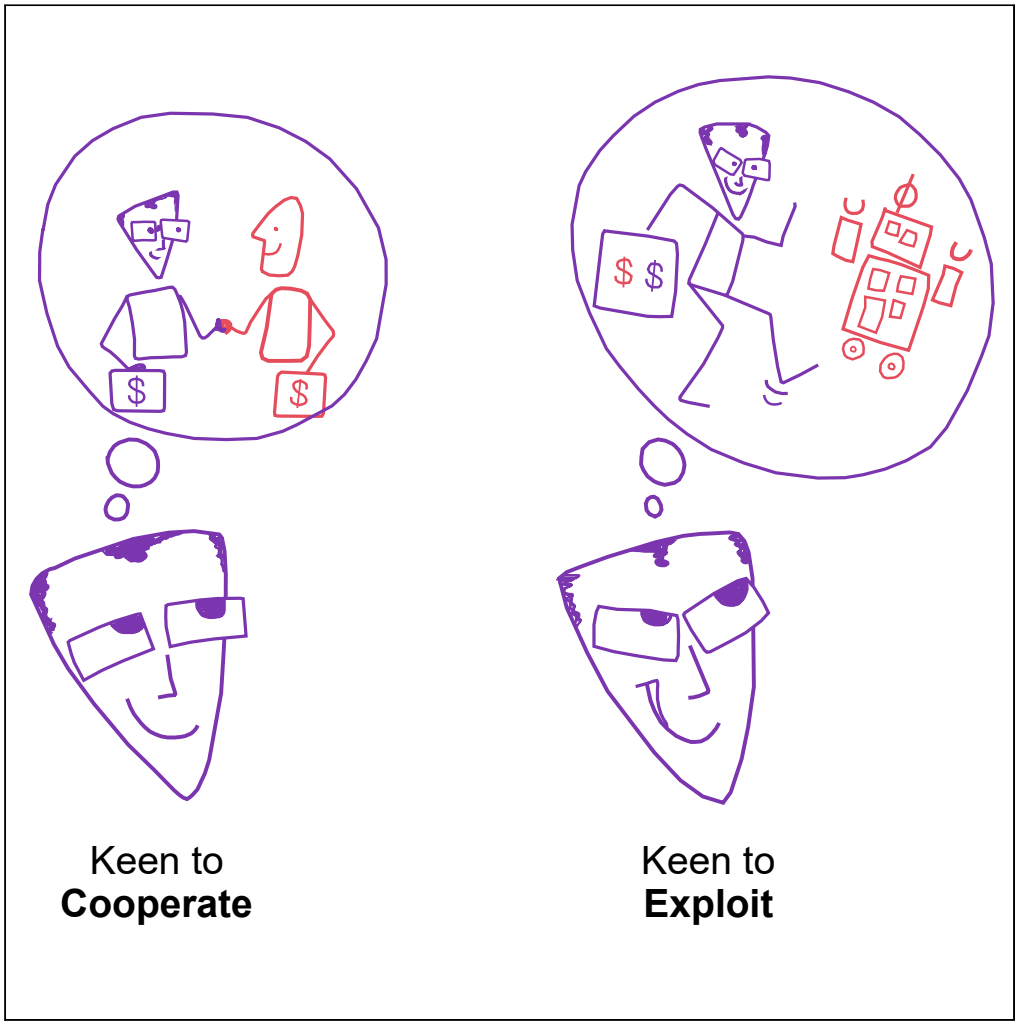
iScience

CellPress
OPEN ACCESS

Article

Algorithm exploitation: Humans are keen to exploit benevolent AI



Jurgis Karpus, Adrian Krüger, Julia Tovar Verba, Bahador Bahrami, Ophelia Deroy

ophelia.deroy@lrz.uni-muenchen.de

Highlights

People predict that AI agents will be as benevolent (cooperative) as humans

People cooperate less with benevolent AI agents than with benevolent humans

Reduced cooperation only occurs if it serves people's selfish interests

People feel guilty when they exploit humans but not when they exploit AI agents

Karpus et al., iScience 24, 102679
June 25, 2021
https://doi.org/10.1016/j.isci.2021.102679

iScience

CellPress
OPEN ACCESS

Article

Algorithm exploitation: Humans are keen to exploit benevolent AI

Jurgis Karpus,[1,3] Adrian Krüger,[1] Julia Tovar Verba,[1] Bahador Bahrami,[2,3,4,5] and Ophelia Deroy[1,5,6,7,*]

