iScience

CellPress
OPEN ACCESS

Article

Rewards and punishments help humans overcome biases against cooperation partners assumed to be machines

Kinga Makovi,[1,5,*] Jean-François Bonnefon,[2] Mayada Oudah,[1] Anahit Sargsyan,[1,3] and Talal Rahwan[4,*]
[1]Social Science Division, New York University Abu Dhabi, Abu Dhabi, UAE
[2]Toulouse School of Economics, CNRS (TSM-R), University of Toulouse Capitole, Toulouse, France
[3]School of Social Sciences and Technology, Technical University of Munich, München, Germany
[4]Computer Science, Science Division, New York University Abu Dhabi, Abu Dhabi, UAE
[5]Lead contact
*Correspondence: km2537@nyu.edu (K.M.), talal.rahwan@nyu.edu (T.R.)
https://doi.org/10.1016/j.isci.2025.112833

SUMMARY

High levels of human-machine cooperation are required to combine the strengths of human and artificial intelligence. Here, we investigate strategies to overcome the machine penalty, where people are less cooperative with partners they assume to be machines, than with partners they assume to be humans. Using a large-scale iterative public goods game with nearly 2,000 participants, we find that peer rewards or peer punishments can both promote cooperation with partners assumed to be machines but do not overcome the machine penalty. Their combination, however, eliminates the machine penalty, because it is uniquely effective for partners assumed to be machines and inefficient for partners assumed to be humans. These findings provide a nuanced road map for designing a cooperative environment for humans and machines, depending on the exact goals of the designer.

INTRODUCTION

Artificial intelligence (AI) is often framed as being in competition with human intelligence, in the sense that it aims at surpassing human performance on some benchmark, or at defeating humans in zero-sum competitions.[1–4] In parallel, though, there has been a growing awareness that intelligent machines and humans may also be able to initiate and sustain cooperation,[5] in order to achieve together more than what they could achieve on their own. Because it takes two to cooperate, human-machine cooperation is both a technological and psychological challenge: we may need to endow machines with cooperative abilities—but, as we do in this article, we may also seek to understand and overcome the reluctance of people to cooperate with partners they assume (correctly or incorrectly) to be machines. Some evidence for this reluctance comes from industrial settings[6] or customer relations,[7,8] but the bulk of the evidence comes from the same kind of lab-based, incentivized games that have long been used to study human-human cooperation.[9,10] Experimental studies have repeatedly pointed to the existence of a "machine penalty"[11]: participants in one-shot[12–15] and repeated[16–18] games (e.g., dictator games, ultimatum games, public good games, and prisoners' dilemmas) show non-zero cooperation with partners they assume to be machines, but this cooperation is significantly lower than what they show to partners they assume to be humans. The machine penalty is conceptually unrelated to the aversion that people have to let machines make autonomous moral decisions,[19] or to the aversion that people have to use forecasting algorithms as decision aids,[20,21] but it does belong with these two phenomena, to the family of behaviors where people show a gamut of less positive reactions to machines than the reactions they show to humans.

A machine penalty occurs when people show lower cooperation with partners they "assume" to be humans. The word "assume" is important here, both conceptually and methodologically, because properly measuring the machine penalty can require deceiving experimental participants into thinking they are interacting with machines. If one shows that cooperation decreases when people interact with machine partners, one cannot be sure whether this is due to the limitations of the machine (the technical challenge, i.e., choosing the appropriate actions to elicit cooperation), or to the machine penalty (the psychological challenge, i.e., the lower willingness people have to cooperate with partners they assume to be machines). This is a strong concern in repeated games, in which idiosyncratic machine behavior can derail cooperation, independently of any psychological reluctance to cooperate with machines. The methodological solution to this problem is to compare the behavior of participants who interact with humans they assume to be humans, to the behavior of participants who interact with humans they assume to be machines. This implies deceiving some participants to believe that they are interacting with machines, while they are, in fact, interacting with humans. Since all cooperation actually happens between humans, a decrease in cooperation when participants assume their partners to be

machines can be unambiguously attributed to the machine penalty. (While experimental deception is primarily motivated here by methodological considerations, it resembles real-world situations in which machines and humans interact under possible misattribution, that is, when people are not sure whether they can assume their partners to be humans or machines, and whether their assumptions are correct; we will return to this point in the discussion section. We also discuss deception in more detail in the STAR Methods section).

Here, we show that the tools of peer reward and peer punishment, which already help instill and sustain cooperation between humans, can also be used to overcome the machine penalty. In human groups, cooperation improves when people can reward others for cooperating and punish them for defecting.[22–24] Punishment is a more delicate tool than reward, though. It does not always work as well as rewards,[25–28] its effects are not always predictable when combined with rewards,[29] and people typically show greater reluctance to deliver punishment, compared to rewards.[30–33] One reason for this reluctance, and for the volatile effects of punishment, is that punishment may backfire among humans. It can be perceived as malicious, create resentment, foster a sense of injustice, and start a cycle of retaliation.[34,35] Importantly, however, these concerns can be eliminated when partners are assumed to be machines, because machines have no perceived emotionality or intentionality.[36,37] As a result, punishment may be less of an unpredictable tool when partners are assumed to be machines. In this article, we show that rewards and punishments both increase cooperation with partners assumed to be machines, and that their combination has unique effectiveness in this context, which results in the elimination of the machine penalty. We provide an analysis of the behavior underlying this unique effectiveness, but we acknowledge from the outset that this analysis is exploratory: we had no *a priori* expectation to observe this unique effectiveness, and our experimental design, while constructed to test the effect of rewards, punishments, and their combination, was not constructed to confirm the mechanistic explanation we offer based on our observation of the descriptive data.

RESULTS

Participants in groups of four interacted through an iterated (also often referred to as repeated) public goods game (IPGG). While single-shot public goods games have been used to study public good provision and its impediments,[38,39] IPGGs have been deployed to study how cooperation evolves over time for the provision of public goods,[40] and their external validity has been demonstrated in the context of fishing[41] and managing common forests.[42] In an IPGG, a group of individuals, often four, are given resources in every round to decide what proportion of those resources, if any, they would allocate to a common pool. The sum of these contributions is multiplied by a factor (here, by 1.6) and divided evenly among group members at the end of each round, and a new round begins with the same individuals. These incentives encourage group members to keep all their resources and free ride on others' contributions. Although this is the dominant strategy, should everyone engage in it, the whole group is worse off, and only low levels of the public good are provided.

As we already stated, one problem when studying the machine penalty is that lower cooperation when people play with machines can be the result of machine behavior, over and above any effect of the machine penalty as a psychological phenomenon. In other words, people may cooperate less with machines not because they are machines but because they behave differently than humans. To remove this confounding fact, we engage in deception. All four-person groups are composed of only people. In some groups, participants are correctly informed that their three partners are humans. In the other groups, participants are told that their three partners are machines. This is the condition in which we anticipate that the machine penalty will be strongest (future work shall explore the gradual change in group composition.). We measure the machine penalty as the decrease in cooperation in groups where participants assume their partners to be machines (see the STAR Methods section for further details).

