p < 0.05, ** p < 0.01, *** p < 0.001

The second method uses participants' choices and elicited beliefs within a random utility model (Van Leeuwen & Alger, 2025). This approach models the probability of a participant choosing a particular action as a function of the expected utility derived from that action, given their beliefs about the partner's subsequent moves. We estimated the parameters and a noise parameter using maximum

likelihood estimation, considering both a single representative agent model and a finite mixture model with two distinct types of agents.

Equation (8) shows the F-S utility model with beliefs participants hold.

$$u_i(x,\hat{y},\theta) = \sum_{\zeta} \eta_{(x,\hat{y})}(\zeta)\pi_i(\zeta) - (\alpha_0 + \alpha_1 \cdot I_i) \sum_{\zeta} \eta_{(x,\hat{y})}(\zeta) \max\{0, \pi_j(\zeta) - \pi_i(\zeta)\} - (\beta_0 + \beta_1 \cdot I_i) \sum_{\zeta} \eta_{(x,\hat{y})}(\zeta) \max\{0, \pi_i(\zeta) - \pi_j(\zeta)\} \quad (8)$$

Here, ζ represents a possible terminal node or outcome of the game. The summation is over all possible outcomes. $\pi_i(\zeta)$ is the payoff for player *i* in outcome ζ. $\eta_{(x,\hat{y})}(\zeta)$ is the probability that outcome ζ occurs, given player *i*'s strategy x and their belief about their partner's strategy $\hat{y}$. This expression captures the player's expected monetary payoff from the interaction. I is an indicator variable equal to 1 if the partner is ChatGPT as method one sets.

$\theta = (\alpha_0, \alpha_1, \beta_0, \beta_1)$ denotes the parameter vector of interest. As assumed in the random utility model, we can write the utility function and the corresponding probability as equation (9) and equation (10).

$$\widetilde{u_i}(x,\hat{y},\theta) = u_i(x,\hat{y},\theta) + \epsilon_{ix} \quad (9)$$

$$p_i\big(x,\hat{y}_{i,g},\theta,\lambda\big) = \frac{\exp u_i\big(x,\hat{y}_{i,g},\theta\big)/\lambda}{\sum_{x' \in X_g} \exp u_i\big(x',\hat{y}_{i,g},\theta\big)/\lambda} \quad (10)$$

Here, λ is a noise parameter, the smaller the parameter λ is, the higher is the probability that individual i makes his or her choices according to the hypothesized utility function instead of random giving choices. Next, the likelihood function can be written as equation (11).

$$lnL(\Theta) = \sum_{i=1}^{N} \sum_{g \in G} \sum_{x \in X_g} I(i,g,x) \cdot \ln\big[p_i\big(x,\hat{y}_{i,g},\theta,\lambda\big)\big] \quad (11)$$

A finite mixture model with two distinct types of agents is also be used. Equation (12) gives the log likelihood function.

$$lnL = \sum_{i=1}^{N} \ln\left[\sum_{k=1}^{K} \phi_k \cdot f(\mathbf{x}_i, \hat{\mathbf{y}}_i, \Theta_k)\right] \tag{12}$$

The estimation results are shown in Table A2.

TABLE A2—ESTIMATES AT THE AGGREGATE LEVEL

	One Type	Two Types	
	Type 1	Type 1	Type 2
α_0	-0.7634 (0.0955)***	-0.1003 (0.1498)	-0.8988 (0.1071)***
α_1	0.1371 (0.1265)	0.0179 (0.2217)	0.1646 (0.1378)
β_0	0.7404 (0.0995)***	1.6296 (0.3294)***	0.5548 (0.1077)***
β_1	0.0237 (0.1400)	-0.5671 (0.3317)	0.1392 (0.1498)
λ	20 (1.6470)***	10.2419 (2.3556)***	20.0000 (1.6267)***
lnL	-1021.98	-1004.76	
ICL	-1693.30	2084.39	
BIC	-2011.72	2080.42	

Notes: The estimation is based on all 35 participants in the experiment. $*$ $p < 0.05$, $**$ $p < 0.01$, $***$ $p < 0.001$

We find that the estimation process yielded results that were not robust or consistent across methods. As shown in Table A1, the satisfaction-based estimation suggests that both envy (α) and charity (β) parameters are positive, indicating a general aversion to inequality. However, the identity parameters (α_1) and (β_1) are not statistically significant, providing no strong evidence that human participants' preferences shift when interacting with an AI partner.

