Predicting the Machine: Intentionality Framing Reduces the Prediction Gap in Human–AI Cooperation

Laura K. Globig (laura.globig@gmail.com)[1], Nadya Hanaveriesa[1], Sydney Levine[1], & Jay J. Van Bavel[1,2]
[1]Department of Psychology & Center for Neural Science, New York University, New York, NY, USA
[2]Norwegian School of Economics, Bergen, Norway

Abstract

Cooperation is sustained by shared norms that regulate expectations about how others will behave. As artificial intelligence (AI) systems increasingly participate in social and economic interactions, a critical question emerges: do the normative expectations that sustain human cooperation extend to artificial players? We propose that failures in human-AI cooperation may reflect difficulty forming accurate predictions about artificial behavior, disrupting norm-based coordination. In Study 1 (N = 794), participants in a repeated Public Goods Game were less sensitive to prosocial norms with AI than human players, accompanied by a systematic prediction gap. In Study 2 (N = 314), framing AI as intentional rather than mechanistic substantially reduced prediction errors and cooperation differences. Trial-by-trial analysis showed that intentionality framing affected initial expectation calibration but not feedback-driven updating, consistent with a shift in participants' initial mental model of the AI. Norm-based human-AI cooperation depends on predictability.

Keywords: artificial intelligence; public goods game; cooperation; social norms; predictive processing

Overview

Cooperation in human societies is sustained not only by individual incentives but by shared norms that regulate expectations about how others will behave (Fehr & Schurtenberger, 2018). Norms stabilize cooperation by aligning predictions with behavior, allowing individuals to coordinate even when defection is tempting (Cialdini et al., 1991). If cooperation depends on calibrated expectations, then interactions with artificial intelligence (AI) may falter not because people lack trust or goodwill, but because they struggle to predict artificial behavior. We test this possibility directly.

Existing work on human-AI cooperation has largely framed behavioral differences in terms of trust, preferences, or strategic exploitation (Karpus et al., 2021; Melo et al., 2016). However, if cooperation with AI is shaped by norms in addition to trust and self-interest, the critical issue becomes whether the expectation-based processes through which norms exert their influence operate when people interact with AI. Do the normative mechanisms that ordinarily structure human cooperation operate effectively when the partner is a machine?

In human-human interaction, social norms guide behavior by shaping expectations about how others are likely to act and about the social consequences of deviating from those expectations (Cialdini et al., 1991). These expectations are typically supported by an intentional construal of social partners: people predict others by attributing goals, beliefs, and stable mental states that render behavior intelligible and forecastable (Heider, 1958; Kelley, 1973). Norms are therefore not simple prescriptions, but predictive structures: they allow individuals to infer what others will do and align their own behavior accordingly. Crucially, this process depends on predictability. Norms exert their strongest influence when people can reliably map observed behavior onto stable expectations about future actions (Hackel et al., 2020). When this mapping is noisy, norm information may fail to translate into cooperative behavior.

One reason norms may exert weaker influence in interactions with AI is reduced accuracy in expectation formation. Norms guide action by stabilizing beliefs about others' likely behavior; miscalibrated expectations can therefore blunt the behavioral impact of norm information even when the norm itself is correctly perceived (Pärnamets et al., 2020). While prior work on trust in AI has examined generalized expectations about reliability or benevolence, expectation accuracy has rarely been quantified as trial-by-trial prediction error (Ishowo-Oloko et al., 2019; Köbis et al., 2021).

Expectation formation may depend on whether an agent is construed as intentional rather than mechanistic. Dennett (1987) argued that humans routinely adopt an *intentional stance* toward other agents, predicting behavior by attributing beliefs, desires, and goals rather than modeling underlying mechanisms. While the intentional stance arises readily for human partners, it is applied less consistently to artificial agents, who are often predicted using rule-based or machine-like models (Maninger & Shank, 2022; Marchesi et al., 2019; Wykowska, 2021). If social prediction is gated by this construal, framing AI as intentional should reduce discrepancies in expectation accuracy between AI and human partners.

We test whether human and AI norms are translated into cooperative behavior differently. We hypothesize that in a public goods game (PGG), (1) people will show weaker norm-dependent cooperation with AI compared to human players, and (2) this attenuation will be accompanied by a deficit in predicting AI behavior.