In all conditions, participants play 20 rounds of a standard IPGG. In the **baseline** condition, at the end of each round, they learn about the amount of public good that was provided, but they do not know what each player provided, and they cannot take any action before the next round begins. In the **feedback** condition, participants do learn what each player contributed at the end of each round, but cannot take any action before the next round begins. In the **reward** condition, participants get the same information as in the **feedback** condition, and then have the option of rewarding at a cost other players before the next round begins. This decision is made for each other player individually, which means that participants can give out zero to three rewards. The **punishment** condition is similar, except that the decision is not to reward, but to punish at a cost. Finally, in the **both** condition, participants can choose either to reward or to punish every other player.

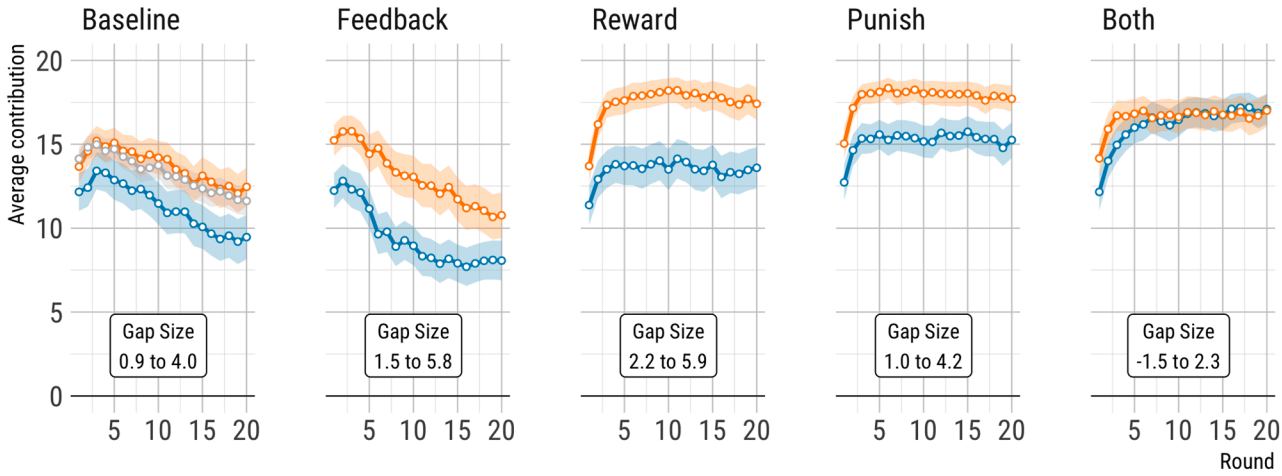
The magnitude of the machine penalty

We start by graphing contributions in the IPGG by experimental condition over the 20 rounds of decision-making in Figure 1. Before calculating the machine penalty, we note that behavior in the groups where participants believe players to be humans is similar to what has been documented in previous research.[23,29] Using random-effects regression models for participants and groups, and fixed effects for rounds, we estimate the machine penalty, that is, the gap in contributions between groups where partners are assumed to be humans, and groups in which partners are assumed to be machines. The text boxes in Figure 1 report the 95% confidence intervals for this gap in each condition. Complete regression tables can be found in the Tables S1–S5.

In the **baseline** condition, a machine penalty of 2.4–4.0 tokens emerges on average across rounds (n = 684 individuals, 171 groups). For exploratory purposes, we included groups in the **baseline** condition in which participants were told that only 1–2 other players were machines, instead of being told that all other players were machines. Cooperation in these groups is shown as a gray line in the baseline panel of Figure 1 and is no different than in groups where all players are assumed to be humans. The machine penalty is conserved in the **feedback** condition, with a gap of 1.5–5.8 tokens (n = 328 individuals, 82 groups). Introducing rewards or punishments improves cooperation compared to **baseline**,[29] but the machine penalty is still

Absolute and relative effects of rewards and punishments

Rewards and punishments independently increase and stabilize contribution to the public good, without closing the gap between groups where partners are assumed to be humans and groups where partners are assumed to be machines. The gap is only closed when rewards and punishment are both available.



The ribbons show the 95% confidence interval of the contribution.
The boxes display the 95% confidence interval of the machine penalty in each group.
The grey line in the Baseline panel shows the average of treatments with intermediate numbers of purported machines
It suggests that the machine penalty only emerges when players believe all other participants are machines

Figure 1. Contributions to the public good over rounds, by experimental condition and purported group composition

conserved, for 2.2–5.9 tokens with rewards (n = 328 individuals, 82 groups) and 1.0–4.2 tokens with punishments (n = 324 individuals, 81 groups). However, when both rewards and punishments are available, the machine penalty is finally suppressed: we find no credible evidence of a gap, with a 95% confidence interval for the gap of −1.5–2.3 (n = 324 individuals, 81 groups).

When partners are assumed to be machines, the effect of rewards and punishment add up: their combination leads to higher contributions (by 1.8 tokens on average, $p = 0.025$) than when they are used in isolation (based on a multilevel model with random intercept for groups and participants, fixed effect of round, comparing the **both** treatment to the pooled **reward** and **punishment** treatments). When partners are assumed to be humans, rewards and punishments do not add up. Their combination does not lead to a credible increase in contributions compared to their use in isolation (decrease of 1 token on average, $p = 0.170$). In other words, combining rewards and punishments is inefficient when partners are assumed to be humans but useful when partners are assumed to be machines—which allows to close the gap in contributions between the two types of groups, when using rewards and punishments in combination. We now explore possible mechanisms for this effect, by considering in turn how participants hand out rewards and punishments across conditions, depending on the purported composition of their group; and how participants respond to rewards and punishments across conditions, also depending on the purported composition of their group.

Frequency of rewards and punishments

There are 240 opportunities to hand out a reward for each group of 4 participants playing for 20 rounds. When only rewards are available (i.e., no opportunity for punishment), participants who assume that their partners are humans use rewards 176 times on average, which corresponds to a 73% reward rate. In the **both** condition where punishments are also available, the reward rate drops to 46%. When participants assumed that their partners were machines, the reward rate is 47% when only rewards are available, and drops to 31% in the **both** condition where punishments are also available. We fitted a multilevel model in which the binary outcome was to hand out a reward, and the predictors were the purported composition of the group, the experimental condition, and their interaction; with random intercepts for groups and participants, and a fixed effect of round. This model detected the expected effects of the purported group composition ($z = -4.20$, $p<0.001$) and experimental condition ($z = -4.76$, $p<0.001$) but no interaction effect ($z = 1.39$, $p = 0.165$) see Table S6. In other words, rewards are less frequent when participants assume their partners to be machines, less frequent when punishments are also made available, but there is no credible evidence that the availability of punishment has different effects in the two types of groups—which means in turn that the frequency of reward giving is an unlikely mechanism to explain the closure of the machine penalty in the **both** condition.

