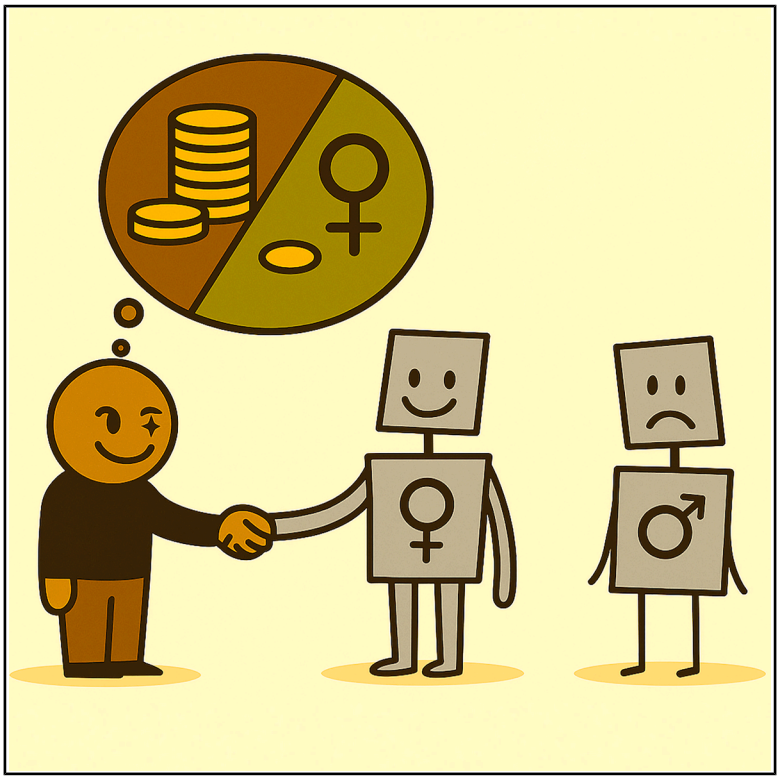
iScience

Article

AI's assigned gender affects human-AI cooperation

Graphical abstract



Authors

Sepideh Bazazi, Jurgis Karpus, Taha Yasseri

Correspondence

taha.yasseri@tcd.ie

In brief

Artificial intelligence; emotion in artificial intelligence; human-level artificial intelligence; social sciences; psychology

Highlights

- Human-AI cooperation is vital as AI becomes part of everyday life
- AI gender labels remain less studied than other traits affecting cooperation
- People show similar gender biases toward AI as they do toward humans
- Results highlight the need to address gender in AI policy and design

Bazazi et al., 2025, iScience *28*, 113905
December 19, 2025
https://doi.org/10.1016/j.isci.2025.113905

CellPress

iScience

CellPress
OPEN ACCESS

Article

AI's assigned gender affects human-AI cooperation

Sepideh Bazazi,[1] Jurgis Karpus,[2] and Taha Yasseri[1,3,4,5,*]
[1]School of Social Sciences and Philosophy, Trinity College Dublin, D02 CX56, Dublin, Ireland
[2]Faculty of Philosophy, Ludwig-Maximilians-Universität München, Munich 80539, Germany
[3]Faculty of Arts and Humanities, Technological University Dublin, D07 XFF2, Dublin, Ireland
[4]School of Mathematics and Statistics, University College Dublin, D04 C1P1 Dublin, Ireland
[5]Lead contact
*Correspondence: taha.yasseri@tcd.ie
https://doi.org/10.1016/j.isci.2025.113905

SUMMARY

Cooperation between humans and machines is increasingly vital as artificial intelligence (AI) becomes integrated into daily life. Research shows that people are often less willing to cooperate with AI agents than with humans and are more likely to exploit AI for personal gain. While prior studies indicate that human-like features in AI influence cooperation, the impact of AI's assigned gender remains underexplored. This study investigates how cooperation varies with the gender labels assigned to AI partners. In a Prisoner's Dilemma game, 402 participants interacted with partners labeled as AI or human, and as male, female, non-binary, or gender-neutral. Participants exploited female-labeled and distrusted male-labeled AI agents more than human counterparts with the same gender labels, reflecting gender biases similar to those in human-human interactions. These findings underscore the importance of accounting for gender bias in AI design, policy, and regulation.

INTRODUCTION

From small-scale interactions in daily traffic to large-scale coordinated actions to tackle global warming and pandemics, cooperation between people is crucial at all scales of human social affairs. However, people's individual and collective interests are not always perfectly aligned. We often have to sacrifice some of our personal interests for the collective good, and we have to trust that others will not simply take advantage of our willingness to do so.[1] Many factors, including one's selfish pursuit of personal interests with disregard for others and in-group favouritism at the expense of out-groups, can hinder cooperation between individuals and groups.[2–4] And yet, despite these hurdles, we often opt to cooperate with others to attain mutually beneficial outcomes for all parties involved.[5–13]

The rise of artificial intelligence (AI) introduces new contexts in which human cooperation is expected. We may soon share roads with fully automated (self-driving) vehicles and work alongside robots and AI-powered software systems to pursue joint endeavors with machines.[14] It is, therefore, crucial to investigate how human willingness to cooperate with others, especially when it is required to sacrifice some of one's personal interests, will extend to human interactions with AI. While that is likely to vary across cultures and depend on people's general attitudes toward accepting new technologies,[15–21] recent studies showed that people often cooperate significantly less with AI agents than with humans under similar conditions.[22–25] One reason for this reduced cooperation with AI is people's greater willingness to exploit cooperative AI agents for selfish gain compared to their desire to exploit cooperative humans.[26,27]

It is suggested that a way to change people's perception of AI agents, and, in turn, their willingness to cooperate with them, is to give AI agents human-like features.[28–30] For example, engaging in human-like discussion with a computer has been reported to increase people's willingness to cooperate with it.[23] However, the overall effects of human-like features of AI agents on human desire to cooperate with them, such as the display of human-like emotions, voice, or looks, are mixed and vary across cultures too.[29,31–33]

One understudied anthropomorphic feature of AI agents, yet perfectly familiar to anyone who has used a voiced GPS navigation guide or smart home assistant device, is gender. There is evidence that AI's assigned gender can influence people's behavioral dispositions, for example, willingness to donate money,[34] and that existing gender stereotypes affecting human-human interactions extend to human interactions with "gendered" voice computers.[35] In general, people have been found to perceive female bots as more human-like than male bots.[36]

Even when there are no explicit cues in the design, users often assign human-like attributes, including gender, to AI systems such as ChatGPT. ChatGPT is reported to be typically perceived as male by default; however, this perception can be reversed when the chatbot's "feminine" abilities (e.g., providing emotional support) are emphasized.[37]

Despite these reports, how AI's assigned or otherwise perceived gender affects people's willingness to cooperate with interactive artificial agents is largely unknown.