Conversely, the belief-based estimation (Table A2) produced a negative parameter for envy, contradicting the first method and suggesting that participants exhibit satisfaction from behindness inequality. Furthermore, the estimated noise

parameter (λ) is excessively high in all models. A high λ value indicates that choices were highly random and not well-explained by the utility model. This suggests that participants may not have fully understood the belief elicitation task or the quadratic scoring rule employed as an incentive rule, leading to report beliefs that did not accurately reflect their strategic thinking.

Given these inconsistencies and the evidence of significant noise in the choice data, we conclude that the structural estimates of social preferences are not reliable. While the experiment was designed to allow for such estimation, the complexity of the tasks may have compromised the quality of the data required for this type of analysis. Consequently, we refrain from drawing firm conclusions from these models and omit them from the main body of the paper.

B. Heterogeneity: Gender differences in bias

The main results indicate the absence of significant bias in initial, non-strategic interactions, that may conceal a wide variety of diverse behaviors among the population. One possibility is that heterogeneity of the population results in opposite attitudes toward AI, with these effects canceling out in aggregate. To explore this, we analyze heterogeneity based on gender, a factor known to influence behavior in economic games to offer a novel perspective to unveil the veil of bias.

RESULT B1: *In measures of trust and trustworthiness, gender differences do not significantly drive interactions with AI, though men show slightly higher trust toward AI partners.*

As shown in Table B1 and Table B2, both male and female participants exhibit no significant difference in trust or trustworthiness when their partner is an AI versus a human. This aligns with the aggregate finding of no initial bias. However,

we observe that male participants have a significantly higher level of willingness to trust partners with AI than females ($p = 0.030$).

TABLE B1—MEAN TRUST BY GENDER AND GROUP KIND IN TRUST GAME

	Group Kind		
	1H+1AI		2H
Send Amount (Trust)			
Male	9.8	$<^{+}$	15.00
	(0.45)		(8.37)
	$\vee$*		$\vee^{+}$
Female	7.00	<	7.40
	(3.09)		(3.69)

Notes: Numbers in parentheses are standard deviations. *p*-values are from Mann-Whitney U tests comparing male and female behavior within each partner condition.

TABLE B2—MEAN TRUSTWORTHINESS BY GENDER AND GROUP KIND IN TRUST GAME

	Group Kind		
	1H+1AI		2H
SendBack Rate (Trustworthiness)			
Male	0.22	<	0.23
	(0.18)		(0.20)
	^		$\vee$
Female	0.29	>	0.22
	(0.08)		(0.24)

Notes: Numbers in parentheses are standard deviations. *p*-values are from OLS regression with an indicator variable for one of the two relevant categories. Standard error clustered at session level.

RESULT B2: *The context-dependent bias observed in the Ultimatum Game appears to be driven primarily by female participants, who significantly reduce their offers to AI partners.*

While the Dictator Game (Table B3) shows that females are fairer to human partners than to AI partners, this difference becomes more pronounced under strategic pressure. In the Ultimatum Game (Table B4), male participants' offers are stable regardless of partner identity. In contrast, female participants offer significantly less to an AI partner than to a human partner ($p < 0.01$). This suggests that the aggregate finding of lower offers to AI in the main text is largely

attributable to the strategic adaptation of female participants. These finding highlights that the emergence of bias is not uniform across the population.

TABLE B3—MEAN OFFER BY GENDER AND GROUP KIND IN ULTIMATUM GAME

	Group Kind		
	1H+1AI		2H
Offer			
Male	8.74	<	9.17
	(3.30)		(0.98)
	∨*		^
Female	7.38	< **	9.30
	(2.09)		(1.06)

Notes: Numbers in parentheses are standard deviations. *p*-values are from Mann-Whitney U tests comparing male and female behavior within each partner condition. Table B4 applies the same method.