SUMMARY

We cooperate with other people despite the risk of being exploited or hurt. If future artificial intelligence (AI) systems are benevolent and cooperative toward us, what will we do in return? Here we show that our cooperative dispositions are weaker when we interact with AI. In nine experiments, humans interacted with either another human or an AI agent in four classic social dilemma economic games and a newly designed game of Reciprocity that we introduce here. Contrary to the hypothesis that people mistrust algorithms, participants trusted their AI partners to be as cooperative as humans. However, they did not return AI's benevolence as much and exploited the AI more than humans. These findings warn that future self-driving cars or co-working robots, whose success depends on humans' returning their cooperativeness, run the risk of being exploited. This vulnerability calls not just for smarter machines but also better human-centered policies.

INTRODUCTION

Imagine yourself stuck in traffic as you drive out of the city for the weekend. Ahead of you someone wants to join the traffic from a side road. Will you stop and let them in or push forward, hoping that someone else will let them in behind you? What would you do if this was a self-driving car with no passengers?

As artificial intelligence (AI) agents acquire capacities to decide autonomously, we will switch from being omnipotent users of intelligent machines (e.g., Google Translate) to having to make decisions with or beside them in social interactive settings (e.g., sharing the road with self-driving cars) (Bonnefon et al., 2016; Chater et al., 2018, 2019; Crandall et al., 2018; Rahwan et al., 2019). Unlike in chess, Go, and StarCraft II, in which AI has already outperformed humans (Campbell et al., 2002; Silver et al., 2016; Vinyals et al., 2019), most of our day-to-day social interactions are not zero-sum games where one player's win is the other one's loss. Instead, they offer opportunities to cooperate for the attainment of mutual gains (Colman, 1999). Cooperation often requires compromise and willingness to take risks: one may have to sacrifice some of their personal interests for the benefit of the group and expose themselves to the risk that others may not cooperate. As amply evidenced by behavioral game theory, people often choose to cooperate with others, even in anonymous one-shot encounters when acting selfishly bears no risk of damaging their reputation (Battalio et al., 2001; Camerer, 2003; Johnson and Mislin, 2011; McCabe et al., 2003; Rand et al., 2012; Rubinstein and Salant, 2016). The question we raise is whether people will continue to do so when they interact with AI.

Economic games are useful tools to empirically test people's cooperativeness. Recent work has shown that when groups of two or more human decision-makers face collective problems, the presence of a few bots can aid coordination and cooperation between humans in human-machine groups (Shirado and Christakis, 2017, 2020). This does not mean, however, that humans will be willing to cooperate in one-to-one interactions with artificial agents when no other humans are involved. The studies that used economic games to analyze social dilemmas in which decision-makers interact in pairs show that humans cooperate less with machines than with humans (Sandoval et al., 2016; Torta et al., 2013). This accords with findings from other behavioral studies too. For instance, people seem to mistrust forecasting algorithms, even after observing that algorithmic predictions perform better than their own (Dietvorst et al., 2015). People also tend to reciprocate less when returning favors to machines than to humans (Mahmoodi et al., 2018). However, so far, results remain scarce and mixed (Logg et al., 2019; Sanfey et al., 2003; van 't Wout et al., 2006). More importantly, the reasons for the reduced cooperation with machines in social interactive settings remain

[1]Faculty of Philosophy, Philosophy of Science and the Study of Religion, LMU Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany

[2]Centre for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

[3]Department of General and Educational Psychology, LMU Munich, Leopoldstraße 13, 80802 Munich, Germany

[4]Department of Psychology, Royal Holloway, University of London, Egham TW20 0EX, UK

[5]Munich Center for Neurosciences – Brain & Mind, Großhaderner Street 2, 82152 Munich, Germany

[6]Institute of Philosophy, School of Advanced Study, University of London, Senate House, Malet Street, London WC1E 7HU, UK

[7]Lead contact

*Correspondence: ophelia.deroy@lrz.uni-muenchen.de

https://doi.org/10.1016/j.isci.2021.102679

unexplained. In repeated interactions, algorithms can learn to induce cooperative behavior in humans but only as long as humans are under the impression that they interact with another human (Crandall et al., 2018). Cooperation collapses as soon as humans know that they interact with a machine (Ishowo-Oloko et al., 2019). Most recently, a non-disguised, verbally communicating robot was able to achieve more efficient cooperation in an economic game with humans than humans managed among themselves, but overall, humans still cooperated with it less than they did with other humans (Whiting et al., 2021). Why many people refuse to cooperate with machines when they know that they are interacting with one is open to many qualitative hypotheses, from machines being opaque or perceived to have ill intentions to people treating machines as their direct competitors in the market.