Mechanistically, these failures could operate at different stages. People might struggle to form accurate expectations about an agent (a calibration deficit), or they might fail to incorporate feedback once an error is observed (a learning deficit). These pathways yield the same observable behavior while reflecting distinct mechanisms (Rescorla & Wagner, 1972; Sutton & Barto, 2018). Distinguishing between them reveals whether an *AI cooperation gap* stems from a noisy

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

internal model of the player or resistance to learning from artificial agents.

Across two pre-registered experiments, we investigated how social norms shape cooperation when people interact with artificial versus human players in a repeated PGG. In Study 1 (N = 794), we examine whether humans apply social norms differently to AI players, and whether any cooperative deficit is driven by an inability to predict AI behavior. In Study 2 (N = 314), we tested a potential underlying mechanism by experimentally manipulating intentionality framing of the AI: whether ascribing goals to the AI reduces prediction error and restores cooperation.

Study 1: Weaker Norm Sensitivity with AI

Participants completed a PGG (Figure 1). On each trial, participants played with three other players and decided how much of an $8 endowment to contribute to a shared pool. Participants first indicated how much they thought the other players would contribute, then chose their own contribution. The total contribution was doubled and evenly redistributed. At the end of each trial, participants saw each player's actual contribution and earnings.

We varied player type between subjects: participants were told the other players were either AI delegates or human delegates playing on behalf of other participants. In reality, all delegate decisions were pre-programmed. We also varied social norms between subjects by manipulating the contribution behavior of other players: groups contributed either generously (prosocial norm) or minimally (antisocial norm).

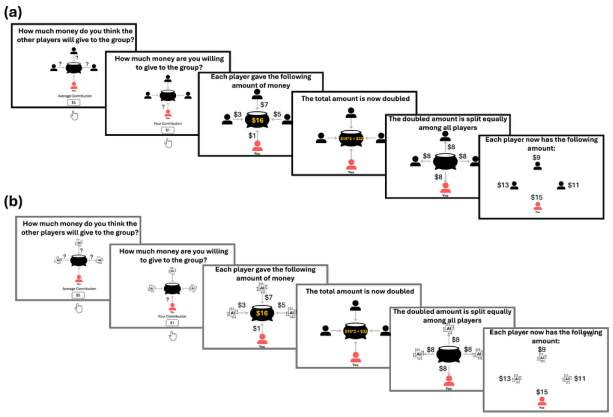

Figure 1. Example trials of the Public Goods Game with **(a)** human delegates or **(b)** AI delegates. On each of 20 trials with new players, participants reported their expectation of others' contributions, then chose their own. The shared pool was doubled and redistributed.

Norm Sensitivity Differs by Player Type

We first examined cooperation as a function of player and norm. A 2 (Player: AI vs. Human) × 2 (Norm: Prosocial vs. Antisocial) between-subjects ANOVA on mean contribution revealed a significant main effect of Norm ($F(1, 790) = 83.23$, $p < .001$, $\eta^2_p = .10$): participants contributed more under prosocial norms ($M = 4.12$, $SE = 0.09$) than antisocial norms ($M = 2.99$, $SE = 0.09$; Figure 2). There was no main effect of Player ($F(1, 790) = 1.13$, $p = .29$).

Critically, we observed a significant Player × Norm interaction ($F(1, 790) = 4.52$, $p = .034$, $\eta^2_p = .006$). Post-hoc comparisons revealed that under prosocial norms, participants contributed significantly more with human players ($M = 4.31$, $SE = 0.13$) than AI players ($M = 3.92$, $SE = 0.12$; $p = .027$). Under antisocial norms, contributions did not differ (AI: $M = 3.06$; human: $M = 2.93$; $p = .44$). The effect of norms on cooperation was weaker for AI than human players.

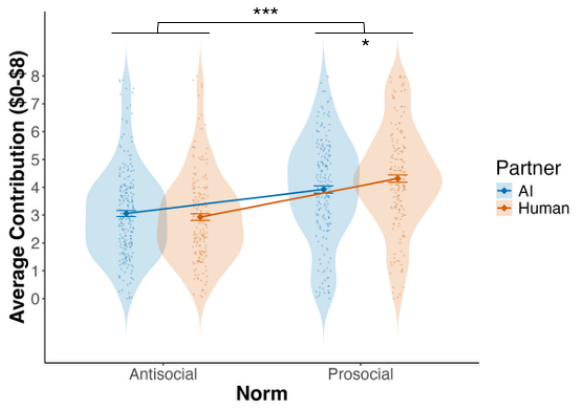

Figure 2. Prosocial norms elicited stronger cooperation with human than AI players. The cooperative boost from prosocial norms was smaller for AI (blue) than human (orange) players. Diamonds: means; bars: ±1 SEM. *** $p < .001$; * $p < .05$.