There are 240 opportunities to mete out punishment for each group of 4 participants playing for 20 rounds. However, in

Reactions to rewards

When only rewards are available, participants assuming partners to be machines attempt to reduce their contribution after receiving 2 or 3 rewards in a given round. This behavior disappears when punishment is available, and does not appear in groups where participants assume partners to be humans.

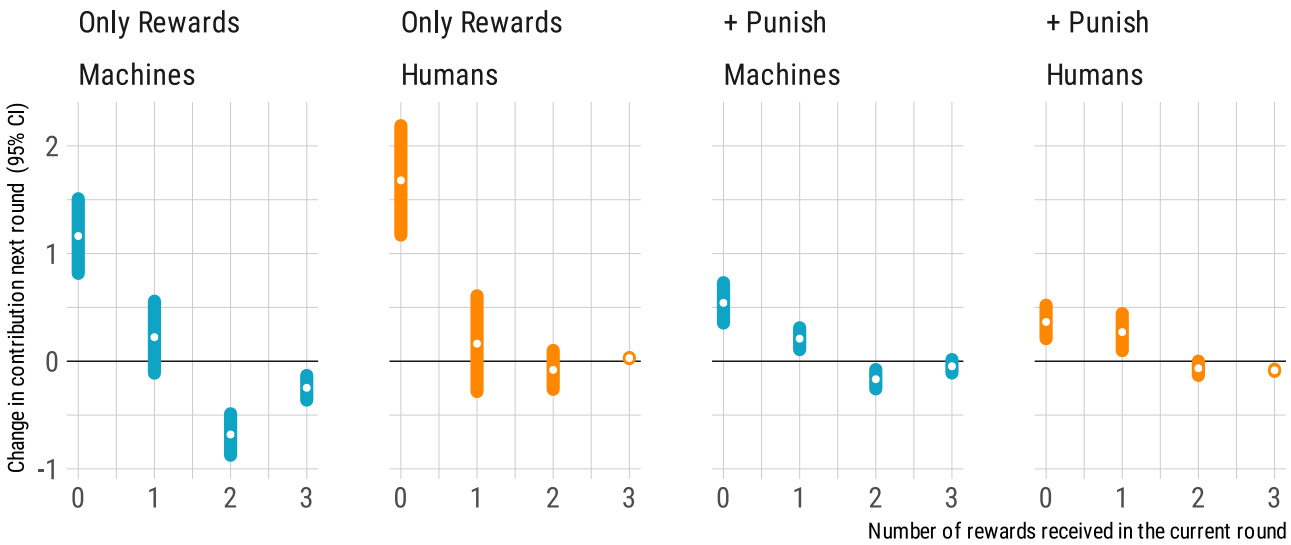


Figure 2. Change in contributions in the next round (95% confidence intervals) conditional on the number of rewards received in the current round, across experimental treatments and group compositions

practice and according to previous research,[28,43] punishment is extremely rare in our experiment. When only punishments are available (i.e., no opportunity for rewards), participants who assume their partners to be humans punish 5 times on average, which corresponds to a 2% punishment rate, and the punishment rate is 3% when rewards are also available. When partners are assumed to be machines, the punishment rate is about 4%, whether rewards are available or not. Due to the extreme rarity of punishment, our multilevel model did not converge—but given this rarity across all conditions and purported group composition, the frequency of punishment is, again, an unlikely mechanism to explain the closure of the machine penalty in the **Both** condition.

Reaction to rewards

In the punishment-only treatment, participants tentatively increase their contributions to avoid punishment. This increase is the same whether they assume their partners to be machines or humans, so while average contributions shift upward overall, the relative difference between contributions in the two purported groups (the machine penalty) remains unchanged. In other words, fear of punishment raises contributions but does not eliminate the machine penalty. When rewards are available, participants tentatively increase their contributions to earn rewards. In a given round, a participant can receive from zero to three rewards from other players. Figure 2 shows how participants react to the number of rewards they receive—specifically, how they adjust their contribution in the next round.

The left panel of Figure 2 displays the behavior of participants when only rewards are available (and punishment is not). Participants who do not receive any rewards increase on average their contribution during the next round, by one or two tokens, regardless of the purported composition of their group. Participants who assume their partners to be human behave cautiously and maintain their contribution level after receiving one to three rewards, suggesting they have found a stable contribution level that earns rewards reliably. In contrast, participants who believe their partners to be machines are more exploratory, and decrease their contribution when receiving two to three rewards, as if they were testing whether they can continue to earn rewards with a lower contribution. Thus, contributions generally increase in the reward-only treatment, but the machine penalty remains because participants are more conservative when playing with purported humans, and more exploratory when playing with purported machines.

As shown in the right panel of Figure 2, this asymmetry disappears when both punishments and rewards are available. While this interpretation is entirely exploratory, it appears that the possibility of punishment discourages participants from testing whether they can earn rewards with lower contributions. As a result, contributions increase because of the appeal of rewards, and participants simultaneously stop testing partners they assume to be machines, which stabilizes contributions to the same levels in the two purported groups.

We tested whether the pattern of behavior described above was detected by a multilevel model. We stress again that this

analysis should be considered exploratory, rather than confirmatory, since we conducted it after discovering the behavior of participants from the descriptive data. The model aims to predict the change in contribution in the next round, with our usual set of predictors: experimental condition, group composition, their interaction, random intercept for groups and participants, and fixed effect of round. Data are restricted to observations where participants receive more than one reward, since this is the relevant subset of observations according to the descriptive data in Figure 2. The model detects an effect of purported group composition (participants increase their contribution less when they assume their partners to be machines, $z = -4.3$, $p<0.001$) but most importantly an interaction effect between experimental condition and purported group composition ($z = 2.3$, $p = 0.022$), reflecting the positive effect of having both punishments and rewards on the contribution change when partners are assumed to be machines (see Table S7). This substantiates the descriptive findings.