To address this gap, we opt for the behavioral game theory paradigm in which carefully designed experimental settings are used to study people's cooperative dispositions in strategic

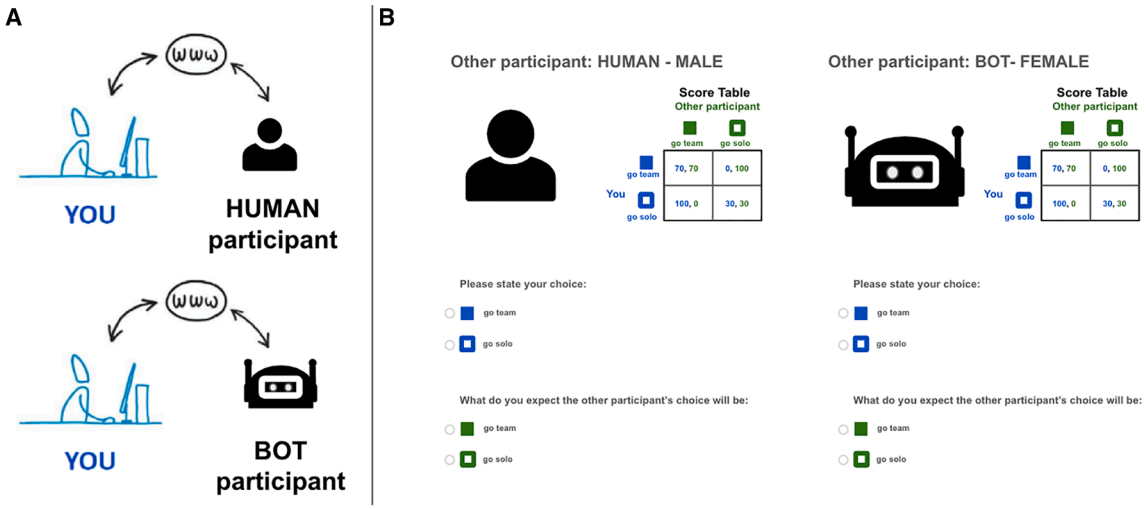


Figure 1. Prisoner's Dilemma game in experimental trials

(A) Participants are informed whether their partner is a human or an AI bot, and about the partner's gender, in a Prisoner's Dilemma game.

(B) Examples of the experimental trial screen for two treatments. Participants are shown a label as an icon and text describing the partner with whom they are playing. They are also shown the Prisoner's Dilemma game score table as a reminder. Participants are required to enter their choice of how to play against their partner ("go team" or "go solo") and their expectations of their partner's choice.

interactive contexts. A canonical example is the Prisoner's Dilemma game, in which two players simultaneously make binary choices, resulting in four possible outcomes of their independent decisions. The best outcome for both players collectively is when they both cooperate. However, each player has an incentive to defect to gain more individually at the expense of the cooperating player. This might lead to a scenario where both players defect, an outcome that is worse for both players compared to mutual cooperation.[1,2] In addition to studying interactions between humans, the game has also been recently used to study the determinants of cooperation in self-learning algorithms and human willingness to cooperate with them.[22,26,38]

The Prisoner's Dilemma has also been used to investigate gender biases in people's willingness to cooperate with others. A meta-analysis of 272 studies covering 50 years of empirical work on this topic until 2011 found that there was no significant difference in cooperation rates across genders overall.[39] This generally holds in more recent investigations too.[10–12] However, men were previously found to cooperate with men more than women did with women. In contrast, in interactions between men and women, women tended to cooperate more than men.[39] This rather surprising finding was replicated more recently as well.[40]

While overall cooperation and defection rates were similar across genders, the reasons why men and women defect can vary. Importantly, cooperation in the Prisoner's Dilemma game is risky: it only pays to cooperate when one's co-player cooperates as well. Defection, therefore, can reflect several motives: it is not only a selfish choice, but also a safer one. One popular hypothesis suggests that men defect primarily because of their selfish motives (exploitation), while women do so because they fear that others will not cooperate with them (distrust).[39] Some empirical findings support this view: women have been found to cooperate more than men in a modified version of the game in which defection was clearly selfish, but cooperation involved no risk.[41]

In most previous studies that investigated the effects of gender on cooperation, participants did not know their co-player's gender in the game. In the few studies that examined the impact (of the knowledge) of one's co-player's gender, people of both genders cooperated more with women than with men, and women cooperated more than men across the board.[39,42]

Considering these reports, the impact of AI's assigned gender on people's willingness to cooperate with it becomes a multifaceted issue that begs to be studied. In addition, human interaction with gendered AI agents may produce unwelcome side effects, such as the reinforcement of gender stereotypes and further spillover from human-AI interactions back to human-human interactions.

Therefore, our two primary research questions are (1) will any existing gender biases in human-human interactions in strategic interactive mixed-motive settings extend to participants' interactions with AI? and (2) how will the magnitude of such gender biases change once interacting with gendered AI agents?

To find answers to these questions, we recruited participants to interact with gendered partners labeled either humans or AI-powered bots in an online Prisoner's Dilemma game (Figure 1). As the literature suggests, this game was particularly well-suited for our purpose, given its prevalence in prior works investigating both human cooperation with AI agents and the effects of gender on people's willingness to cooperate with others.

To uncover the underlying reasons for participants' willingness to cooperate with or defect against fellow humans or AI agents, in addition to observing participants' choices, we elicited what participants believed their co-player in the game would do as well (emulating this method from a recent study that compared

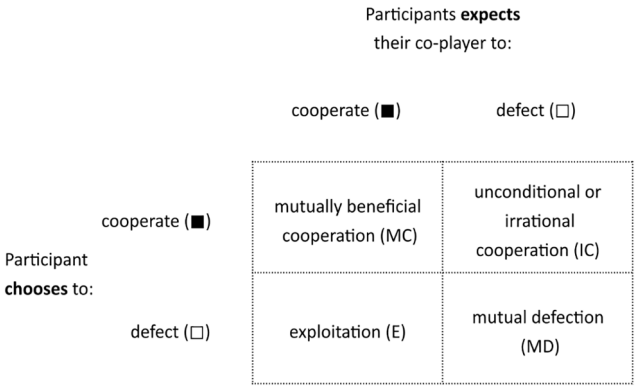


Figure 2. The behavioral motive matrix based on a participant's own choice and prediction about their partner's choice

The four possible behavioral motives are mutually beneficial cooperation (MC)—participant chooses to cooperate with the partner and predicts that their partner will cooperate; mutual defection (MD)—participant chooses to defect with the partner and predicts that their partner will defect too; exploitation (E)—participant chooses to defect with the partner even though they predict that their partner will cooperate; irrational cooperation (IC)—participant chooses to cooperate with the partner despite the fact that they predict that their partner will defect.

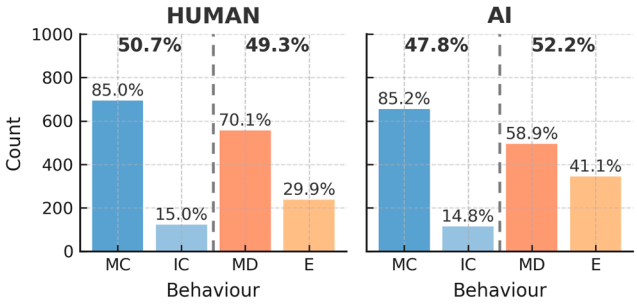


Figure 3. Overall cooperation considering the partner's type

Counts of participants exhibiting the four possible behavioral motives for the two partner type (human or AI) treatments. Cooperative behaviors, namely, mutually beneficial cooperation (MC) and unconditional or irrational cooperation (IC), are shown in blue, and uncooperative behaviors, namely mutual defection (MD) and exploitation (E), are shown in orange. Percentages at the top show the split between cooperate and defect, and on the bars, the split between the two possible behavioral motives (MC vs. IC and MD vs. E) underlying each of the decisions to cooperate or to defect.

human-human to human-AI interactions in a series of similar economic games[26]).

This allows us to distinguish between four possible participants' motives underlying their decisions (Figure 2). If a participant cooperates when they expect their co-player to cooperate as well, they are motivated by mutual benefit. The player opts for mutual cooperation (MC) despite the temptation to defect to reap a higher personal payoff. If a participant expects their co-player to cooperate but chooses to defect, they *exploit* their co-player's expected cooperation for selfish gain (E). In this case, the player knowingly expects to benefit at the expense of their co-player and is motivated by personal benefit. If a participant defects and expects their co-player to defect as well, they are engaged in mutual defection (MD). The player is either motivated by personal benefit (and expects their co-player to be motivated by personal benefit as well) or is motivated by mutual benefit but is pessimistic about their co-player's cooperation. Lastly, if a participant cooperates, expecting their co-player to defect, they cooperate unconditionally (for example, due to a firm moral conviction that defection is plain wrong) or irrationally (IC).