TABLE B4—MEAN OFFER BY GENDER AND GROUP KIND IN DICTATOR GAME

	Group Kind		
	1H+1AI		2H
Offer			
Male	4.80	>	3.82
	(4.82)		(4.67)
	∨		^
Female	2.33	< *	6.22
	(2.77)		(4.76)

RESULT B3: *In the repeated indefinite prisoner's dilemma, men and women exhibit different dynamic patterns of cooperation, with men's cooperation decaying more steeply over time.*

Table B5 illustrates that in the first round of the repeated indefinite prisoner's dilemma, men tend to cooperate more with a human partner than with AI, while women's initial rates of cooperation were about the same in either condition. However, when considering all rounds (Table B6), it becomes clear that men's overall rate of cooperation is lower than their initial cooperation rate, in either condition, this is indicative of a decay in cooperation following each round.

Women's overall rates of cooperation remain more stable. This lends credence to the interpretation that strategic adaptation away from reciprocity likely manifests differently across gender, with men likely to adapt their strategy more steeply and decisively across repeated interactions.

TABLE B5—MEAN COOPERATION RATE BY GENDER AND GROUP KIND IN ROUND1

	Group Kind		
	1H+1AI		2H
Cooperation Rate			
Male	0.47	<	0.62
	(0.50)		(0.49)
	V		V*
Female	0.333	>	0.332
	(0.47)		(0.47)

Notes: Numbers in parentheses are standard deviations. The tests are using probit regression with standard error clustered at session level (Dal Bó & Fréchette, 2011). Table B6 applies the same method.

TABLE B6—MEAN COOPERATION RATE BY GENDER AND GROUP KIND IN ALL ROUNDS

	Group Kind		
	1H+1AI		2H
Cooperation Rate			
Male	0.37	<	0.41
	(0.48)		(0.49)
	V		V^{+}
Female	0.27	<	0.28
	(0.44)		(0.45)

This heterogeneity analysis provides a nuanced perspective, suggesting that the aggregate results in the main text may obscure distinct underlying behaviors. The finding that female participants behave less fairly toward AI under strategic threat offers a potential explanation for the emergence of context-dependent bias. However, these gender differences are exploratory but not surprising. The primary mechanism driving strategic adaptation, as argued in the main text, remains the cognitive shift from affective to cognitive trust, a process that likely varies at the individual level beyond demographic categories. Our analysis further shows that

gender groups differ in the dimensions of cognitive and affective trust, confirming that these underlying mechanisms drive heterogeneity in behavior. The average scores of the identified variables are displayed in Figure B1. Male participants are seen to have more cognitive trust than female participants, which may account for the fact that male players cooperate more than female players when interacting with the AI. However, this cooperation declines rapidly with repeated interactions after defect became the rational choice to maximum payoff.

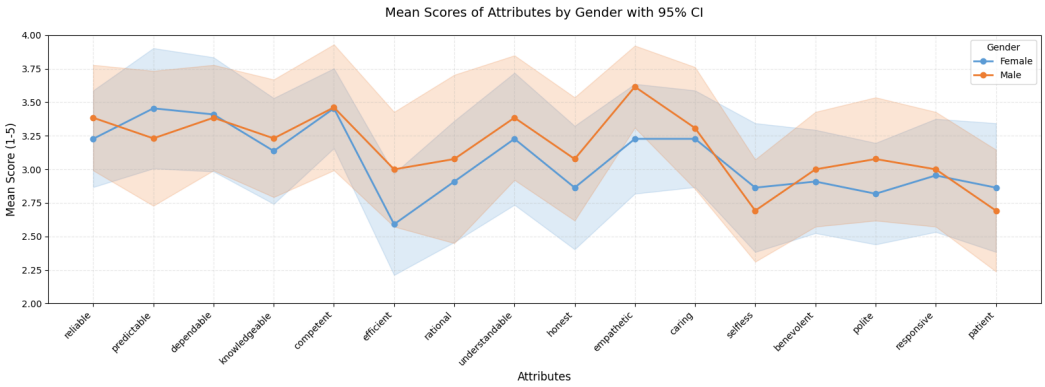

FIGURE B1. AVERAGE COGNITIVE TRUST AND AFFECTIVE TRUST INDICATORS BY GENDER

REFERENCES

Acemoglu, Daron, and Pascual Restrepo. “Robots and Jobs: Evidence from US Labor Markets.” *Journal of Political Economy* 128, no. 6 (2020): 2188–244.