DISCUSSION

To ensure that humans and machines cooperate smoothly, we may need to not only endow machines with cooperative abilities but also understand how to encourage humans to cooperate with machines—in particular, we may want to encourage humans to cooperate with machines to the same extent that they cooperate with other humans, which requires us to overcome the machine penalty. Some previous attempts at removing the machine penalty tried to humanize machines, following the logic that people would cooperate more with machines if the machines were more human-like.[44] For example, in the context of human-robot cooperation, robots were given eyes or emotional displays[45,46]; stylized, stereotyped gender cues such as body shape or hair length[47,48]; or a fully humanoid appearance.[49,50] This strategy has not proven to be very effective[11] and can be problematic, in particular when it exploits and possibly amplifies pre-existing gender biases.[51] An extreme form of the humanization strategy is to allow machines to pretend to be humans, for example, in online interactions in which they are allowed a human avatar.[52] This can be very effective,[16] but it goes against the transparency requirements present in many emerging codes of ethics for AI.[53–55]

Here, we explored a different strategy: since cooperation between humans can be improved with peer rewards and punishments, can these peer rewards and punishments also improve cooperation between humans and partners they assume to be machines? First, we showed that peer rewards and punishments can largely increase this cooperation in absolute terms, but not in relative terms. That is, reward and punishment increase cooperation with partners assumed to be machines, but do not close the gap in cooperation rate with partners assumed to be humans. This is already a useful set of results, if the goal is simply to increase cooperation. If the goal is to overcome the machine penalty, though, the situation is more complicated. We showed that neither rewards nor punishments, used in isolation, could close or even narrow the gap in cooperation with partners that are assumed to be humans vs. machines. However, the machine penalty is overcome when rewards and punishments are used in combination. Exploring the mechanisms for this effect, we found that combining rewards and punishments was inefficient for partners assumed to be humans (i.e., not increasing cooperation any further than rewards or punishments already have in isolation) but useful for cooperation with partners assumed to be machines. These results have implications for designing cooperative human-machine environments, which will depend on the mix of humans and machines in the environment, the assumptions people make about this mix, and the type of cooperation which is the most important to promote. Future work should also explore the impact of the incentive structure of the game deployed here by modifying e.g., the cost of punishment and/or rewards and their impact on those who receive them relative to the gains one can make in the public goods game. In addition, it should also explore if machines can elicit better outcomes with how they play repeated games, which intentionally falls outside of the scope of the present study.

Real-world environments can include a hybrid population of humans and machines, with the added complication that people may not be sure about whether they can assume others to be humans or machines. For example, machines can account for 1%–15% of users on social media platforms such as X, formerly known as Twitter,[56,57] and generate a disproportionate amount of content on important topics, such as climate,[58] vaccines,[59] and religious tensions.[60] People may have the suspicion that other users are machines,[61] but they are not very good at distinguishing human and machine users accurately.[62,63] In such contexts of uncertainty, one may want to increase cooperation between all users across the board, without seeking to close the machine penalty. In this case, one should use either reward or punishments since their combination adds complications without efficiency for human-human cooperation—and the choice will probably be rewards only, if one wants to avoid the emotional and social frictions generated by punishments.

If the goal is instead to specifically encourage cooperation between humans and partners they assumed to be machines (correctly or incorrectly), then one has to decide whether it is important or not to also close the machine penalty. If it is a requirement, that is, to make people cooperate the same regardless of whether they assume their partners to be humans or machines, then one should use a combination of rewards and punishments, as indicated by our results. If it is not a requirement to close the machine penalty, and one simply seeks to make people more cooperative with partners they assume to be machines, without reaching human-human levels of cooperation, then either reward, punishment, or their combination is appropriate. In summary, our results provide a road map for smooth human-machine cooperation, identifying the different routes we can take depending on the specific goals and constraints of this cooperation.

Limitations of the study

The study has a few key limitations. First, as discussed at length, in order to disentangle the psychological and technical reasons behind the machine penalty, we employed deception. Other work might address this problem with a different research design. Additionally, some of the results we presented here are

exploratory in nature. We hope that future work will explore these further in a confirmatory study. Furthermore, given funding limitations and the extensive resources required to carry out a large-scale data collection for a study we presented here, our work does not explore how the results might change, should the incentive structure of the public goods game be dramatically altered. This should also be investigated in future work.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Kinga Makovi (km2537@nyu.edu).

Materials availability

The study is based on an original experiment. All materials necessary to replicate the procedures are included in the supplemental information.

Data and code availability

- The data associated with the studies in this manuscript have been deposited at OSF registry and are publicly available as of the date of publication.
- No custom code was generated to analyze data collected for these studies. Standard techniques have been employed using the R statistical software (version: 3.6.3)
- We pre-registered our study on OSF registry (https://osf.io/2wjrv).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank Robert Gordon, Russell Coke, and Julie Liu for research support in developing the interface for the experiments. K.M. acknowledges the support of the Research Enhancement Funds received from NYUAD, funding from the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen, under the NYUAD Research Institute Award CG001; K.M. and T.R. acknowledge discretionary research support received from NYUAD; J.F.B. acknowledges support from the Agence Nationale de la Recherche (ANR-19-PI3A-0004 and ANR-17-EURE-0010), and the Research Foundation TSE-Partnership.

AUTHOR CONTRIBUTIONS

K.M. and T.R. conceived the study; K.M., M.O., T.R., and A.S. designed the experiments; J.-F.B., K.M., and T.R. designed the analysis; M.O. and A.S. collected the data; J.-F.B. and K.M. performed the analysis; J.F.B. and K.M. wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

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

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - The magnitude of the machine penalty
  - Frequency of rewards and punishments
  - Reaction to rewards

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.112833.

Received: October 30, 2024
Revised: February 10, 2025
Accepted: June 3, 2025
Published: June 6, 2025

REFERENCES

1. Campbell, M., Hoane Jr, A.J., and Hsu, F.-h. (2002). Deep blue. Artif. Intell. *134*, 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1.
2. Schaeffer, J., Burch, N., Björnsson, Y., Kishimoto, A., Müller, M., Lake, R., Lu, P., and Sutphen, S. (2007). Checkers is solved. Science *317*, 1518–1522. https://doi.org/10.1126/science.1144079.
3. Bowling, M., Burch, N., Johanson, M., and Tammelin, O. (2015). Heads-up limit hold'em poker is solved. Science *347*, 145–149. https://doi.org/10.1126/science.1259433.
4. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature *529*, 484–489. https://doi.org/10.1038/nature16961.
5. Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., and Graepel, T. (2021). Cooperative AI: machines must learn to find common ground. Nature *593*, 33–36. https://doi.org/10.1038/d41586-021-01170-0.
6. Villani, V., Pini, F., Leali, F., and Secchi, C. (2018). Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. Mechatronics *55*, 248–266. https://doi.org/10.1016/j.mechatronics.2018.02.009.
7. Luo, X., Tong, S., Fang, Z., and Qu, Z.F. (2019). Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. Mark. Sci. *38*, 937–947. https://doi.org/10.1287/mksc.2019.1192.
8. Giroux, M., Kim, J., Lee, J.C., and Park, J. (2022). Artificial intelligence and declined guilt: Retailing morality comparison between human and AI. J. Bus. Ethics *178*, 1027–1041. https://doi.org/10.1007/s10551-022-05056-7.
9. Makovi, K., Sargsyan, A., Li, W., Bonnefon, J.-F., and Rahwan, T. (2023). Trust within human-machine collectives depends on the perceived consensus about cooperative norms. Nat. Commun. *14*, 3108. https://doi.org/10.1038/s41467-023-38592-5.
10. 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.
11. Bonnefon, J.-F., Rahwan, I., and Shariff, A. (2024). The moral psychology of Artificial Intelligence. Annu. Rev. Psychol. *75*, 653–675. https://doi.org/10.1146/annurev-psych-030123-113559.
12. 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.
13. Nielsen, Y.A., Thielmann, I., Zettler, I., and Pfattheicher, S. (2022). Sharing money with humans versus computers: On the role of honesty-humility and (non-)social preferences. Soc. Psychol. Personal. Sci. *13*, 1058–1068. https://doi.org/10.1177/19485506211055622.
14. Von Schenk, A., Klockmann, V., and Köbis, N. (2023). Social Preferences Toward Humans and Machines: A Systematic Experiment on the Role of Machine Payoffs. Perspect. Psychol. Sci. *20*, 161–185. https://doi.org/10.1177/17456916231194949.
15. Oudah, M., Makovi, K., Gray, K., Battu, B., and Rahwan, T. (2024). Perception of experience influences altruism and perception of agency influences trust in human–machine interactions. Sci. Rep. *14*, 12410. https://doi.org/10.1038/s41598-024-63360-w.
16. 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.