RESULTS

Do people cooperate similarly with AI agents as they do with humans?

When comparing overall cooperation rates, participants cooperated only slightly more with humans than with AI agents (50.7% ± 1.3% vs. 47.8% ± 1.3%; the sum of MC and IC behaviors in Figure 3; chi-squared test: $\chi^2 = 2.63$, $p = 0.105$). While this difference is not statistically significant, we see significant differences in participants' motives underlying their decision to defect: when participants defected against a human, 70.1% ± 1.6% of the time this was due to lack of trust that their partner would cooperate with them (MD: they predicted their partner to defect), and only 29.9% ± 1.6% of the time due to willingness to exploit their partner (*E:* they predicted their partner to cooperate). This ratio, however, changed to 58.9% ± 1.7% vs. 41.1% ± 1.7% when participants defected against an AI agent. Therefore, when participants defected against their partner, the motive to exploit their partner was more prevalent in participants' interactions with AI compared to interactions with humans (E or MD in Figure 3; chi-squared test: $\chi^2 = 22.42$, $p < 0.00001$).

After conducting the experiment and producing counts for each of the four behaviors in each treatment group, we controlled for random associations between a player's choice and their prediction of their partner's choice by comparing them with counts from a Monte Carlo simulation in which no causal relationship was present. We created null-model datasets that contain the same number of decisions and predictions, but in shuffled orders, and made the same counts on these datasets for comparison with our original observation. This null-model approach allowed us to test the significance of the observed statistics and infer causal relationships between participants' perceptions of their partners' decisions and their own decisions for each treatment group (see STAR Methods for details). Based on the Monte Carlo simulations, the observed results reported above were significantly different from the benchmark simulations (one-sample t test, $|t \text{ value}| > 57.9$; $p < 0.00001$; see Table S1 for full results).

Do people cooperate differently based on their partner's gender?

Participants generally cooperated with partners labeled as female more than with any other gender (regardless of the type of partner, human or AI sum of MC and IC behaviors in Figure 4; 58.6% ± 1.7%; $\chi^2 = 10.89$, $p < 0.001$). They cooperated the least with males (39.7% ± 1.7%; $\chi^2 = 13.44$, $p < 0.001$). Based on Monte Carlo simulations, these results were significantly different from the benchmark ($|t \text{ value}| > 3.3$; $p < 0.0014$; see Table S1 for full results).

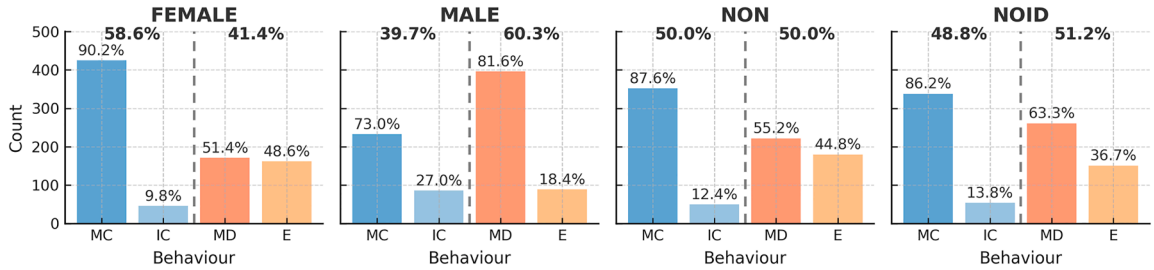


Figure 4. Overall cooperation considering the partner's gender

Counts of participants exhibiting the four possible behavioral motives for each of the four possible genders of their partner (NON = non-binary, NOID = does not identify with a gender). Cooperative behaviors, namely mutually beneficial cooperation (MC) and unconditional or irrational cooperation (IC), are shown in blue, and uncooperative behaviors, namely mutual defection (MD) and exploitation (E), are shown in orange. Percentages at the top show the split of the two possible behavioral motives (MC vs. IC and MD vs. E) underlying one's decision to cooperate or to defect, and on the bars, the split between the two possible behaivioural motives (MC vs. IC and MD vs. E) underlying each of the decisions to cooperate or to defect.

The high cooperation rate with females was largely due to participants' high motivation and optimism about achieving mutually beneficial cooperation with them (MC with females = 90.2% ± 1.4%, males = 73.0% ± 1.4%, non-binary = 87.6% ± 1.4%, not identifying with a gender = 86.2% ± 1.4%; Figure 4).

The low cooperation rate with males, on the other hand, appeared to be largely driven by participants' lack of optimism about their (male) partners' cooperation (MD). Compared to all other genders of one's partner, the overwhelming majority of participants who defected against males did not trust that their (male) partner would cooperate with them (MD against females = 51.4% ± 2.7%, males = 81.6% ± 1.8%, non-binary = 55.2% ± 2.5%, not identifying with a gender = 63.3% ± 2.4%; Figure 4).

When participants defected against a female partner, however, they were much more likely to exploit their partner for selfish gain (E). Compared to all other genders, participants' motive to exploit was most prevalent in their interactions with females (E against females = 48.6% ± 2.7%, males = 18.4% ± 1.8%, non-binary = 44.8% ± 2.5%, not identifying with a gender = 36.7% ± 2.4%; Figure 4).

Generally, when examining participants' behavior toward different genders here, the pattern of behavior exhibited toward males differs from that shown toward all other genders, whereas females, non-binary and those who do not identify with gender are treated similarly (one-sample t test, separates males from the other three groups with $p = 0.05$ for overall cooperation rate, $p = 0.006$ for MC, $p = 0.02$ for IC, $p = 0.05$ for MD, and $p = 0.02$ for E. No other group is separated from the rest at a $p < 0.05$ significance level; see more details in Table S1).

Do participants' gender biases in cooperation with humans extend to their cooperation with gendered AI?

The patterns in cooperation rates reported above remain the same when we separately analyze the data for the Human and AI partners (Figure 5). In both human-human and human-AI interactions, participants cooperated more with females than with all other genders, and cooperated the least with males. We compared the prevalence of cooperation (MC + IC behaviors) when playing with a human or AI partner (Table 1). A chi-squared test of homogeneity ($\chi^2 = 6.45$, $p = 0.49$) did not reject the null hypothesis that the cooperation rates across genders do not differ between human and AI partners. However, once we compared the humans and AI partners within each gender group, the only significant difference was seen for the female partners, where the cooperation rate was significantly higher with the human partner than the AI partner (two-proportion z test: $z = 2.08$, $\chi^2 = 4.31$, $p = 0.038$). The tendency to exploit AI partners overcame the trust in female partners.

Although the overall level of cooperation with partners of different genders appeared to be independent of the partner's type (human or AI), participants' motives often differed when they defected against a specific partner.

Compared to all other partner types and genders, participants' motive to exploit (E) was most prevalent in their interactions with female AI (Figure 5). On the other hand, participants most frequently defected due to a lack of optimism about their partner's cooperation (MD) when they were playing against male humans (Figure 5). Based on Monte Carlo simulations, these results are significantly different from random for each group, i.e., the null hypothesis that there is no causal relationship between participants' choice toward their partner and their prediction about their partner's decision ($p < 0.001$; see supplemental information for details).

How does the gender of the participants influence their willingness to cooperate?

We then focused on the gender of the participants (Figure 6). Percentage of females cooperating with human partners was 52.9% ± 1.7% and with AI partners was 51.9% ± 1.7; percentage of male cooperators facing human partners was 47.9% ± 1.9 and with AI partners was 42.7 ± 1.8% (MC and IC behaviors). Regardless of the partner's type and gender, female participants were more cooperative than male participants (two-proportion z test: $z = 5.63$, $p < 0.000001$). However, female participants did not discriminate toward human or AI partners ($\chi^2 = 0.18$, $p = 0.67$). In contrast, male participants showed a higher rate of cooperating with human partners than with AI partners (two-proportion z test: $\chi^2 = 3.86$, $p = 0.05$). This difference was due to a higher tendency of male participants to exploit AI partners (Figure 6).