Participants Predicted AI Worse than Humans

We next examined whether reduced norm sensitivity was accompanied by differences in predicting player behavior. We computed trial-by-trial absolute prediction error as the difference between predicted and actual mean contributions, fitting a linear mixed-effects model with Player and Norm as fixed effects and random participant intercepts.

This revealed a significant main effect of Player ($b = -0.23$, $SE = 0.07$, $t(789) = -3.12$, $p = .002$): participants exhibited larger prediction errors with AI ($M = 1.74$, $SE = 0.04$) than human players ($M = 1.56$, $SE = 0.04$; Figure 3a). Neither the main effect of Norm nor the interaction was significant.

Directional (signed) prediction error analysis showed a main effect of Player ($b = -0.24$, $p = .048$): participants overestimated AI contributions ($M = 0.11$) relative to humans ($M = -0.01$). A strong main effect of Norm ($b = -1.51$, $p < .001$) reflected systematic under- and overestimation across norm contexts: participants overestimated in antisocial environments and underestimated in prosocial environments (Figure 3b). Across norms, expectations about AI players were biased toward greater cooperation than was actually observed.

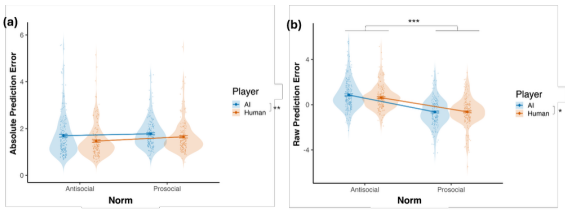

Figure 3. Participants were worse at predicting AI than human behavior. **(a)** Larger absolute prediction errors for AI than human players across norm contexts. **(b)** Participants overestimated AI relative to human contributions; predictions were biased toward overestimation under antisocial norms and underestimation under prosocial norms. ** $p < .01$; * $p < .05$.

Trial-by-Trial Updating Did Not Differ by Player

To test whether prediction errors shaped decisions, we examined trial-by-trial updating. Using a Rescorla-Wagner framework, we modeled changes in expected contributions as a function of prior-trial prediction errors. We observed robust error-driven belief updating ($b = 0.20$, $SE = 0.01$, $t(1814) = 14.38$, $p < .001$), and crucially, this updating did not differ between AI and human players (Player × Error interaction: $p = .32$).

Modeling absolute change in contribution as a function of absolute prediction error revealed that larger errors predicted larger subsequent adjustments ($b = 0.10$, $SE = 0.02$, $p < .001$), again with no Player × Error interaction ($p = .46$). Together, these analyses indicate that prediction errors drove both belief and behavioral updating, and that this error-driven learning operated equivalently for AI and human players. The cooperation differences we observe were accompanied by a calibration deficit, raising the possibility that expectation accuracy contributes to norm-driven cooperation.

Study 2: Intentionality Framing Closes the Gap

Study 1 reveals a prediction gap accompanying weaker norm sensitivity for AI. If this gap arises because people fail to adopt the intentional stance toward artificial agents, framing AI as goal-directed should improve expectation accuracy and restore cooperation. We tested this with a new sample. The task was identical to Study 1, except that player type and norm varied within subjects, allowing a more sensitive within-person test of differential prediction accuracy.

Intentionality framing varied between subjects. Participants were randomly assigned to interact with AI agents described either as goal-directed and capable of planning (high intentionality) or as executing stochastic, mechanistic rules (low intentionality), providing a causal test of the role of perceived intentionality.

Manipulation Check

Participants in the high intentionality condition attributed greater intentionality to AI. Scores on the IDAQ capable-of-intentions subscale (Waytz et al., 2014) were higher in the high intentionality condition ($M = 45.9$, $SEM = 2.69$) than the low intentionality condition ($M = 26.6$, $SEM = 2.57$, $t(313) = 5.17$, $p < .001$, $d = 0.58$; Figure 4). The manipulation successfully increased intentionality attribution to the AI.