17. Crandall, J.W., Oudah, M., Rahwan, I., Tennom, Ishowo-Oloko, F., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.F., Cebrian, M., Shariff, A., and Goodrich, M.A. (2018). Cooperating with machines. Nat. Commun. *9*, 233. https://doi.org/10.1038/s41467-017-02597-8.
18. Sandoval, E.B., Brandstetter, J., Obaid, M., and Bartneck, C. (2016). Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game. Int. J. Soc. Robot. *8*, 303–317. https://doi.org/10.1007/s12369-015-0323-x.
19. Bigman, Y.E., and Gray, K. (2018). People are averse to machines making moral decisions. Cognition *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003.
20. Dietvorst, B.J., Simmons, J.P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. *144*, 114–126. https://doi.org/10.1037/xge0000033.
21. Burton, J.W., Stein, M.-K., and Jensen, T.B. (2020). A systematic review of algorithm aversion in augmented decision making. J. Behav. Decis. Mak. *33*, 220–239. https://doi.org/10.1002/bdm.2155.
22. Balliet, D., Mulder, L.B., and Van Lange, P.A.M. (2011). Reward, punishment, and cooperation: a meta-analysis. Psychol. Bull. *137*, 594–615. https://doi.org/10.1037/a0023489.
23. Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. Nature *415*, 137–140. https://doi.org/10.1038/415137a.
24. Wu, J., Luan, S., and Raihani, N. (2022). Reward, punishment, and prosocial behavior: Recent developments and implications. Curr. Opin. Psychol. *44*, 117–123. https://doi.org/10.1016/j.copsyc.2021.09.003.
25. Van Dijk, E., Molenmaker, W.E., and de Kwaadsteniet, E.W. (2015). Promoting cooperation in social dilemmas: The use of sanctions. Curr. Opin. Psychol. *6*, 118–122. https://doi.org/10.1016/j.copsyc.2015.07.006.
26. Noussair, C.N., van Soest, D., and Stoop, J. (2015). Punishment, reward, and cooperation in a framed field experiment. Soc. Choice Welfare *45*, 537–559. https://doi.org/10.1007/s00355-014-0841-8.
27. Ohtsuki, H., Iwasa, Y., and Nowak, M.A. (2009). Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature *457*, 79–82. https://doi.org/10.1038/nature07601.
28. Rand, D.G., and Nowak, M.A. (2011). The evolution of antisocial punishment in optional public goods games. Nat. Commun. *2*, 434–437. https://doi.org/10.1038/ncomms1442.
29. Rand, D.G., Dreber, A., Ellingsen, T., Fudenberg, D., and Nowak, M.A. (2009). Positive interactions promote public cooperation. Science *325*, 1272–1275. https://doi.org/10.1126/science.1177418.
30. Dreber, A., Rand, D.G., Fudenberg, D., and Nowak, M.A. (2008). Winners don't punish. Nature *452*, 348–351. https://doi.org/10.1038/nature06723.
31. Jordan, J.J., Hoffman, M., Bloom, P., and Rand, D.G. (2016). Third-party punishment as a costly signal of trustworthiness. Nature *530*, 473–476. https://doi.org/10.1038/nature16981.
32. Molenmaker, W.E., de Kwaadsteniet, E.W., and van Dijk, E. (2014). On the willingness to costly reward cooperation and punish non-cooperation: The moderating role of type of social dilemma. Organ. Behav. Hum. Decis. Process. *125*, 175–183. https://doi.org/10.1016/j.obhdp.2014.09.005.
33. Molenmaker, W.E., de Kwaadsteniet, E.W., and van Dijk, E. (2016). The impact of personal responsibility on the (un) willingness to punish non-cooperation and reward cooperation. Organ. Behav. Hum. Decis. Process. *134*, 1–15. https://doi.org/10.1016/j.obhdp.2016.02.004.
34. Heffner, J., and FeldmanHall, O. (2019). Why we don't always punish: Preferences for non-punitive responses to moral violations. Sci. Rep. *9*, 13219. https://doi.org/10.1038/s41598-019-49680-2.
35. Raihani, N.J., and Bshary, R. (2019). Punishment: one tool, many uses. Evol. Hum. Sci. *1*, e12. https://doi.org/10.1017/ehs.2019.12.
36. Bigman, Y.E., Wilson, D., Arnestad, M.N., Waytz, A., and Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. J. Exp. Psychol. Gen. *152*, 4–27. https://doi.org/10.1037/xge0001250.
37. Hidalgo, C.A., Orghian, D., Canals, J.A., De Almeida, F., and Martin, N. (2021). How Humans Judge Machines (MIT Press).
38. Fowler, J.H., and Christakis, N.A. (2010). Cooperative behavior cascades in human social networks. Proc. Natl. Acad. Sci. USA *107*, 5334–5338. https://doi.org/10.1073/pnas.0913149107.
39. Tavoni, A., Dannenberg, A., Kallis, G., and Löschel, A. (2011). Inequality, communication, and the avoidance of disastrous climate change in a public goods game. Proc. Natl. Acad. Sci. USA *108*, 11825–11829. https://doi.org/10.1073/pnas.1102493108.
40. Rand, D.G., Nowak, M.A., Fowler, J.H., and Christakis, N.A. (2014). Static network structure can stabilize human cooperation. Proc. Natl. Acad. Sci. USA *111*, 17093–17098. https://doi.org/10.1073/pnas.1400406111.
41. Fehr, E., and Leibbrandt, A. (2011). A field study on cooperativeness and impatience in the Tragedy of the Commons. J. Publ. Econ. *95*, 1144–1155. https://doi.org/10.1016/j.jpubeco.2011.05.013.
42. Rustagi, D., Engel, S., and Kosfeld, M. (2010). Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management. Science *330*, 961–965. https://doi.org/10.1126/science.1193649.
43. Herrmann, B., Thöni, C., and Gächter, S. (2008). Antisocial Punishment Across Societies. Science *319*, 1362–1367. https://doi.org/10.1126/science.1153808.
44. Nielsen, Y.A., Pfattheicher, S., and Keijsers, M. (2022). Prosocial behavior toward machines. Curr. Opin. Psychol. *43*, 260–265. https://doi.org/10.1016/j.copsyc.2021.08.004.
45. Hsieh, T.-Y., and Cross, E.S. (2022). People's dispositional cooperative tendencies towards robots are unaffected by robots' negative emotional displays in prisoner's dilemma games. Cogn. Emot. *36*, 995–1019. https://doi.org/10.1080/02699931.2022.2054781.
46. De Kleijn, R., van Es, L., Kachergis, G., and Hommel, B. (2019). Anthropomorphization of artificial agents leads to fair and strategic, but not altruistic behavior. Int. J. Hum. Comput. Stud. *122*, 168–173. https://doi.org/10.1016/j.ijhcs.2018.09.008.
47. Bernotat, J., Eyssel, F., and Sachse, J. (2021). The (fe)male robot: how robot body shape impacts first impressions and trust towards robots. Int. J. Soc. Robot. *13*, 477–489. https://doi.org/10.1007/s12369-019-00562-7.
48. Eyssel, F., and Hegel, F. (2012). (S)he's got the look: Gender stereotyping of robots. J. Appl. Soc. Psychol. *42*, 2213–2230. https://doi.org/10.1111/j.1559-1816.2012.00937.x.
49. Złotowski, J., Sumioka, H., Nishio, S., Glas, D.F., Bartneck, C., and Ishiguro, H. (2016). Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. Paladyn. J. Behav. Rob. *7*, 55–66. https://doi.org/10.1515/pjbr-2016-0005.
50. Cominelli, L., Feri, F., Garofalo, R., Giannetti, C., Meléndez-Jiménez, M.A., Greco, A., Nardelli, M., Scilingo, E.P., and Kirchkamp, O. (2021). Promises and trust in human–robot interaction. Sci. Rep. *11*, 9687. https://doi.org/10.1038/s41598-021-88622-9.
51. Fossa, F., and Sucameli, I. (2022). Gender Bias and Conversational Agents: an ethical perspective on Social Robotics. Sci. Eng. Ethics *28*, 23. https://doi.org/10.1007/s11948-022-00376-3.
52. Nightingale, S.J., and Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proc. Natl. Acad. Sci. USA *119*, e2120481119. https://doi.org/10.1073/pnas.2120481119.
53. Leong, B., and Selinger, E. (2018). Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism. In Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAT'19) (ACM), pp. 1–10. https://doi.org/10.1145/3287560.3287591.
54. O'Leary, D.E. (2019). GOOGLE'S Duplex: Pretending to be human. Intell. Syst. Account. Financ. Manag. *26*, 46–53. https://doi.org/10.1002/isaf.1443.

55. Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. Nat. Mach. Intell. *1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2.

56. Schuchard, R., Crooks, A.T., Stefanidis, A., and Croitoru, A. (2019). Bot stamina: examining the influence and staying power of bots in online social networks. Appl. Netw. Sci. *4*, 55. https://doi.org/10.1007/s41109-019-0164-x.

57. Varol, O., Ferrara, E., Davis, C., Menczer, F., and Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. In Proceedings of the Eleventh International AAAI Conference on Web and Social Media, *vol 11*, pp. 280–289. https://doi.org/10.1609/icwsm.v11i1.14871.

58. Marlow, T., Miller, S., and Robert, J.T. (2020). Bots and online climate discourses: Twitter discourse on President Trump's announcement of U.S. withdrawal from the Paris Agreement. Clim. Policy *21*, 765–777. https://doi.org/10.1080/14693062.2020.1870098.

59. Yuan, X., Schuchard, R.J., and Crooks, A.T. (2019). Examining Emergent Communities and Social Bots Within the Polarized Online Vaccination Debate in Twitter. Social Media + Society *5*. https://doi.org/10.1177/2056305119865465.

60. Albadi, N., Kurdi, M., and Mishra, S. (2019). Hateful People or Hateful Bots? Detection and Characterization of Bots Spreading Religious Hatred in Arabic Social Media. Proc. ACM Hum. Comput. Interact. *3*, 1–25. https://doi.org/10.1145/3359163.

61. Abascal, M., Makovi, K., and Sargsyan, A. (2021). Unequal treatment toward copartisans versus non-copartisans is reduced when partisanship can be falsified. PLoS One *16*, e0244651. https://doi.org/10.1371/journal.pone.0244651.

62. Kats, D., and Sharif, M. (2022). "I Have No Idea What a Social Bot Is": On Users' Perceptions of Social Bots and Ability to Detect Them. In HAI '22: Proceedings of the 10th International Conference on Human-Agent Interaction, pp. 32–40. https://doi.org/10.1145/3527188.3561928.

63. Schweitzer, S., Dobson, K.S.H., and Waytz, A. (2024). Political Bot Bias in the Perception of Online Discourse. Soc. Psychol. Personal. Sci. *15*, 234–244. https://doi.org/10.1177/19485506231156020.

64. Hendriks, A. (2012). SoPHIE–Software Platform for Human Interaction (University of Osnabrueck).

65. Litman, L., Robinson, J., and Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. Behav. Res. Methods *49*, 433–442. https://doi.org/10.3758/s13428-016-0727-z.

66. Hauser, D.J., Moss, A.J., Rosenzweig, C., Jaffe, S.N., Robinson, J., and Litman, L. (2023). Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. Behav. Res. Methods *55*, 3953–3964. https://doi.org/10.3758/s13428-022-01999-x.

67. Chandler, J., Paolacci, G., and Hauser, D. (2020). Data quality issues on MTurk. In Conducting Online Research on Amazon Mechanical Turk and Beyond, L. Litman and J. Robinson, eds. (Sage Academic Publishing), pp. 95–120, chap. 5. https://doi.org/10.4135/9781506391151.n9.

68. Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P.D., Jewell, R., and Winter, N.J.G. (2020). The shape of and solutions to the MTurk quality crisis. Political Sci. Res. Methods *8*, 614–629. https://doi.org/10.1017/psrm.2020.6.