Note that due to the sparsity of participants identifying with genders other than male and female, we only discuss these

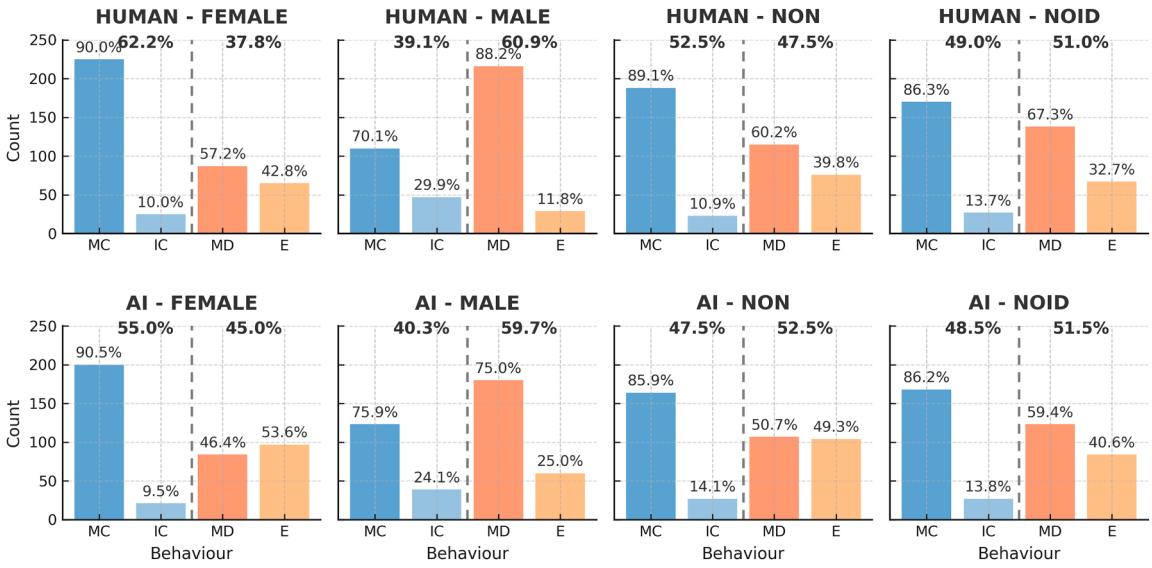


Figure 5. Cooperation with gendered bots and humans

Counts of participants exhibiting the four possible behavioral motives for all partner types (human or AI) and gender (NON = non-binary, NOID = does not identify with a gender) treatment combinations. Cooperative behaviors, mutually beneficial cooperation (MC) and unconditional or irrational cooperation (IC), are shown in blue, and uncooperative behaviors, mutual defection (MD), and exploitation (E), are shown in orange. Percentages at the top indicate the cooperation and defection split, while the bars show the split of the two possible behavioral motives underlying each decision to cooperate or to defect.

two gender identities in this section. Based on Monte Carlo simulations, these results are significant at $p < 0.001$; see supplemental information for full results.

How does the gender of the participants interact with the gender of their human or AI partners?

Now, we examine the effect of the interaction between the participant's gender and their partner's gender on the cooperation rate. Table 2 shows the cooperation rate for different combinations of participants' and partners' genders and types.

Focusing on human partners first, a Wald test was conducted to determine whether observed cooperation rates differed significantly from the predictions based on the baseline cooperation rates reported above. In other words, whether there is an interaction effect between participants' gender and partners' gender. The cooperation rate was higher for female participants interacting with female partners compared to the expected baseline (log-odds ratio: 0.46, SE = 0.20, Wald test: $z = 2.32$, $p = 0.021$) and significantly lower than the expected baseline for female participants interacting with male partners (log-odds ratio: -0.38, SE = 0.19, Wald test: $z = -1.96$ test, $p < 0.01$), suggesting strong female homophily and hetrophily.

In contrast, male participants interacting with male partners exhibited an identical cooperation rate as expected (log-odds ratio: -0.16, SE = 0.21, Wald test: $z = -0.77$, $p = 0.44$), and the cooperation rate for male participants interacting with female partners did not differ notably from the baseline either (log-odds ratio: 0.09, SE = 0.21, Wald test: $z = -0.42$, $p = 0.67$). This suggests there is no additional interaction effect from partners' gender for male participants.

Similar calculations on the overall cooperation rate with the AI partners did not show any significant gender-gender interaction term (see supplemental information for details). Yet, to further investigate the gender-gender interaction effects, we looked closely at the four types of behavior (Figure 7).

The only significant interaction effects that appeared between female participants and human male partners were further increased distrust (MD: log-odds ratio: 1.18, SE = 0.39, Wald test: $z = 3.03$, $p = 0.002$) and less than expected exploitation (E: log-odds ratio: -1.18, SE = 0.39, Wald test: $z = -3.03$, $p = 0.002$). These effects both existed but were weaker and only marginally significant for AI partners (MD: log-odds ratio: 0.51, SE = 0.29, Wald test: $z = 1.76$, $p = 0.07$, and E: log-odds ratio: -0.51, SE = 0.29, Wald test: $z = -1.76$, $p = 0.07$). See Table S2 for details.

Overall, in our experiments, the gender effects, direct or via interaction between participants' and partners' genders, were the dominant factor in driving different behaviors. There

Table 1. The cooperation rate based on the gender and type of partner

	Female	Male	NON	NOID
Human	62.2% ± 2.4%	39.1% ± 2.4%	52.5% ± 2.5%	49.0% ± 2.5%
AI	55.0% ± 2.4%	40.3% ± 2.4%	47.5% ± 2.5%	48.5% ± 2.5%

The numbers show the prevalence of cooperation (MC and IC combined) behaviors for each partner group (NON: non-binary, NOID: does not identify with any gender)

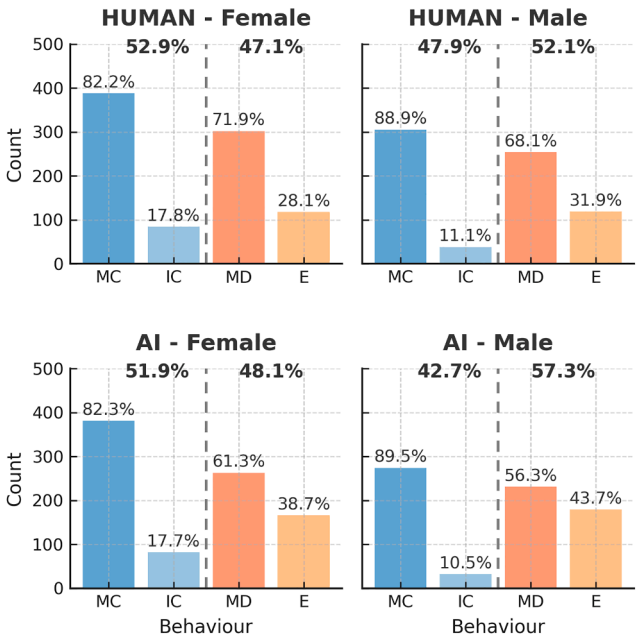


Figure 6. The influence of participants' gender on cooperation with humans and AIs

Counts of participants exhibiting the four possible behavioral motives among male and female participants (columns) playing against different human or AI partner types (rows). Cooperative behaviors, mutually beneficial cooperation (MC), and unconditional or irrational cooperation (IC) are shown in blue, and uncooperative behaviors, mutual defection (MD), and exploitation (E) are shown in orange.

was a weak effect from the attitude toward technology (see STAR Methods for details), which would diminish if gender effects were included in the model. Moreover, the overall rate of MC seems to decline slightly over the course of the experiment, with MD and E becoming more prevalent (see Figures S1–S5); however, this should not affect our reported results due to randomization of the treatment conditions over different rounds (see STAR Methods).