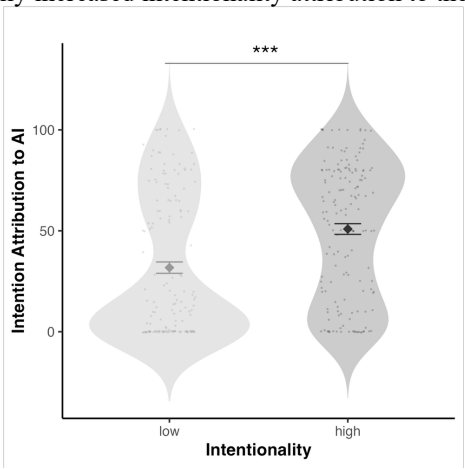

Figure 4. Intentionality manipulation was successful. Participants in the high intentionality condition assigned greater intent to the AI than the low intentionality condition. *** $p < .001$.

Intentionality Reduced Cooperation Gap

We conducted a 2 (Player: AI vs. Human) × 2 (Norm: Prosocial vs. Antisocial) × 2 (Intentionality: Low vs. High) mixed ANOVA with Player and Norm as within-subject factors. Replicating Study 1, we observed a robust main effect of Norm ($F(1, 313) = 237.88$, $p < .001$, $\eta^2_p = .43$; Figure 5). There was also a main effect of Player ($F(1, 313) = 17.51$, $p < .001$, $\eta^2_p = .05$): participants contributed more with humans than AI. There was no main effect of Intentionality ($p = .746$).

Crucially, we observed a significant Player × Intentionality interaction ($F(1, 313) = 8.86$, $p = .003$, $\eta^2_p = .03$), indicating cooperation differences between AI and human players depended on intentionality framing. Planned comparisons revealed that under low intentionality, participants contributed less with AI than human players averaged across norms (AI: $M = 2.84$; human: $M = 3.31$; $\Delta M = -0.47$, $t(313) = -5.02$, $p < .001$). Under high intentionality, contributions did not differ (AI: $M = 3.09$; human: $M = 3.17$; $\Delta M = -0.08$, $p = .39$). Framing AI as intentional attenuated cooperation differences while preserving sensitivity to social norms.

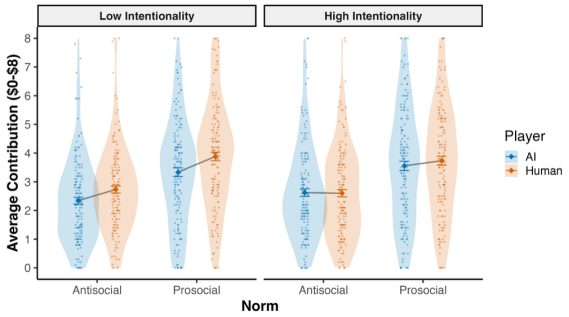

Figure 5. Intentionality framing attenuates cooperation differences between AI and human players. Under low intentionality, participants contributed less with AI across both norms; under high intentionality, contributions did not differ. AI: blue; human: orange. Diamonds: means; bars: ±1 SEM.

Intentionality Reduced the Prediction Gap

Did intentionality alter the gap in prediction accuracy? Absolute prediction error was analyzed using a linear mixed-effects model with Player, Norm, and Intentionality as fixed effects and random intercepts plus random slopes for Player and Norm at the participant level.

This revealed a significant Player × Norm × Intentionality interaction (b = 0.17, SE = 0.08, $t(11485)$ = 2.03, p = .04). Under low intentionality, in prosocial environments participants showed larger errors with AI (M = 1.85) than humans (M = 1.72; $t(417)$ = 3.04, p = .002); no AI-human difference in antisocial environments. Under high intentionality, prediction errors did not differ between AI and human players in either antisocial (p = .058) or prosocial (p = .28) environments (Figure 6a).

Directional prediction error analysis revealed a Player × Intentionality interaction (b = −0.24, p = .022). Under low intentionality, participants underestimated AI more than humans in prosocial environments (AI: M = −1.14; human: M = −0.95; p = .009). Under high intentionality, this directional bias was no longer reliable in either norm context (Figure 6b). Ascribing intentionality to AI selectively attenuates the prediction error gap.

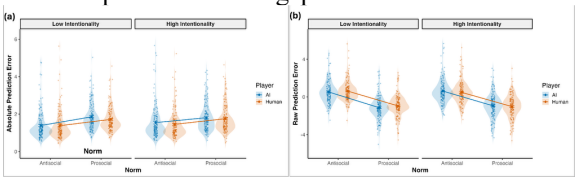

Figure 6. Reduced prediction error gap when AI is framed as intentional. **(a)** Larger absolute prediction errors for AI than human players in prosocial environments under low intentionality; gap attenuated under high intentionality. **(b)** Directional bias of underestimating AI more than humans in prosocial environments under low intentionality; not reliable under high intentionality.