69. Ortmann, A., and Hertwig, R. (2002). The Costs of Deception: Evidence from Psychology. Exp. econ. *5*, 111–131. https://doi.org/10.1023/A:1020365204768.

70. Charness, G., Samek, A., and van de Ven, J. (2022). What is considered deception in experimental economics? Exp. econ. *25*, 385–412. https://doi.org/10.1007/s10683-021-09726-7.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Analyzed data	OSF Registry	https://osf.io/2wjrv
Software and Algorithms		
SoPHIELABS	SophieLabs	https://www.sophielabs.com
R v3.6.3	R Core Team	https://www.r-project.org

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The experiment follows a between-person design, and participants were recruited for different conditions on different days (i.e., we treat this as random assignment in terms of what actions were available to participants as well as their group composition). Participants were blinded to the complete set of experimental conditions. Please refer to the Figures S1–S19 to see the user interface used in the experiments. The experiment has been approved by the NYU Abu Dhabi IRB (#HRPP-2019-93). All participants entered the experiment after providing informed consent online.

Participants whose data we analyze (i.e., passed comprehension checks, were in complete groups, and did not drop out) across all conditions have the following demographic characteristics: 51% identified as female (all other participants identified as male: 48% or other 1%), and 82% identified as White (all other participants have identified as American Indian or Alaska Native, Asian or Asian American, Black or African American, Hispanic or Latino/a, Middle Eastern or North African, Other, or identified with multiple of these categories), with an average age of 38.49 ($sd = 11.73$). Our sample is described based on these demographic characteristics, as well as education, region and income in Table S13 that shows no meaningful differences across experimental conditions. The sample size by experimental condition is the following: 684 respondents in **Baseline**, i.e., 171 groups; 328 in **Feedback**, i.e., 82 groups; 324 in **Punishment**, i.e., 81 groups; 328 in **Reward**, i.e., 82 groups' and 324 in **Both**, i.e., 81 groups.

All experiments are programmed in SOPHIE.[64] We recruit participants on Amazon Mechanical Turk using the services of CloudResearch (previously TurkPrime,[65]). Only MTurk workers 18 years or older, located in the United States—as specified on their MTurk account and by their IP address—could see the "Human Intelligence Task" (HIT). To be eligible, workers also needed to have at least 100 HITs approved and a 95% approval rating. We also excluded workers from suspicious geolocations and those on the "universal exclude list," both managed by CloudResearch. In addition to these filters, we recruit from CloudResearch Approved participants to enhance data quality when these filters became available,[66] as these individuals have exhibited high levels of engagement and attention in prior tasks managed by CloudResearch.[67,68] In the description of the HIT prospective participants were asked to complete the study using a computer as the experiment was not optimized for phones and tablets. We organize all HITs in a survey group to prevent the same individuals to take part multiple times as well as maintain an external list of MTurk IDs of those who have already participated to screen out repeat participation.

Data collection for the **Baseline** condition took place between the 5th of December, 2021 and the 21st of February, 2022; and for all other conditions took place between the 22nd of May, 2022 and the 23rd of November, 2022. We aimed to collect data from 40 groups in each experimental condition determined by the information available about contributions and the potential actions group-members might take and the identity of group members signaled where all group members completed 20 round in the IPGG. To honor the informed consent obtained online, all groups where one or more participants dropped out finished all 20 interactions, and participants who finished the study were compensated as described in the consent form, while dropouts were replaced by actual bots. Groups with dropouts are not analyzed. As previously described, participants who failed to answer all comprehension check questions correctly on two attempts have not been analyzed, nor data from those who have not had three other group members to start the IPGG. When reporting the sample size these individuals are not counted as they are excluded form analyses. The demographic background of the individuals who were removed as a result of these steps are shown in Tables S8–S12, and contrasted to the same information of those who finished the study.

One potential concern could be that some groups do not finish the study due to participants who get punished dropping out. To examine this possibility, we looked to see how many participants' behavior follows this pattern. In the **Punishment** condition, only 4, or 2%, of game dropouts meet this criteria, in the **Both** condition, these statistics are 9 and 3%, respectively, highlighting the rarity of punishment. In the middle of the fielding of the **Punishment** condition we implemented a screen where dropouts could specify the reason why they dropped out, and this screen was active for all respondents in the **Both** condition. Out of the 13 respondents who dropped out after being punished two were "away from keyboard," two experienced connection problems, two claimed they were "kicked out" of the game, one had trouble with the submit button, one stated they were bored, and only one referenced "compensation" as their reason for dropping out (the only potential respondent who may have dropped out as a result of being punished). The

rest of the respondents (four) did not answer this question. Based on this, we believe that selective dropout as a result of punishment is unlikely; therefore, we do not believe that the reported results are biased due to this dynamic.

Participants have been compensated by a participation fee of $1.50 in the **Baseline** condition, and by $2.00 in all other conditions, which was later increased to $2.50 to compensate for the prolonged participation time. In addition, participants also earned a bonus from the IPGG and as a result of punishment and reward decisions where applicable. Average total bonuses have been $3.20 in the **Baseline** condition, $2.95 in the **Feedback** condition, $4.21 in the **Reward** condition, $3.05 in the **Punishment** condition, $3.83 in the **Both** condition, respectively. As previously specified, those whose groups have not been composed received their show-up payment, and those who failed to answer comprehension check questions correctly were extended compensation HITs for $0.10.

In all conditions, participants play 20 rounds of a standard IPGG. In the **Baseline** condition, at the end of each round, they learn about the amount of public good that was provided, but they do not know what each player provided, and they cannot take any action before the next round begins. In the **Feedback** condition, participants do learn what each player contributed at the end of each round, but cannot take any action before the next round begins. In the **Reward** condition, participants get the same information as in the **Feedback** condition, and then have the option of rewarding at a cost other players before the next round begins. This decision is made for each other player individually, which means that participants can give out zero to three rewards. The **Punishment** condition is similar, except that the decision is not to reward, but to punish at a cost. Finally, in the **Both** condition, participants can choose either to reward or to punish every other player.

METHOD DETAILS

Our design and parameter choices follow those of Rand and colleagues,[29] where the group size is set to 4, the multiplication factor is 1.6, the cost of rewarding or punishing a member is set to 4MUs, and the impact on the rewarded or punished member is + 12 MUs, and −12 MUs, respectively. The main difference between our work and that of Rand and colleagues is the identity of group members participants are informed about, as we add machines in the group.

To mathematically formulate the game, let N be the number of players in a group, E be the endowment each player receives in each round, c_i the contribution of player i to the public pool ($0 \leq c_i \leq E$), γ is the multiplication factor of the public good ($\gamma = 1.6$), r_i is the payoff of player i. The total contribution C by all players to the public pool is: $C = \sum_{i=1}^{N} c_i$. Each player receives an equal share of the public pool value ($V = \gamma \times C$), regardless of their contribution. Therefore, $r_i = E - c_i + \frac{V}{N}$ in each round.