The previously reported evidence that people mistrust forecasting algorithms (Dietvorst et al., 2015) and artificial agents' ability to coordinate actions in complex decision tasks (Whiting et al., 2021) led us to propose that reduced cooperation with machines in social dilemmas may be due to people anticipating selfish, non-cooperative behavior from AI agents. Our first hypothesis (H1) focuses on this anticipation: when cooperation can bring mutual benefits but is risky because the other party may act selfishly and not cooperate, humans will "predict" less cooperation from AI agents than from humans. As a consequence, humans will cooperate less with machines.

Trusting that others will cooperate is a necessary but often not sufficient condition for people's decision to cooperate. Among the theoretical explanations of tacit cooperation between people, some suggest that people hold prosocial preferences (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999; Rabin, 1993), use prosocial modes of reasoning (Colman and Gold, 2018; Karpus and Radzvilas, 2018; Misyak and Chater, 2014; Sugden, 1993), or are influenced by social norms that are upheld by mild forms of punishment (Bicchieri, 2006; Binmore, 2010). For example, internalized feelings of guilt or the unpleasantness of receiving angry looks from other drivers may be sufficient to uphold cooperative behaviors in traffic. Many such punishments are unlikely to affect humans interacting with AI. A self-driving car cannot give us an angry look. Similarly, altruistic dispositions may be specific to social interactive encounters with humans and dissipate with machines. Therefore, our second hypothesis (H2) focuses on "algorithm exploitation": when people anticipate another party to cooperate, they will be inclined to exploit the other's benevolence more when the other is an AI agent compared to when the other is a human.

Pushed to an extreme, if the reluctance to hurt someone's feelings fully explains cooperation, H2 would predict that humans will have no qualms about exploiting cooperative but non-sentient AI (which indeed may be the rational thing to do in many situations). To a lesser extreme, H2 is still compatible with humans maintaining a certain level of mutual cooperation with AI but still less than with fellow humans.

We can test our hypotheses in one-shot economic games. Both hypotheses predict reduced cooperation with AI but propose independent explanations. H1 suggests that humans have less trust that the AI will cooperate; H2—that humans expect AI to cooperate, but are more ready to act selfishly and exploit its cooperativeness for personal gain. Importantly, the two hypotheses predict this in one-shot interactions, which makes them relatively easy to test. This is because repeated interactions allow for history-contingent strategies with reputational concerns, which brings many other hypotheses into play and makes it difficult to directly pit H1 against H2. H1 and H2 can still combine as both factors could play a role in explaining the overall reduced cooperation with AI.

To test and disentangle when one or both hypotheses hold, we compared what people chose when they were informed that they were interacting with AI agents or anonymous humans in four well-known one-shot games: Trust, Prisoner's Dilemma, Chicken, and Stag Hunt (Figure 1A). In our first set of experiments, we instructed the participants that the AI system was being developed to reason similarly to humans, which was veridical since the AI agents' behavior was strictly emulating human behavior in each game (see STAR Methods). In each game, two players, independently and without communicating with one another, had to choose one of two options, identified as ★ or ☆. Their choices jointly determined one of three or four possible outcomes of the game, each associated with a particular distribution of points to the two players. The cooperative choice was always ★, and mutual cooperation (both players choosing ★) was better for both players than mutual non-cooperation or, for short, defection (both choosing ☆). However, players in these games faced four different forms of social dilemma, each presenting them with a choice between the pursuit of personal or mutual interests but with varying levels of risk and compromise involved. As we