Calibration, Not Learning

We sought to isolate the cognitive locus of the intentionality effect. Reduced prediction error could arise from improved initial calibration (better expectations before feedback) or improved feedback integration (better learning after errors). We modeled trial-by-trial contribution updating as a function of prior-trial prediction error.

Prediction errors robustly drove behavioral updating: larger errors predicted larger subsequent corrections (b = −0.23, SE = 0.02, p < .001). Crucially, learning slopes did not differ by Player or Intentionality (all ps > .11). Once a discrepancy was detected, participants learned from AI just as efficiently as humans. The deficit was not in incorporating feedback but in where they started.

Decomposing raw prediction error into systematic bias (mean directional error) and stochastic variability (noise), bias analyses revealed a Player × Intentionality interaction (b = −0.24, p = .040): intentionality framing selectively reduced the systematic tendency to misestimate AI. Stochastic variability did not differ between conditions (ps > .19). Attributing intentionality to AI restores cooperation by correcting systematic calibration of social expectations, leaving error-driven learning intact.

Discussion

Across two pre-registered studies, we demonstrate that interactions with artificial agents are characterized by a selective attenuation of norm-driven cooperation. Rather than reflecting a generalized *algorithm aversion* (Dietvorst et al., 2015) or global withdrawal of social preferences (Karpus et al., 2021; Melo et al., 2016), our results identify a calibration deficit that co-occurs with attenuated norm sensitivity: participants struggled to form stable, accurate expectations about AI behavior, and this prediction gap tracked the reduction in norm-driven cooperation.

In Study 1, while participants were sensitive to the normative environment, the cooperation boost from prosocial norms was significantly smaller for AI than human players, consistent with evidence that artificial agents are treated as a distinct social category that elicits reduced fairness (Köbis et al., 2021; Weiß et al., 2020). This divergence was accompanied by larger prediction errors. Without a precise expectation of what an AI *will* do, the behavioral force of social norms is attenuated.

Study 2 provides the mechanistic key: the attribution of intentionality. Adopting the *intentional stance* (Dennett, 1987) renders behavior predictable by treating actions as governed by stable internal states (Devaine et al., 2014). Framing AI agents as possessing goals and mental states increases trust, reliance, and perceived predictability (Waytz et al., 2014). Our intentionality manipulation closed the prediction error gap and restored cooperation to human-human levels. Critically, this effect was restricted to expectation formation. Once an error occurred, participants updated their behavior with equal efficiency across conditions. Humans do not have a fundamental learning deficit when interacting with AI; they lack the predictive

scaffold (the intentional stance) that ordinarily renders behavior legible.

Multiple lines of evidence converge on expectation calibration as the mechanism underlying the cooperation gap. Prediction errors were larger for AI than human players across both studies. Learning rates did not differ by player type, ruling out differential feedback integration. Intentionality framing simultaneously reduced both prediction error and cooperation differences, consistent with a shared underlying process. Decomposition of prediction error revealed that intentionality selectively affected systematic bias rather than stochastic noise, directly implicating calibration.

These findings carry practical implications for designing collaborative human-AI systems. Fostering cooperation may not require increasing the warmth or human-likeness of AI (Sampaio et al., 2023) but rather increasing behavioral legibility. Interventions that support calibrated expectations, whether through intentionality framing or transparency, may activate the normative mechanisms that drive collective action. While effects were modest in absolute magnitude, small biases compound across repeated interactions (Matz et al., 2017; Peysakhovich & Rand, 2016), and population-scale deployments could amplify them. In practice, this could mean designing AI systems that explicitly communicate goals and decision-relevant states, provide transparent behavioral histories that users can interpret through an intentional lens, or adopt interaction patterns that invite mentalizing.

Several boundary conditions warrant attention. Our manipulation varied observed contribution behavior, operationalizing descriptive norms; this does not capture injunctive norms or sanctioning dynamics. While intentionality attributions stabilize expectations, they may also lead to over-trust or misattribution of moral agency to systems that remain algorithmic (Gray et al., 2012; Maninger & Shank, 2022). Our studies used text-based framing in a stylized economic game; whether these effects generalize to embodied AI agents or richer interaction contexts remains open.