In the conditions when punishments are allowed, each player will pay a cost δ where $\delta_i = \sum_{j \neq i} 4 \times D_{ij}$ for punishing other players in the group: D_{ij} is player i's decision on whether to punish player j, $D_{ij} = 1$ when player i punished player j and 0 otherwise. Furthermore, player i will pay a cost for the total number of punishments received $P_i = \sum_{j \neq i} -12 \times D_{ji}$.

In the conditions when rewards are allowed, each player will pay a cost δ^* where $\delta_i^* = \sum_{j \neq i} 4 \times D_{ij}^*$ for rewarding other players in the group: D_{ij}^* is player i's decision on whether to reward player j, $D_{ij}^* = 1$ when player i rewarded player j and 0 otherwise. Furthermore, player i will receive a benefit for the total number of rewards received $R_i = \sum_{j \neq i} 12 \times D_{ji}^*$.

In the games with punishment, reward or both, the player i's individual payoff: $r_i = E - c_i + \frac{V}{N} - \delta_i - \delta_i^* + P_i + R_i$ per round. Note that in the condition where both punishment and reward were available, a player can not be punished and reward by another player at the same time.

Prior to participating in the IPGG, participants are informed that: (i) the activity would take a maximum of 40 rounds—in fact everyone played only 20 rounds to avoid end of game effects; (ii) each round lasts up to 60 s; (iii) the composition of the group would not change (i.e., the same players keep interacting across rounds). Participants are also familiarized with how they might recognize machines and people in their group using images. Machines are introduced as "computer programs that make decisions." In the **Baseline** condition, all participants start with no personal funds, while in all other conditions they receive 50 monetary units (MUs), to enable all participants to punish and/or reward if these actions are available in their conditions. Before each round begins, all participants receive 20MUs to decide over, and 20MUs = $0.10. After reading the instructions, participants see three visual examples with calculations that show a situation where (i) all group-members contributed all their funds; (ii) no group-members contributed anything; (iii) half the group members contributed everything while half contributed nothing. Following these, participants are asked four comprehension check questions, which they have to answer all correctly on two attempts, otherwise, they may not take part in the study. Where applicable, participants are then introduced to the second stage of decision making. They are told that rewards and punishments are costly: they cost 4MUs, but benefit the rewarded by 12MUs, or harm the punished by 12MUs. Participants are told that no personal funds can go below zero (i.e., no participant would loose their participation fee). After instructions for the second stage of decision making, participants see an example where a group member rewards/punishes two other group members and are rewarded/punished by two group members. In the **Both** condition the example shows a group member punishing one and rewarding one, and experiencing these same actions as well. Similarly to the contribution stage, four comprehension check questions follow, and only participants who answered them all correctly on a maximum of two attempts may proceed. Future work should explore how the results would change by changing the incentive structure for participants, such as modifying the cost of punishment and reward and its impact on the budgets of those punished and rewarded, as well as how these relate to the money one could gain in the first stage of the game through public good provision.

When participants completed all these steps they are placed in a waiting room where they are told that a beep will alert them when the activity starts. After 5 min of waiting, participants are offered their show-up fee if they would like to leave. After 10 min, participants

are automatically given a completion code. While in the waiting room, participants see an image of the one or two decision making screens. The screens are annotated to draw their attention to the progression of a round; i.e., how many group members have already made a decision and which stage they are in, the slider or radio buttons to make contributions and reward/punishment decisions where applicable; the amount they have in their personal account which accumulates after every round of decision making; a depiction of their group that signals group composition; a timer that measures the seconds in a decision making round; and a jar that represents the common pot, and visually fills after each round of contribution. If participants are in any other condition than the **Baseline**, contributions in the prior round are also highlighted. Lastly, they all see a summary screen to convey how much they earned after both rounds (where applicable). The activity is followed by questions related to it, and concludes with information about participants' demographics and a screen that summarizes all earnings.

Participants' identity is signaled via images, which is constantly reinforced throughout the experiment. This signal is meant to convey who is in a participant's group, while all participants were, in fact, humans. This inevitably entails deception. In some disciplines, like Psychology and Sociology this form of deception, coupled with a debriefing of subjects after the experimental game is standard practice, while in others, such as economics, it is seldom used and is against disciplinary norms.[69] In the present case, it was essential to create a realistic representation of human behavior in a human-machine group where a person believes to be playing with machines, while the machines' behavior is human-like, preserving the variation that exists in these settings. For this reason, using data from the human-only groups does not meet the requirements of the human-machine scenario, as we show that people behave in these settings differently. In other words, this research would not be possible without taking this methodological approach, and many economists have also recently been of the opinion that deception should be allowed in cases like these.[70] Of course, deception is not without ethical considerations, such as causing potential embarrassment upon debriefing subjects who find out that they played with other people instead of machines. In addition, the common use of deception on the platforms where many researchers collect data may lead to weakening treatment effects where research subjects expect to be deceived to begin with, and assume to play with machines, rather than people generally when interacting with purported humans.

QUANTIFICATION AND STATISTICAL ANALYSIS

All analyses were conducted using the R programming language (v3.6.3) and the lme4 library. No formal sample size estimation was conducted; the sample size ($N = 1988$) was determined based on prior studies and logistical feasibility. Means and 95% confidence intervals are reported throughout. Statistical significance was defined as $p<0.05$, two-tailed.

The magnitude of the machine penalty

In Figure 1 we graph mean contributions in the IPGG by experimental condition over the 20 rounds of decision making. The ribbons show the 95% confidence interval of the contribution. Using random-effects regression models for participants and groups, and fixed effects for rounds, we estimate the machine penalty, that is, the gap in contributions between groups where partners are assumed to be humans, and groups in which partners are assumed to be machines. These estimates are reported in Figure 1 and are labeled "Gap Size."

Frequency of rewards and punishments

We calculate the frequency of rewards and punishments by considering all the 240 opportunities to hand out a reward or punishment for each group of 4 participants playing for 20 rounds in the conditions where this is applicable. We use a simple percentage of these possible events when punishments were meted out or rewards were given. We also fitted a multilevel model in which the binary outcome was to hand out a reward, and the predictors were the purported composition of the group, the experimental condition, and their interaction; with random intercepts for groups and participants, and a fixed effect of round. The same approach could not be applied to model punishments due to the extreme rarity of punishment, our multilevel model did not converge.

Reaction to rewards

Figure 2 shows how participants react to the number of rewards they receive—specifically, how they adjust their contribution on average in the next round, as well as the 95% confidence interval for this adjustment. We conducted a multilevel model to predict the change in contribution in the next round, with our usual set of predictors: experimental condition, group composition, their interaction, random intercept for groups and participants, and fixed effect of round.