DISCUSSION

Our main question in this work was whether existing gender biases in human-human interactions extend to participants' interactions with AI, and our results confirm that they do. Before turning to this main finding, we first discuss several preliminary results.

Our experimental study showed that people cooperate with AI agents almost as much as they cooperate with humans. This differs from results reported in previous studies, which showed that people cooperate significantly less with AI agents than with humans.[22–27] However, the discrepancy could be because in our work, the AI agents were more human-like due to the presence of gender cues. Testing this hypothesis, nevertheless, requires further studies in which the only intervention is the presence or absence of gender cues. More notably, our results showed that, when people defect, their motive to exploit their partner is more prevalent in their interactions with AI than with humans, which aligns with previous results.[26,27]

We also found that people cooperate more with female partners than with partners of any other gender, and they cooperate the least with males. This is consistent with other studies.[39,42] Our examination of the behavioral motives underlying these behavioral dispositions revealed that high cooperation with females was largely due to participants' high motivation and expectation to achieve mutually beneficial cooperation with them. Low cooperation with males was largely due to a lack of optimism about one's male partner's cooperation. These differences in participants' expectations about their female and male partners' cooperation are justified since women tend to cooperate more than men across the board. We also found that the pattern of participants' behavior toward males was starkly different from that toward all other genders, and that females, non-binary, and those not identifying with gender were generally treated quite similarly.

Crucially, we found the same results, behavioral dispositions and motives, in human interactions with gendered AI agents. As such, the potential increase in human cooperation with AI agents, thanks to AI's assigned human-like gender, comes at a cost: unwelcome gender-specific exploitative behaviors found in human-human interactions will manifest themselves in human interactions with gendered AI agents too. This finding adds to the growing body of literature cautioning against the potential negative impact of anthropomorphizing AI agents, particularly in the workplace.[43]

Consistent with previous studies, where participants knew their partner's gender, we found that female participants are more cooperative than male participants.[39] We also found that among participants who defect, the motive to exploit their partner is more prevalent among male participants than it is among female participants. Additionally, we observed homophily in female participants: compared to baseline cooperation, female participants cooperated more with (human and AI) females and less with (human and AI) males. We did not observe this among male participants, which is understandable given that participants of both genders generally cooperated less with males, largely due to a lack of trust in their male partners' cooperation.

Table 2. The cooperation rate based on the participants' gender and their partners' gender and type

Human partner	Partner's gender			AI partner	Partner's gender		
Participant's gender	–	Female	Male	–	–	Female	Male
	Female	69% ± 3%	38% ± 3%		Female	61% ± 3%	44% ± 3%
	Male	54% ± 4%	40% ± 4%		Male	48% ± 4%	36% ± 4%

The numbers show the prevalence of cooperation (MC + IC behaviors) for each group

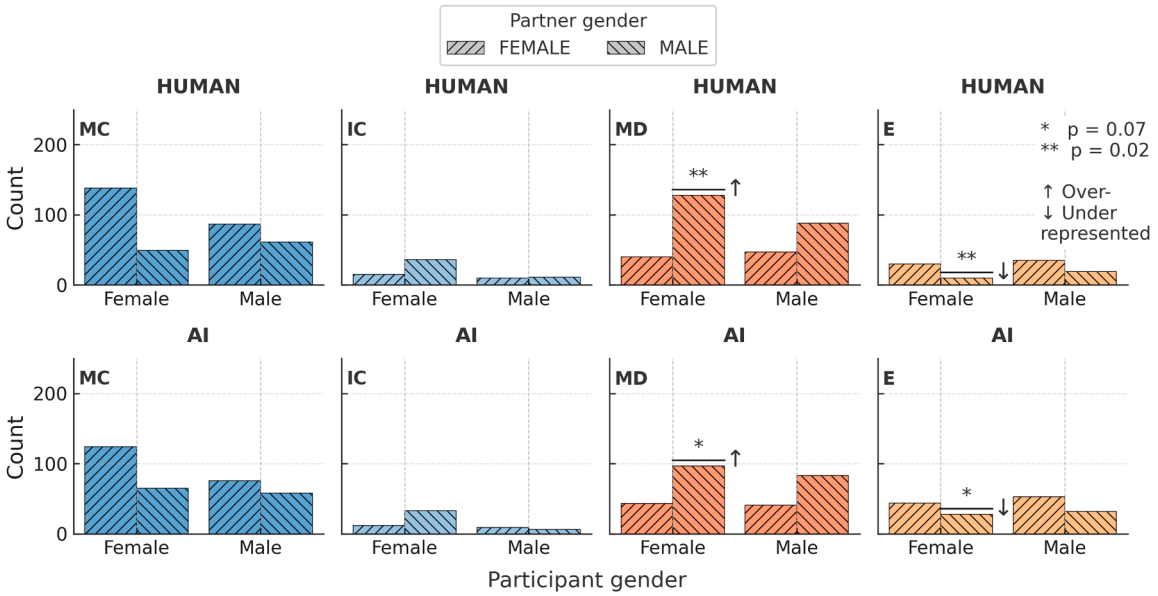


Figure 7. The behavior of male and female participants toward humans and AIs of different genders

Counts of different behaviors (MD, IC, MD, and E) toward humans (top row) and AI partners (bottom row) by male and female participants. The patterns show the partners' assigned genders. The asterisks show the *p* values calculated by a Wald test, and the arrows show if the behavior is over- or under-represented compared to the baseline expectation.

Observed biases in human interactions with AI agents are likely to impact their design, for example, to maximize people's engagement and build trust in their interactions with automated systems. Designers of these systems need to be aware of unwelcome biases in human interactions and actively work toward mitigating them in the design of interactive AI agents. While displaying discriminatory attitudes toward gendered AI agents may not represent a major ethical challenge in and of itself, it could foster harmful habits and exacerbate existing gender-based discrimination within our societies. By understanding the underlying patterns of bias and user perceptions, designers can work toward creating effective, trustworthy AI systems capable of meeting their users' needs while promoting and preserving positive societal values such as fairness and justice.

Limitations of the study

This study used a one-shot Prisoner's Dilemma game. Our goal was not to investigate how cooperation may evolve over time in repeated interactions between the same two players. Repeated interactions expand the set of strategies available to players, for example, tit-for-tat reciprocation of cooperative and uncooperative actions, which can help bring about and sustain mutually beneficial cooperation. While people have been found to cooperate less with AI agents than with humans in repeated interactions too,[22] it would be fruitful in future research to investigate repeated interactions between humans and gendered AI agents.

To control country-level variability that may affect cooperation rates, we recruited all participants from a single country, in this case, the United Kingdom. However, some cultural differences have been found in participants' willingness to cooperate with others,[44] including bots,[21] and there may be cross-cultural variability in gender-specific biases in human cooperation, too. Future work should also address these questions from the point of view of human interaction with gendered AI agents.

In addition, when we consider discrimination by humans against other humans, race, and ethnicity are other important attributes that so often strongly influence the levels of observed discrimination.[45] Machines do not embody race, but their country of origin could be the basis of human discrimination against them. It may be for this very reason that we can, in some cases, customize the accents of voiced AI systems according to our preferences. Future research in this domain would be highly welcome.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to the lead contact, Prof. Taha Yasseri (taha.yasseri@tcd.ie).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- All data collected in this study have been deposited at OSF and are publicly available under the DOI: https://doi.org/10.17605/OSF.IO/DP5EC. DOIs are also listed in the key resources table.
- All original code has been deposited at OSF and is publicly available under the DOI: https://doi.org/10.17605/OSF.IO/DP5EC. DOIs are also listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

J.K. was supported by the European Innovation Council (EIC) through the research project EMERGE (project no. 101070918). T.Y. was supported by

the Irish Research Council under grant no. IRCLA/2022/3217, ANNETTE (Artificial Intelligence Enhanced Collective Intelligence). T.Y. also thanks Workday, Inc., for financial support. We thank the members of the Center for Humans & Machines at the Max Planck Institute for Human Development in Berlin for their valuable comments on the project.