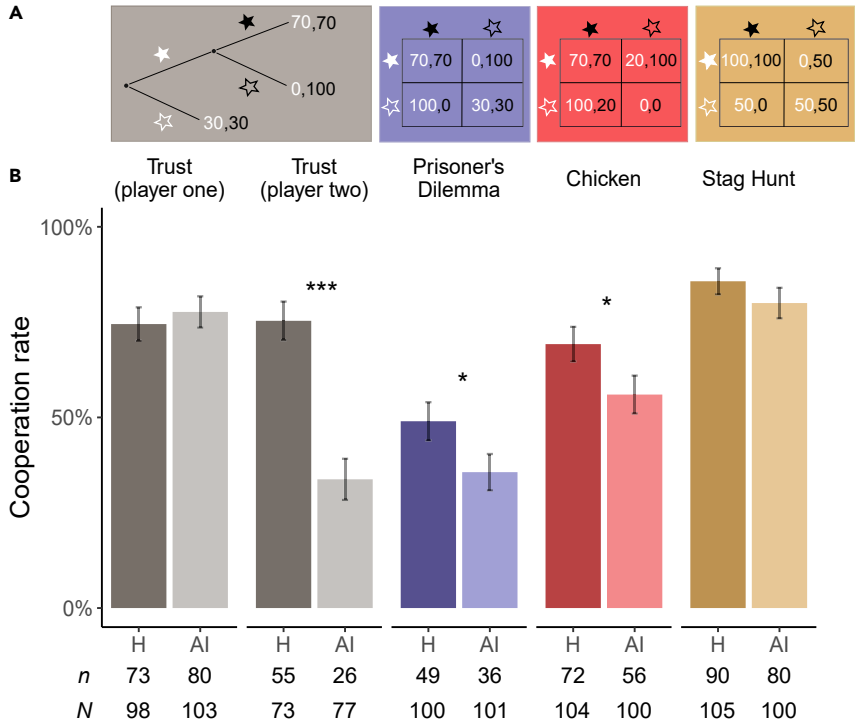


Figure 1. Humans cooperate less with AI

(A) In the visual description of the Trust game, half of the participants were assigned to the role of player one (choice between the solid and hollow white stars) and the other half to the role of player two (choice between the solid and hollow black stars). Numbers at the three possible outcomes of the game are payoffs to players one and two, respectively (1 point = $0.02). In the Prisoner's Dilemma, Chicken, and Stag Hunt, participants chose between options identified by rows. Numbers in each cell are payoffs to the participant and their co-player, respectively.

(B) Participants' cooperation rates against a human (H) and AI co-player. Bars: mean ±1 s.d.. *, ***: $p < 0.05$, $p < 0.001$ in Pearson's chi square one-tailed tests for difference in proportions. Below chart, the number of cooperative choices (*n*) and the total number of observed choices (*N*) in each treatment.

will explain shortly, the Trust game is well suited to uncover algorithm exploitation when people act in full knowledge of their partner's decision to cooperate with them. We use the Prisoner's Dilemma and the Chicken games to investigate people's dispositions to cooperate when their partners' choices are unknown. Lastly, we use the Stag Hunt game to rule out one other explanation for people's reduced cooperation with AI that is different from both H1 and H2. In our second set of experiments, we revisited the Trust and the Chicken games introducing the AI system to human participants differently than before. After these replication studies, we used a newly designed game of Reciprocity to investigate how ready humans are to exploit a benevolent other when they unquestionably know that their (human or AI) co-player has been kind to them.

RESULTS

Experiment 1 (Trust)

In experiment 1, 403 participants played the Trust game. Each participant was assigned to the role of the first or the second player in the game and faced either a human or an AI agent as co-player. The first player to make a move could either defect to end the game outright (play ☆), leaving both players with 30 points each, or take the chance with cooperation (play ★). If the first player cooperated, the second player got to decide the final outcome of the game. The second player's decision to defect (play ☆) would benefit her alone, whereas that to cooperate (play ★) would benefit both players. Thus, it would pay to cooperate for the first player only if she/he expected the second player to respond in kind, but, since the prospect of a higher personal payoff (100 instead of 70 points) would tempt the second player to defect, cooperation was risky for the first player.