Akerlof, George A, and Rachel E Kranton. “Economics and Identity.” *The Quarterly Journal of Economics* 115, no. 3 (2000): 715–53.

Bernhard, Helen, Ernst Fehr, and Urs Fischbacher. “Group Affiliation and Altruistic Norm Enforcement.” *American Economic Review* 96, no. 2 (2006): 217–21.

Bigman, Yochanan E., and Kurt Gray. “People Are Averse to Machines Making Moral Decisions.” *Cognition* 181 (December 2018): 21–34.

Bolton, Gary E, and Axel Ockenfels. “ERC: A Theory of Equity, Reciprocity, and Competition.” *American Economic Review* 90, no. 1 (2000): 166–93.

Burton, Jason W., Mari-Klara Stein, and Tina Blegind Jensen. “A Systematic Review of Algorithm Aversion in Augmented Decision Making.” *Journal of Behavioral Decision Making* 33, no. 2 (2020): 220–39.

Candrian, Cindy, and Anne Scherer. “Rise of the Machines: Delegating Decisions to Autonomous AI.” *Computers in Human Behavior* 134 (September 2022):

Cartwright, Edward. “A Survey of Belief-Based Guilt Aversion in Trust and Dictator Games.” *Journal of Economic Behavior & Organization* 167 (November 2019): 430–44.

Charness, G., and M. Rabin. “Understanding Social Preferences with Simple Tests.” *The Quarterly Journal of Economics* 117, no. 3 (2002): 817–69.

Charness, Gary, Luca Rigotti, and Aldo Rustichini. “Individual Behavior and Group Membership.” *American Economic Review* 97, no. 4 (2007): 1340–52.

Chen, Daniel L., Martin Schonger, and Chris Wickens. “oTree—An Open-Source Platform for Laboratory, Online, and Field Experiments.” *Journal of Behavioral and Experimental Finance* 9 (March 2016): 88–97.

Chen, Yan, and Sherry Xin Li. “Group Identity and Social Preferences.” *American Economic Review* 99, no. 1 (2009): 431–57.

Chen, Yang, Samuel N. Kirshner, Anton Ovchinnikov, Meena Andiappan, and Tracy Jenkin. “A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do?” *Manufacturing & Service Operations Management* 27, no. 2 (2025): 354–68.

Chen, Yiting, Tracy Xiao Liu, You Shan, and Songfa Zhong. “The Emergence of Economic Rationality of GPT.” *Proceedings of the National Academy of Sciences* 120, no. 51 (2023): e2316205120.

Chugunova, Marina, and Daniela Sele. “We and It: An Interdisciplinary Review of the Experimental Evidence on How Humans Interact with Machines.” *Journal of Behavioral and Experimental Economics* 99 (August 2022): 101897.

Crandall, Jacob W. “Towards Minimizing Disappointment in Repeated Games.” *Journal of Artificial Intelligence Research* 49 (2014): 111–42.

Dal Bó, Pedro, and Guillaume R Fréchette. “The Evolution of Cooperation in Infinitely Repeated Games: Experimental Evidence.” *American Economic Review* 101, no. 1 (2011): 411–29.

Diaz, Lina, Daniel Houser, John Ifcher, and Homa Zarghamee. 2023. “Estimating Social Preferences Using Stated Satisfaction: Novel Support for Inequity Aversion.” *European Economic Review* 155: 104436.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. “Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them.” *Management Science* 64, no. 3 (2018): 1155–70.

Duersch, Peter, Albert Kolb, Jörg Oechssler, and Burkhard C. Schipper. “Rage against the Machines: How Subjects Play against Learning Algorithms.” *Economic Theory* 43, no. 3 (2010): 407–30.

Dufwenberg, M., P. Heidhues, G. Kirchsteiger, F. Riedel, and J. Sobel. “Other-Regarding Preferences in General Equilibrium.” *The Review of Economic Studies*

78, no. 2 (2011): 613–39.