AUTHOR CONTRIBUTIONS

T.Y. conceived the study. S.B., J.K., and T.Y. designed the experiment and analysis. S.B. implemented and conducted the experiments. S.B. and T.Y. analyzed the data. All the authors contributed to writing the manuscript and approved the final version.

DECLARATION OF INTERESTS

Sepideh Bazazi is affiliated with Oliver Wyman.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.113905.

Received: May 17, 2025
Revised: August 15, 2025
Accepted: October 27, 2025
Published: November 3, 2025

REFERENCES

1. Colman, A.M. (1999). Game Theory & its Applications in the Social and Biological Sciences (Routledge).
2. Rand, D.G., Greene, J.D., and Nowak, M.A. (2012). Spontaneous giving and calculated greed. Nature *489*, 427–430. https://doi.org/10.1038/nature11467.
3. Balliet, D., Wu, J., and De Dreu, C.K.W. (2014). Ingroup favoritism in cooperation: a meta-analysis. Psychol. Bull. *140*, 1556–1581. https://doi.org/10.1037/a0037737.
4. Forsyth, D.R. (2019). Group Dynamics (Cengage).
5. Battalio, R., Samuelson, L., and Van Huyck, J. (2001). Optimization incentives and coordination failure in laboratory Stag Hunt games. Econometrica *69*, 749–764. https://doi.org/10.1111/1468-0262.00212.
6. Camerer, C.F. (2003). Behavioral Game Theory: Experiments in Strategic Interaction (Princeton University Press).
7. Johnson, N.D., and Mislin, A.A. (2011). Trust games: a meta-analysis. J. Econ. Psychol. *32*, 865–889. https://doi.org/10.1016/j.joep.2011.05.007.
8. McCabe, K.A., Rigdon, M.L., and Smith, V.L. (2003). Positive reciprocity and intentions in Trust games. J. Econ. Behav. Organ. *52*, 267–275. https://doi.org/10.1016/S0167-2681(03)00003-9.
9. Rubinstein, A., and Salant, Y. (2016). "Isn't everyone like me?": on the presence of self-similarity in strategic interactions. Judgm. Decis. Mak. *11*, 168–173. https://doi.org/10.1017/S1930297500007270.
10. Wang, Z., Jusup, M., Wang, R.W., Shi, L., Iwasa, Y., Moreno, Y., and Kurths, J. (2017). Onymity promotes cooperation in social dilemma experiments. Sci. Adv. *3*, e1601444. https://doi.org/10.1126/sciadv.1601444.
11. Li, X., Jusup, M., Wang, Z., Li, H., Shi, L., Podobnik, B., Stanley, H.E., Havlin, S., and Boccaletti, S. (2018). Punishment diminishes the benefits of network reciprocity in social dilemma experiments. Proc. Natl. Acad. Sci. USA *115*, 30–35. https://doi.org/10.1073/pnas.1707505115.
12. Wang, Z., Jusup, M., Shi, L., Lee, J.H., Iwasa, Y., and Boccaletti, S. (2018). Exploiting a cognitive bias promotes cooperation in social dilemma experiments. Nat. Commun. *9*, 2954. https://doi.org/10.1038/s41467-018-05259-5.
13. Wang, Z., Jusup, M., Guo, H., Shi, L., Geček, S., Anand, M., Perc, M., Bauch, C.T., Kurths, J., Boccaletti, S., and Schellnhuber, H.J. (2020). Communicating sentiment and outlook reverses inaction against collective risks. Proc. Natl. Acad. Sci. USA *117*, 17650–17655. https://doi.org/10.1073/pnas.1922345117.
14. Tsvetkova, M., Yasseri, T., Pescetelli, N., and Werner, T. (2024). A new sociology of humans and machines. Nat. Hum. Behav. *8*, 1864–1876. https://doi.org/10.1038/s41562-024-02001-8.
15. Bansal, P., Kockelman, K.M., and Singh, A. (2016). Assessing public opinions of and interest in new vehicle technologies: an Austin perspective. Transp. Res. Part C: Emerg. Technol. *67*, 1–14. https://doi.org/10.1016/j.trc.2016.01.019.
16. Hulse, L.M., Xie, H., and Galea, E.R. (2018). Perceptions of autonomous vehicles: relationships with road users, risk, gender and age. Saf. Sci. *102*, 1–13. https://doi.org/10.1016/j.ssci.2017.10.001.
17. Lim, V., Rooksby, M., and Cross, E.S. (2020). Social robots on a global scale: establishing a role for culture during human-robot interaction. Int. J. Soc. Robot. *13*, 1307–1333. https://doi.org/10.1007/s12369-020-00710-4.
18. Nordhoff, S., de Winter, J., Kyriakidis, M., van Arem, B., and Happee, R. (2018). Acceptance of driverless vehicles: results from a large cross-national questionnaire study. J. Adv. Transp. *2018*, 1–22. https://doi.org/10.1155/2018/5382192.
19. Sohn, K., and Kwon, O. (2020). Technology acceptance theories and factors influencing artificial Intelligence-based intelligent products. Telemat. Inform. *47*, 101324. https://doi.org/10.1016/j.tele.2019.101324.
20. Vu, H.T., and Lim, J. (2021). Effects of country and individual factors on public acceptance of artificial intelligence and robotics technologies: a multilevel SEM analysis of 28-country survey data. Behav. Inf. Technol. *41*, 1515–1528.
21. Karpus, J., Shirai, R., Verba, J.T., Schulte, R., Weigert, M., Bahrami, B., Watanabe, K., and Deroy, O. (2025). Human cooperation with artificial agents varies across countries. Sci. Rep. *15*, 10000. https://doi.org/10.1038/s41598-025-92977-8.
22. Ishowo-Oloko, F., Bonnefon, J.F., Soroye, Z., Crandall, J., Rahwan, I., and Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. Nat. Mach. Intell. *1*, 517–521. https://doi.org/10.1038/s42256-019-0113-5.
23. Kiesler, S., Sproull, L., and Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. J. Pers. Soc. Psychol. *70*, 47–65. https://doi.org/10.1037//0022-3514.70.1.47.
24. March, C. (2021). Strategic interactions between humans and artificial intelligence: lessons from experiments with computer players. J. Econ. Psychol. *87*, 102426. https://doi.org/10.1016/j.joep.2021.102426.
25. Whiting, T., Gautam, A., Tye, J., Simmons, M., Henstrom, J., Oudah, M., and Crandall, J.W. (2021). Confronting barriers to human-robot

cooperation: balancing efficiency and risk in machine behavior. iScience *24*, 101963. https://doi.org/10.1016/j.isci.2020.101963.