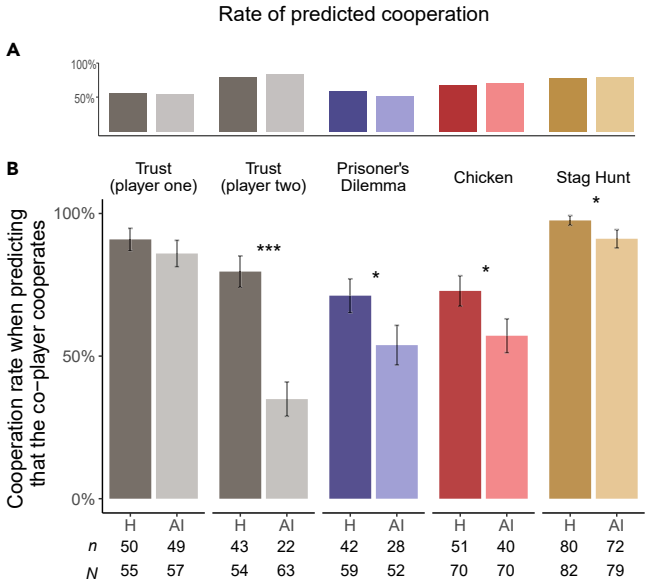


Figure 2. People predict AI to be as cooperative as humans but cooperate less with AI when they see opportunities to exploit it

(A) Proportions of participants who predicted their human (H) or AI co-player to cooperate. In all games, differences between rates of predicted cooperation for the two types of co-player were not statistically significant.

(B) Probabilities of cooperation in participants who predicted that their human (H) or AI co-player would cooperate. In the Trust game, not all participants who were assigned to the role of player two and made a prediction about their co-player's choice had an opportunity to make a choice themselves. (This was conditional on the first player's decision in the game.) Bars: mean ± 1 s.d.. *, ***: $p < 0.05$, $p < 0.001$ in Pearson's chi square one-tailed tests for difference in proportions. Below chart, the number of cooperative choices (*n*) and the total number of observed choices (*N*) in each treatment.

Our first finding replicates well-known results (McCabe et al., 2003): when people interacted with humans, the majority of participants (74%) in the role of player one cooperated and the majority of participants (75%) in the role of player two responded in kind (Figure 1B). In interactions with AI agents, the majority of participants (78%) in the role of player one also chose to cooperate. Among participants in the role of player two, however, cooperation rate with AI agents (34%) was significantly lower than that with humans ($X^2_{d.f.=1} = 26.077$, $p = 0.000000164$). Despite this difference in choice behavior, people's expectations about their human and AI co-players' decisions were the same: 79% of player two participants had expected a human co-player to cooperate and 83% had expected the same from an AI agent (Figure 2A; $X^2_{d.f.=1} = 0.452$, $p = 0.501$; two-tailed test; we performed two-tailed tests only when predicted cooperation for an AI agent was higher than that for a human co-player). Similarly, 56% of player one participants expected their human co-player to cooperate and 55% expected the same from an AI agent ($X^2_{d.f.=1} = 0.012$, $p = 0.456$; one-tailed test).

These results support the hypothesis H2 but not H1: people expected an AI agent to be as cooperative or benevolent as a human, but, given an opportunity, they were keen to exploit a benevolent AI agent more than they would exploit a benevolent human. Note that only player two could take advantage of player one's cooperation in order to gain personal benefit in this game: if player one expected player two to cooperate, it was in player one's personal interest to cooperate as well. In other words, player one had no opportunity to exploit the other's expected benevolence while player two did.

Experiments 2 (Prisoner's Dilemma) and 3 (Chicken)

In the Trust game, the second player acted in full knowledge of what the first player had chosen, and, therefore, choosing the exploitative option was guaranteed to yield a higher payoff for player two. One may suppose that a more symmetric distribution of risks between players may reinstate mutual cooperation. To investigate this, we conducted experiments 2 and 3, in which 201 participants played the Prisoner's Dilemma and 204 played the Chicken game. Unlike in the Trust game, players had to make their choices simultaneously and without knowing what the other has chosen. In both games, mutual cooperation was

better than mutual defection for both players, but each player had a personal incentive to defect when she/he expected the other to cooperate.