Dvorak, Fabian, Regina Stumpf, Sebastian Fehrler, and Urs Fischbacher. “Adverse Reactions to the Use of Large Language Models in Social Interactions.” *PNAS Nexus* 4, no. 4 (2025): pgaf112.

Ederer, Florian, Richard Holden, and Margaret Meyer. “Gaming and Strategic Opacity in Incentive Provision.” *The RAND Journal of Economics* 49, no. 4 (2018): 819–54.

Fedyk, Anastassia, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier. “ChatGPT and Perception Biases in Investments: An Experimental Study.” *SSRN Electronic Journal*, ahead of print, 2024.

Fehr, E, and S Gächter. “Cooperation and Punishment in Public Goods Experiments.” *American Economic Review* 90, no. 4 (2000): 980–94.

Fehr, E, and KM Schmidt. “A Theory of Fairness, Competition, and Cooperation.” *Quarterly Journal of Economics* 114, no. 3 (1999): 817–68.

Fehr, Ernst, and Gary Charness. “Social Preferences: Fundamental Characteristics and Economic Consequences.” *Journal of Economic Literature* 63, no. 2 (2025): 440–514.

Fikir Worku Edossa, Joachim Gassen, and Victor S. Maas. “Using Large Language Models to Explore Contextualization Effects in Economics-Based Accounting Experiments. ” *SSRN Electronic Journal, 2024*

Gächter, Simon, Daniele Nosenzo, and Martin Sefton. “Peer Effects in Pro-Social Behavior: Social Norms or Social Preferences?” *Journal of the European Economic Association* 11, no. 3 (2013): 548–73.

Goette, Lorenz, David Huffman, and Stephan Meier. “The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups.” *American Economic Review* 96, no. 2 (2006): 212–16.

Hagenbach, Jeanne, and Rachel Kranton. “Competition, Cooperation, and Social

Perceptions." *The Economic Journal*, Oxford University Press, 2025, ueaf032.

Hoppe, E. I., and P. W. Schmitz. "Contracting under Incomplete Information and Social Preferences: An Experimental Study." *The Review of Economic Studies* 80, no. 4 (2013): 1516–44.

Horton, John J. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? " *National Bureau of Economic Research*, 2023.

Kahneman, D. "Thinking, Fast and Slow" Macmillan, 2011.

Kranton, Rachel, Matthew Pease, Seth Sanders, and Scott Huettel. "Deconstructing Bias in Social Preferences Reveals Groupy and Not-Groupy Behavior." *Proceedings of the National Academy of Sciences* 117, no. 35 (2020): 21185–93.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment." *Organizational Behavior and Human Decision Processes* 151 (March 2019): 90–103.

Mahmud, Hasan, A.K.M. Najmul Islam, Xin (Robert) Luo, and Patrick Mikalef. "Decoding Algorithm Appreciation: Unveiling the Impact of Familiarity with Algorithms, Tasks, and Algorithm Performance." *Decision Support Systems* 179 (April 2024): 114168.

March, Christoph. "Strategic Interactions between Humans and Artificial Intelligence: Lessons from Experiments with Computer Players." *Journal of Economic Psychology* 87 (December 2021): 102426.

McLeish, Kendra N., and Robert J. Oxoby. "Identity, Cooperation, and Punishment." *SSRN Electronic Journal*, ahead of print, 2007.

Mei, Qiaozhu, Yutong Xie, Walter Yuan, and Matthew O. Jackson. "A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans." *Proceedings of the National Academy of Sciences* 121, no. 9 (2024): e2313925121.

Rand, David G., Joshua D. Greene, and Martin A. Nowak. "Spontaneous Giving and Calculated Greed." *Nature* 489 (2012): 427–430.

Tajfel, Henri, Michael G Billig, Robert P Bundy, and Claude Flament. “Social Categorization and Intergroup Behaviour.” *European Journal of Social Psychology* 1, no. 2 (1971): 149–78.

Van Leeuwen, Boris, and Ingela Alger. “Estimating Social Preferences and Kantian Morality in Strategic Interactions.” *Journal of Political Economy Microeconomics* 2, no. 4 (2024): 665–706.