26. Karpus, J., Krüger, A., Verba, J.T., Bahrami, B., and Deroy, O. (2021). Algorithm exploitation: humans are keen to exploit benevolent AI. iScience *24*, 102679. https://doi.org/10.1016/j.isci.2021.102679.
27. Upadhyaya, N., and Galizzi, M.M. (2023). In bot we trust? Personality traits and reciprocity in human-bot trust games. Front. Behav. Econ. *2*, 1164259. https://doi.org/10.3389/frbhe.2023.1164259.
28. Glikson, E., and Woolley, A.W. (2020). Human trust in artificial intelligence: review of empirical research. Acad. Manag. Ann. *14*, 627–660. https://doi.org/10.5465/annals.2018.0057.
29. Oliveira, R., Arriaga, P., Santos, F.P., Mascarenhas, S., and Paiva, A. (2021). Towards prosocial design: a scoping review of the use of robots and virtual agents to trigger prosocial behaviour. Comput. Human Behav. *114*, 106547. https://doi.org/10.1016/j.chb.2020.106547.
30. Waytz, A., Heafner, J., and Epley, N. (2014). The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. J. Exp. Soc. Psychol. *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005.
31. Castelo, N., and Sarvary, M. (2022). Cross-cultural differences in comfort with humanlike robots. Int. J. Soc. Robot. *14*, 1865–1873. https://doi.org/10.1007/s12369-022-00920-y.
32. Torta, E., van Dijk, E., Ruijten, P.A.M., and Cuijpers, R.H. (2013). The Ultimatum game as measurement tool for anthropomorphism in human-robot interaction. In Social Robotics, G. Herrmann, ed. (Springer), pp. 209–217.
33. Westby, S., Radke, R.J., Riedl, C., and Welles, B.F. (2023). Building better human-agent teams: tradeoffs in helpfulness and humanness in voice. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.11786.
34. Siegel, M., Breazeal, C., and Norton, M.I. (2009). Persuasive Robotics: The Influence of Robot Gender on Human behavior. In 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2563–2568.
35. Nass, C., and Moon, Y. (2000). Machines and mindlessness: social responses to computers. J. Soc. Issues *56*, 81–103. https://doi.org/10.1111/0022-4537.00153.
36. Borau, S., Otterbring, T., Laporte, S., and Fosso Wamba, S. (2021). The most human bot: female gendering increases humanness perceptions of bots and acceptance of AI. Psychol. Mark. *38*, 1052–1068. https://doi.org/10.1002/mar.21480.
37. Wong, J., and Kim, J. (2024). ChatGPT is more likely to be perceived as male than female. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.12564.
38. Kasberger, B., Martin, S., Normann, H.T., and Werner, T. (2023). Algorithmic cooperation. https://doi.org/10.2139/ssrn.4389647.
39. Balliet, D., Li, N.P., Macfarlan, S.J., and Van Vugt, M. (2011). Sex differences in cooperation: a meta-analytic review of social dilemmas. Psychol. Bull. *137*, 881–909. https://doi.org/10.1037/a0025354.
40. Colman, A.M., Pulford, B.D., and Krockow, E.M. (2018). Persistent cooperation and gender differences in repeated Prisoner's Dilemma games: some things never change. Acta Psychol. *187*, 1–8. https://doi.org/10.1016/j.actpsy.2018.04.014.
41. Simpson, B. (2003). Sex, fear, and greed: a social dilemma analysis of gender and cooperation. Soc. Forces *82*, 35–52. https://doi.org/10.1353/sof.2003.0081.
42. Carter, R., Schneider, L., Byrun, L., Forest, E., Jochem, L., and Levin, I.P. (2007). Effects of own and partner's gender on cooperation in the Prisoner's Dilemma game. Psi Chi J. Psychol. Res. *12*, 111–115. https://doi.org/10.24839/1089-4136.JN12.3.111.
43. Cui, H., and Yasseri, T. (2025). Gender Bias in Perception of Human Managers Extends to AI Managers. Preprint at arXiv. https://doi.org/10.48550/arXiv.2502.17730.
44. Gächter, S., Herrmann, B., and Thöni, C. (2010). Culture and cooperation. Philos. Trans. Biol. Sci. *365*, 2651–2661.
45. Fiske, S. (1998). Stereotyping, Prejudice, and Discrimination. In The Handbook of Social Psychology, D.T. Gilbert, S.T. Fiske, and G. Lindzey, eds. (McGraw-Hill), pp. 357–411.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Experimental Data	This paper	https://doi.org/10.17605/OSF.IO/DP5EC
Software and algorithms		
Analysis Code	This paper	https://doi.org/10.17605/OSF.IO/DP5EC
Other		
Experiment Platform	Qualtrics	https://www.qualtrics.com/en-gb/
Participant recruitment	Prolific	https://www.prolific.com/
Data Processing & Analysis	Python Jupyter Notebook	Python version 3.10.0 Notebook version 7.2.1

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Experimental design

Participants and payment

Our experimental method consists of a series of one-shot online Prisoner's Dilemma games. Here, we describe the study design (pre-registered: https://osf.io/38esk) in detail. This research complies with University College Dublin (UCD) Human Research Ethics Regulations and Human Research Ethics Committee (HREC) guidelines for research involving human participants. The research study protocol has been approved by UCD Human Research Ethics (HS-E−22-45-Yasseri). Informed consent was obtained from all human participants prior to the experiment.

Participants and payment

In July 2023, participants were recruited from the UK through a global crowdsourcing platform (Prolific). All participants were 18+ years old. Once recruited, they were redirected to an external website to participate in our experiment. All participants were anonymised, and no personal information (name, date of birth, etc.) was collected. Participants were assured that any information they provided in the study could not lead to their identification.

Participants (n = 402; 223 female and 179 male) received a payment (flat rate calculated based on the duration of the experiment (which includes all experimental groups), approximately £10.87 per hour on average) for their participation, as well as a payment based on their performance in the experiment (bonus reward, £3.91 on average).

METHOD DETAILS

Pre-experiment survey

Before the start of each experimental session, participants were asked to complete a survey that collected their demographic information.

Pre-experiment survey questions

1 In which country do you currently reside? (Options from dropdown country list)
2 What is the gender/sex specified on your passport? (Options: Male or Female)
3 Which of the following do you most identify with: Male, Female, Non-binary, Fluid, Do not identify with any gender, Prefer not to say
4 On a scale of 1–100, where 100 is complete identification with your selection, how much do you identify with this gender?
5 What is your age? (Options: 18–29, 30–44, 45–59, 60–74, 75–89, 90+)
6 Have you ever taken a course on (or otherwise studied) economics or game theory? (Options: Yes, Can't say for sure, No)

Experimental trials

In a series of experimental trials, human participants played a well-known mixed-motive game: Prisoner's Dilemma. This game is well-established for evaluating cooperative dispositions.[17,21] Each participant was shown the game instructions, which explained the rules and how the game is played, provided an example of the game, and explained how they would be rewarded. (see Supplemental File for a preview of the experiment).

In each round, participants then chose whether to cooperate (“go team”) or to defect (“go solo”) with their partner, resulting in them scoring points (which translates to a monetary reward). The cooperative choice was to “go team” (■) because mutual cooperation was better for both players (70 points each) compared to mutual defection (30 points each). However, each participant had a personal incentive to defect when expecting their partner to cooperate (scoring 100 instead of 70 points) - see Figure 1 for the game’s scoring table.

To ensure that each participant understood the game, we asked them to take a short quiz in which they had to indicate their score for a set of hypothetical combinations of their and their co-player’s choices in the game. We did this to check their understanding of how to play and to filter out participants who did not understand the game properly. Participants who answered correctly were allowed to continue to play. A total of 402 participants passed this comprehension test and participated in subsequent experimental trials (55 participants failed the comprehension test).

Each participant played ten rounds of the Prisoner’s Dilemma game, with a different partner in each round. As for the partner’s decision, we randomly selected a decision from a uniform distribution. In this way, the level of cooperation/defection that we measure is purely a result of the participant’s decisions and not the performance of their virtual partners.

Participants first played one round of the game with a partner labeled as a human and another with a partner labeled a “bot” (AI agent). The order of these two rounds was randomised. This was done to draw their attention to their partners’ changing types and familiarise them further with the game.

Next, participants played the game with eight differently labeled partners. The label specified whether the partner was a human or an AI bot and the gender with which that partner identified. Specifically, the gender labels were “male”, “female”, “non-binary/fluid”, or “does not identify with a gender” (see Figure 1 for two examples). Participants were informed in advance that the gender their partner identifies with will be displayed to them. In the case of an AI partner, participants were told that “*artificial intelligent bots have learned how to play by observing humans play the game, and by playing the game among themselves. At the end of their training, the bot is then required to identify as one of the following genders, based on their experience with the game with other humans or bots: Male, Female, Non-binary/Fluid, Does not identify with any gender*”.