Conclusion

Human cooperation with artificial agents is characterized by a selective attenuation of norm-driven behavior, rooted not in a generalized failure of social preference but in a specific calibration bottleneck of expectation formation. Participants remain highly sensitive to social context, yet the mapping from normative signals to cooperative action is diluted by the inherent noise of predicting artificial behavior. By identifying the attribution of intentionality as the mechanistic lever that stabilizes these expectations without altering how participants update behavior, we suggest that AI can be effectively integrated into collective action environments when it is rendered behaviorally legible and triggers intuitive norm-following heuristics. Pinpointing the locus of divergence in expectation formation provides a foundation for designing agents that sustain human cooperative norms in mixed-agent ecosystems.

Methods

We report how we determined sample sizes, all data exclusions, all manipulations, and all measures. The research was approved by the New York University IRB (FY2025-9956). Anonymized data and analysis code are available on the Open Science Framework: https://osf.io/qgz3a. Pre-registrations for Study 1 and Study 2 are available at https://aspredicted.org/xr2pp8.pdf and https://aspredicted.org/t2uy3p.pdf.

Study 1

Participants. Sample size was determined by a priori power analysis (GPower 3.1) to achieve 80% power (β = 0.20, α = 0.05) to detect small-to-medium effects. We recruited 1,092 participants via Prolific. As pre-registered, we excluded participants giving identical responses on every trial (n = 93) or indicating in post-task checks that they did not believe their players were real (n = 205). The final sample was 794 (Mage = 44.16; female = 400, male = 385, other/prefer not to say = 9), all fluent English speakers in the US. Participants received approximately $12/hour plus a performance bonus up to $8.

Public Goods Game. Participants completed a repeated PGG adapted from Hackel et al. (2020). On each round, participants were paired with three new players and received an $8 endowment. Participants chose an integer contribution between $0 and $8 to a shared pool, which was doubled and evenly redistributed. Player type was manipulated between subjects (AI vs. human delegates); all player behavior was generated by pre-programmed algorithms. Social norms were manipulated between subjects: prosocial groups consistently contributed high amounts (~$6.5), antisocial groups contributed low amounts. Each round, participants reported their expectation of others' contributions, made their own contribution under a 10-second deadline, and received feedback on all players' contributions and earnings. One round was randomly selected for payment.

Procedure. Participants provided informed consent, received instructions, and completed comprehension checks (with up to three attempts). Following 20 PGG rounds, participants completed a survey assessing perceived realism, trust toward AI and humans, and perceived closeness via Inclusion of Other in the Self diagrams.

Analysis. Mean contributions were entered into a 2 (Player) × 2 (Norm) between-subjects ANOVA with post-hoc pairwise comparisons. Trial-by-trial absolute and directional prediction errors were analyzed using linear mixed-effects models with Player and Norm as fixed effects and random participant intercepts. To assess belief updating, we modeled changes in predicted contribution from trial t to t+1 as a function of signed prediction error at t, with prediction level as a covariate. Behavioral adjustment was modeled as the absolute change in contribution as a function of absolute prediction error, controlling for current contribution. All analyses were conducted in R.

Study 2

Participants. Sample size was determined by power analysis (GPower 3.1) for 80% power to detect small-to-medium effects. We recruited 396 participants via Prolific. As pre-registered, we excluded participants giving identical responses on every trial (n = 10) or who did not believe their players were real (n = 71). The final sample was 314 (Mage = 41.54; female = 159, male = 154, other = 2). Participants received ~$12/hour plus a bonus up to $2.

Public Goods Game. The PGG was identical to Study 1 with two differences. First, player type and social norm varied within participants across blocks (10 trials per block, 40 trials total) with counterbalanced order. Second, AI blocks were preceded by an intentionality framing manipulation that varied between participants, describing the AI as either low in intentionality (operating via fixed algorithms without goals) or high in intentionality (capable of forming goals and adapting behavior in socially meaningful ways). Framing did not alter payoff structure, contribution distributions, or feedback. All other aspects matched Study 1.