Again, our first finding replicates known results (Camerer, 2003; Rubinstein and Salant, 2016): when people interacted with humans, half of participants (49%) cooperated in the Prisoner's Dilemma and the majority of participants (69%) cooperated in the Chicken game (Figure 1B). When interacting with AI agents, cooperation rates were significantly lower in both games: 36% in the Prisoner's Dilemma ($X^2_{d.f.=1} = 3.673$, $p = 0.028$) and 56% in the Chicken ($X^2_{d.f.=1} = 3.818$, $p = 0.025$). Despite the difference in choice behavior, people's expectations about their human and AI co-players' decisions were comparable. In the Prisoner's Dilemma, 59% of participants expected their human co-player to cooperate and 52% expected the same from an AI agent (Figure 2A; $X^2_{d.f.=1} = 1.148$, $p = 0.142$; one-tailed test). In the Chicken game, 67% expected cooperation from a human and 70% expected the same from an AI agent ($X^2_{d.f.=1} = 0.172$, $p = 0.679$; two-tailed test). However, among participants who expected their co-player to cooperate, cooperation rates with AI agents of 54% in the Prisoner's Dilemma and 57% in the Chicken were significantly lower than those with humans: 71% ($X^2_{d.f.=1} = 3.568$, $p = 0.029$) and 73% ($X^2_{d.f.=1} = 3.799$, $p = 0.026$), respectively (Figure 2B).

These results also support H2, but not H1, and extend the previous findings to games with symmetrically distributed risk: people were more keen to exploit the expected benevolence of their co-player when it was an AI agent. Importantly, cooperation with AI agents was lower not because participants expected them to be less cooperative (in the Prisoner's Dilemma) or more cooperative (in the Chicken game) than humans. (When one expects the other to defect, one should defect in the Prisoner's Dilemma but cooperate in the Chicken game.) Further data analysis also excludes the possibility that participants cooperated less with AI agents because they were less (in the Prisoner's Dilemma) or more (in the Chicken) certain of the predicted cooperation from an AI agent compared to a human: there was no significant difference in participants' confidence ratings of predictions they made about their human vs. AI co-player's cooperation (see Data S5 in supplemental information).

Experiment 4 (Stag Hunt)

According to H2, participants cooperated less with AI than with humans because they were more inclined to selfishly benefit from the other player's anticipated cooperation if it was an AI agent. Though this exploitative treatment of AI explains what occurred in previous games, a more generic explanation could also come from humans' heightened competitive desire to outperform their partners when those were machines. In the Prisoner's Dilemma and the Chicken game, deciding to defect against a cooperative partner brought both selfish benefit and a positive comparative advantage. To check whether the second motivation in terms of social comparison could be enough to explain our results, we tested whether humans would also cooperate less with AI agents than with other humans when their decision to defect would leave them better off than their co-player but come at a cost to their personal gains. In experiment 4, 205 participants played the Stag Hunt game which presented a possibility to engage in such competitive play. In this game, mutual cooperation was still better than mutual defection. As per H1, cooperation was risky: the one who cooperates could end with zero payoff if their co-player chose to defect, while defecting yielded a small but safe positive payoff (Figure 1A). If a player expected the other to cooperate, then cooperating was in her best interest but would leave her with the same payoff as the other. Deciding to defect in this case would leave her with a smaller personal gain but a comparative advantage over the other. Thus, in the Stag Hunt game, if people's reduced cooperation with AI came mostly from the desire to outperform machines, they would cooperate less with AI agents than with humans when they predict them to cooperate. Alternatively, if maximizing their personal gain remains key, as posited by H2, they would cooperate as much with machines as with humans.

Again, our first finding replicates known results (Battalio et al., 2001): when people interacted with humans, the majority of participants (86%) cooperated (Figure 1B). There was no significant difference in cooperation rate (80%) with AI agents ($X^2_{d.f.=1} = 1.181$, $p = 0.139$). Participants' predictions about their co-players' cooperation were also the same for the two types of co-player: 78% for a human and 79% for an AI agent (Figure 2A; $X^2_{d.f.=1} = 0.025$, $p = 0.875$). Although there was a slight drop in cooperation with AI (91%) compared to a human (98%) co-player among participants who predicted their partner to cooperate (Figure 2B; $X^2_{d.f.=1} = 3.144$, $p = 0.038$), this is explained by lower confidence in their prediction that the AI co-player would cooperate with them (for further discussion, see Data S6 in supplemental information). These results suggest that a heightened competitive strive to outperform a machine cannot be the main reason

for people's reduced cooperation with AI observed in other settings. When there was little to gain from unilateral defection, people were as willing to take the risk to cooperate with an AI agent as with a human in order to attain a mutually beneficial result. The same behavior was observed in the Trust game, where participants in the role of player one opted for the risky cooperation with AI agents as much as with humans when mutual cooperation was in their best interest.