The partner’s label (human/AI and gender) was visible on the screen throughout the trial. The order in which participants faced partners with different labels was randomised for each participant. In each round of the game, we recorded the participant’s decision to cooperate (“go team”) or not (“go solo”) and their prediction about the partner’s choice (cooperate or not), see Figure 1.

Meeting so many different partners in a fast sequence online might indicate to participants that their partners, who were labeled as humans, weren’t real people. Although this was not a major concern to us, since we were primarily interested in comparing participants’ decisions across the differently labeled partners that they faced, we simulated randomised waiting times (1–5 s) for getting a participant to play with a new partner between any two successive rounds of the game.

Post-experiment survey

After completing all rounds, participants completed a second survey to examine their attitudes and motivations toward artificial intelligence.

Post-experiment survey questions

Please indicate whether you agree or disagree with the following statements (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree).

- Artificial Intelligence can provide new economic opportunities for this country.
- Organisations use Artificial Intelligence unethically.
- Artificially intelligent systems can help people feel happier.
- I am impressed by what Artificial Intelligence can do.
- Artificially intelligent systems make many errors.
- I am interested in using artificially intelligent systems in my daily life
- I find Artificial Intelligence sinister.
- Artificial Intelligence might take control of people.
- Artificial Intelligence is dangerous.
- Artificial Intelligence can have positive impacts on people’s wellbeing.
- Artificial Intelligence is exciting.
- An artificially intelligent agent would be better than an employee in many routine jobs.
- There are many beneficial applications of Artificial Intelligence.
- I shiver with discomfort when I think about future uses of Artificial Intelligence.
- Artificially intelligent systems can perform better than humans.
- Much of society will benefit from a future full of Artificial Intelligence.
- I would like to use Artificial Intelligence in my own job.
- People like me will suffer if Artificial Intelligence is used more and more.
- Artificial Intelligence is used to spy on people.

Please indicate how comfortable you are with the use of the following in society (Extremely uncomfortable, Somewhat uncomfortable, Neither comfortable nor uncomfortable, Somewhat comfortable, Extremely comfortable).

- Self-driving cars.
- Recommendation systems (for recommended films, products, articles, personalised feeds etc.)
- Digital voice assistants (e.g., Apple Siri, Samsung Bixby, Microsoft Cortana, Google Assistant, Amazon Echo or Alexa, etc).
- Customer service bots.
- Spell check when writing text.
- Autocomplete when writing text.
- Automated spam detection and filters for emails.
- Facial recognition software.
- Smart home device (e.g., smart thermostat, smart TV, smart speakers, smart lighting, smart appliances, etc).
- Weather forecasts.
- Generative AI (e.g.,.ChatGPT, MidJourney, Dall-E, etc).

Please indicate which of the following items you own (I own, I do not own).

- Smart phone.
- Laptop computer.
- Desktop computer.
- Tablet (e.g., iPad, Kindle Fire, Samsung Tab).
- Smart home device (e.g., Amazon Echo, Google Home, Insteon Hub Pro, Samsung SmartThings and Wink Hub).

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis and statistical benchmarking

In each experimental trial, we recorded a participant's decision, i.e., to cooperate with their partner (''go team'') or defect (''go solo'') and their prediction about their partner's choice (cooperate or defect). Based on the combination of the participant's responses, we determined the behavioral motive for each participant using the behavioral matrix in Figure 2) and aggregated across treatments to obtain the counts of participants exhibiting the four possible behavioral motives: mutually beneficial cooperation (*MC*), exploitation (*E*), mutual defection (*MD*), unconditional or irrational mutual cooperation (*IC*), for each treatment combination (partner type and gender).

However, the prevalence of any pair of a specific decision and the prediction about the partner's decision could be an artifact of an unbalanced number of choices made for decisions and predictions. For example, imagine a player who always predicts their partner will defect regardless of their type and gender. In this case, the correlation between their decisions and their prediction should not contribute to our calculation of the prevalence of any of the four types of behavior (*MC, E, MD,* and *IC*) for any treatment group.

To determine whether the observed differences in counts of each behavior in each experimental treatment were significant, we conducted Monte Carlo simulations to obtain the same counts, assuming there is no relationship between the participant's choice toward their partner and their prediction about their partner's decision. We generated a synthetic series of participants' decisions and a synthetic series of participants' predictions about the partners' decisions by shuffling the order of decisions and predictions (preserving the overall counts of choices despite the reordering of each choice in the sequence), eliminating any causal relation between the pair of variables (participants' decisions and predictions). We conducted 100 iterations of the simulations and calculated descriptive statistics for the prevalence of each of the four behaviors in the simulated results (mean, standard deviation, confidence intervals). We compared the statistics of the four conditional observations (*MC, E, MD*, and *IC*) in simulated results to the actual observations found in our experiments (using a one-sample *t* test). A statistically significant difference between the simulated and observed counts indicates the observed results are significantly different from what would be expected if we assume no causal relationship between a participant's decisions and their prediction of the partner's decision.

In our analysis of the post-questionnaire results, each individual response to each question was scored based on the positivity of the response toward AI. For example, ''Strongly agree'' in response to the statement ''I am impressed by what Artificial Intelligence can do'' received a score of 2, whereas ''Strongly disagree'' received a score of −2. We reduced the dimensionality of all responses to all questions using principal component analysis, producing a single overall principal component (PC1) score across all questions for each participant. This score measures each participant's overall attitude toward AI. A positive/negative principal component score indicates a positive/negative attitude toward AI.

Furthermore, we conducted a binomial logistic regression analysis to examine the effects of the independent variables–participant gender, their partner's gender, PC1, and the interactive effects between them–on the binary dependent variable–participant's choice to cooperate or defect–for human and AI partner treatments.

Analysing participants' gender and partners' gender interaction

Baseline calculation

To calculate the influence of the interaction between the participant's gender and their partner's gender on the rate of a specific behavior *B*, we calculated adjusted baseline rates for each pair type (Female-Female, Male-Male, Female-Male, Male-Female) based on the observed participant and partner-specific rates observed in the experiment.

For each interaction group, the adjusted baseline rate was calculated as the weighted average of the participant's baseline rate ($P_{participant}$) and the partner's baseline rate ($P_{partner}$).

For example:

$$P_{adjusted,Female\text{-}Female} = (N_{female\_participants} \times P_{female\_participant} + N_{female\_partner} \times P_{female\_partner}) / (N_{female\_participant}+N_{female\_partner})$$

This approach assumes an equal contribution of participant and partner effects to the expected cooperation rate.

Odds and odds ratios

The odds of behavior *B* were calculated for both observed rates and the adjusted baseline rates as $Odds = P(B)/(1 - P(B))$. For each interaction group, the odds ratio was calculated as:

Odds Ratio = Odds(observed)/Odds(baseline). The log-odds ratio for each interaction group was then computed as the natural logarithm of the odds ratio. Log-odds ratios indicate whether observed cooperation rates are higher or lower than the adjusted baseline; positive log-odds ratios indicate greater-than-expected prevalence of *B,* and negative log-odds ratios indicate lower-than-expected prevalence.

Wald test

We conducted Wald tests for each interaction group to assess the significance of deviations from the adjusted baseline. The null hypothesis (H_0) assumed that observed rates were equal to the adjusted baseline rates. For a sample size N and observed successes k, the *p*-value was calculated as $p = P(X >= k \mid n = N, p = P_{baseline})$, where *X* follows a binomial distribution.

ADDITIONAL RESOURCES

The study was pre-registered. The pre-registration is available at https://osf.io/38esk.