Analysis. Manipulation check: IDAQ capable-of-intentions and ToM intentions subscale scores were compared between conditions via independent-samples t tests. Mean contributions were entered into a 2 (Player) × 2 (Norm) × 2 (Intentionality) mixed ANOVA. Absolute and directional prediction errors were analyzed via linear mixed-effects models with Player, Norm, and Intentionality as fixed effects, random intercepts, and random slopes for Player and Norm. Trial-by-trial updating was modeled with prior directional prediction error as predictor and Player and Intentionality as moderators. Bias and noise were decomposed by computing, per participant within each Player × Norm cell, the mean directional error (bias) and the standard deviation of residual errors around that mean (noise); each was analyzed via linear mixed-effects regression with fixed effects for Player, Norm, and Intentionality.

Acknowledgments

This project was funded with support from the Templeton World Charity Foundation (S.L., J.J.V.B.; doi.org/10.54224/31570) and the National Science Foundation (#2334148, J.J.V.B.). L.K.G. is supported by the Roddenberry Foundation and the Dana Foundation. J.J.V.B. and L.K.G. are also supported by a grant from Google.org. We thank the Social Identity and Morality Lab for helpful comments. The authors used AI assistance solely for editorial polishing (spelling, grammar, and minor stylistic refinements); no AI tools were used to generate scientific content, analyses, or interpretations, and all content remains the authors' responsibility.

References

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in Experimental Social Psychology* (Vol. 24, pp. 201-234). Elsevier. https://doi.org/10.1016/S0065-2601(08)60330-5

Dennett, D. C. (1987). *The intentional stance*. The MIT Press.

Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: Does mentalizing make a difference when we learn? *PLOS Computational Biology*, *10*(12), e1003992. https://doi.org/10.1371/journal.pcbi.1003992

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114-126. https://doi.org/10.1037/xge0000033

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458-468. https://doi.org/10.1038/s41562-018-0385-5

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101-124. https://doi.org/10.1080/1047840X.2012.651387

Hackel, L. M., Wills, J. A., & Van Bavel, J. J. (2020). Shifting prosocial intuitions: Neurocognitive evidence for a value-based account of group-based cooperation. *Social Cognitive and Affective Neuroscience*, *15*(4), 371-381. https://doi.org/10.1093/scan/nsaa055

Heider, F. (1958). *The psychology of interpersonal relations*. John Wiley & Sons. https://doi.org/10.1037/10628-000

Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency-efficiency tradeoff in human-machine cooperation. *Nature Machine Intelligence*, *1*(11), 517-521. https://doi.org/10.1038/s42256-019-0113-5

Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., & Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience*, *24*(6), 102679. https://doi.org/10.1016/j.isci.2021.102679

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107-128. https://doi.org/10.1037/h0034225

Köbis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, *5*(6), 679-685. https://doi.org/10.1038/s41562-021-01128-2

Maninger, T., & Shank, D. B. (2022). Perceptions of violations by artificial and human actors across moral foundations. *Computers in Human Behavior Reports*, *5*, 100154. https://doi.org/10.1016/j.chbr.2021.100154

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00450

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, *114*(48), 12714-12719. https://doi.org/10.1073/pnas.1710966114

Melo, C. D., Marsella, S., & Gratch, J. (2016). People do not feel guilty about exploiting machines. *ACM Transactions*

on Computer-Human Interaction, *23*(2), 1-17. https://doi.org/10.1145/2890495

Pärnamets, P., Shuster, A., Reinero, D. A., & Van Bavel, J. J. (2020). A value-based framework for understanding cooperation. *Current Directions in Psychological Science*, *29*(3), 227-234. https://doi.org/10.1177/0963721420906200

Peysakhovich, A., & Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, *62*(3), 631-647. https://doi.org/10.1287/mnsc.2015.2168

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (Vol. 2, pp. 64-99). Appleton-Century-Crofts.

Sampaio, W. M., Freitas, A. L., Rêgo, G. G., Morello, L. Y. N., & Boggio, P. S. (2023). Effects of co-players' identity and reputation in the public goods game. *Scientific Reports*, *13*(1), 13520. https://doi.org/10.1038/s41598-023-40730-4

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). The MIT Press.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113-117. https://doi.org/10.1016/j.jesp.2014.01.005

Weiß, M., Rodrigues, J., Paelecke, M., & Hewig, J. (2020). We, them, and it: Dictator game offers depend on hierarchical social status, artificial intelligence, and social dominance. *Frontiers in Psychology*, *11*, 541756. https://doi.org/10.3389/fpsyg.2020.541756

Wykowska, A. (2021). Robots as mirrors of the human mind. *Current Directions in Psychological Science*, *30*(1), 34-40. https://doi.org/10.1177/0963721420978609