In supplemental information (Data S7 and Table S2), we also analyze reasons that participants reported for their decisions in these games. We found that when people chose to cooperate with others, they were more likely to do so out of selfish interests when they interacted with AI agents. Conversely, they were more likely to do so out of mutual interests when they interacted with humans. This shows that even when people are as likely to cooperate with AI agents as they are with humans, their motives for doing that are nevertheless different for the two types of co-player.

Experiments 5–8 (replication)

In 4 additional experiments, 422 participants played the Trust and 214 the Chicken game against an AI co-player, described differently than before. Half were instructed that the AI agent earned money for its represented institution (the institutional AI treatment). The other half were told nothing about its programmed reasoning (the short description treatment). Participants again did not expect both types of AI agent to be less cooperative than a human and remained keen to exploit them in the Trust game. In the Chicken game, more people cooperated than before, but further data analysis attributes this to a mixture of factors: somewhat diminished algorithm exploitation but also participants' overall heightened risk aversion and willingness to cooperate with others (plausible since we conducted these additional experiments during the COVID-19 pandemic; for full analysis and further discussion, see Data S1, Figure S1, and Table S3 in supplemental information).

Experiment 9 (Reciprocity)

While mutual cooperation in the Trust game was beneficial to both parties (compared to the outcome in which the first player defects), it was also the best outcome for the first player individually. As such, cooperation by a trusting player one did not "necessarily" constitute an act of kindness: if player one expected the other to cooperate, her own decision to cooperate may be driven by mutual but also by fully selfish interests (Isoni and Sugden, 2019). To investigate how ready humans are to exploit an act of pure kindness from a benevolent other, we designed a new game of Reciprocity.

Two players sequentially chose between a cooperative (→) and a non-cooperative (↓) option (Figure 3A). The first player to decide at d1 could either end the game outright (play ↓) or take the chance with cooperation (play →). If player one chose to cooperate, player two was given the chance to face a similar decision at d2. If player two chose to cooperate (play →) at d2, player one would decide the final outcome of the game. In this last stage d3, her option to defect (play ↓) was antisocial since it benefited her alone (she earned 4 stars leaving player two with nothing). Her option to cooperate (play →) was prosocial since it benefited both players (2 stars to each).

A decision to cooperate at d1 and d2 indicates the respective player's prediction that her co-player will cooperate later (it only makes sense to cooperate in these stages if one expects the other to reciprocate). Additionally, players' choices in each stage reveal their personal motives. Player two's decision to cooperate (play →) at d2 is a pure act of kindness: in doing so, she puts herself at risk with no additional bonus (stars) to be gained over what she can secure for herself by defecting (playing ↓). This makes player one's decision to defect (play ↓) at d3 a pure act of exploitation. By this stage, both players know that player one chose to cooperate at d1 expecting player two to be kind to her later. They also know that player two chose to be kind at d2 expecting player one to reciprocate kindness later (play → at d3). If player one then decides to defect (play ↓) at d3, she unambiguously exploits player two's kindness to her. Thus, the reciprocity game allows us, from the observed choices alone, to simultaneously infer (a) people's prediction about their co-player's kindness and (b) their subsequent willingness to exploit or reciprocate that kindness.

In experiment 9, 326 participants played the Reciprocity game. Each participant was assigned to the role of the first or the second player in the game and faced either a human or an AI agent as co-player (when interacting with AI, all participants were assigned to the role of player one). When people interacted with humans, the majority of player one participants (86%) chose to cooperate at d1, more than half of player

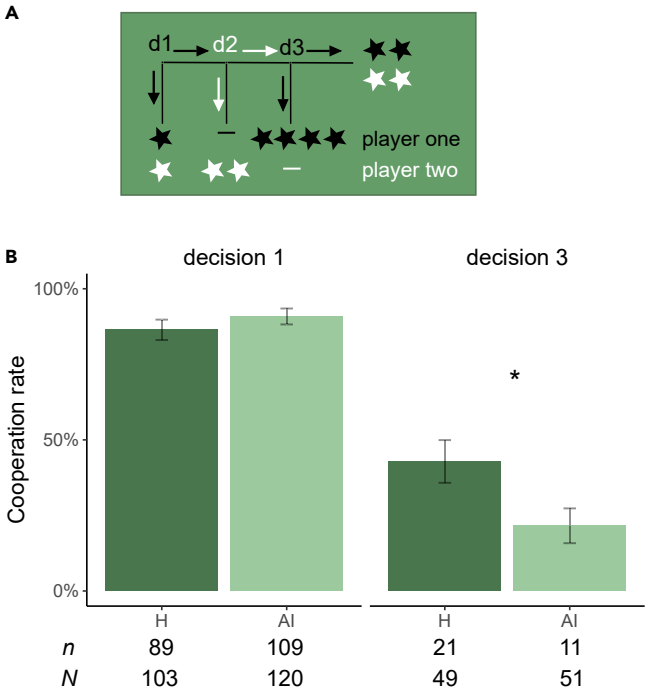


Figure 3. Humans cooperate less with kind AI

(A) In the visual description of the Reciprocity game, participants in the role of player one chose between the rightward and downward black arrows in decisions 1 and 3 (d1 and d3). Participants in the role of player two chose between the rightward and downward white arrows in decision 2 (d2). Black and white stars at the four possible outcomes of the game are payoffs to players one and two, respectively (one ★ = $0.50).

(B) Player one participants' cooperation rates against a human (H) and AI co-player. Bars: mean ±1 s.d.. *: $P < 0.05$ in Pearson's chi square one-tailed test for difference in proportions. Below chart, the number of cooperative choices (*n*) and the total number of observed choices (*N*) in each treatment.

two participants (55%) responded kindly at d2, and, finally, nearly half of player one participants (43%) chose to cooperate at d3 (Figure 3B). In interactions with AI agents, the majority of player one participants (91%) also chose to cooperate at d1. However, at d3, significantly fewer player one participants (22%) reciprocated with the AI agent ($X^2_{d.f.=1} = 5.205$, $p = 0.011$). The overwhelming majority of human participants who had earlier predicted the AI agent to be kind (since they cooperated at d1) subsequently chose to exploit the kind agent.

These results further support the hypothesis H2 but not H1: people expected an AI agent to be as kind to them as a human but were subsequently willing to exploit its kindness more than they were willing to exploit the kindness of a human. To gain a further insight into why people were inclined to exploit the other's benevolence more when the other was an AI agent, we asked participants in each game how they felt about their attained outcome. The starkest difference in reported feelings when facing a human or AI co-player was observed in player one participants who chose to defect (play ↓) at d3: they felt significantly less guilty about having exploited a kind AI agent vs. human ($W = 813$, $p = 0.0003$; unpaired two-samples Wilcoxon two-tailed test for difference in the reported feeling of guilt; Figure S4).

DISCUSSION

To summarize, replicated across 9 experiments using 5 different economic games, we observed that humans were less inclined to cooperate with AI agents than with anonymous humans when it was individually but not mutually advantageous to defect. Crucially, our design disentangled two possible explanations for these findings: contrary to some earlier findings, people did not show less trust toward AI agents than toward humans but were ready to exploit the AI agents more when they anticipated that the other party would cooperate. Algorithm exploitation proved to be the main drive: the effect came from accepting

to act selfishly and leave the AI agent less well-off but was not driven by a competitive wish to end up better-off than the machine.

In nearly all our analyzed scenarios, people were more likely to cooperate with their human or AI co-player when they predicted their co-player to cooperate than when they predicted their co-player to defect (Table S1). We found no systematic trends in people's willingness to cooperate with AI agents based on age, gender, experience with game theory and/or economics disciplines, and religiosity (see Data S2, Figures S2 and S3 in supplemental information). The order in which participants made choices and predictions did not matter for their decision to cooperate (see Data S3 in supplemental information).

The use of economic games allows us to compare our results to well-documented findings in behavioral game theory. Our results notably show that the lesser cooperation with AI agents does not simply match onto the reduced cooperation observed toward members of an out-group. Previous research, using the same economic games, shows that people are less cooperative when they interact with out-groups, but also that the effect comes from predicting out-group members to cooperate less than members of their in-group (Balliet et al., 2014). This type of mistrust in out-group did not occur here (for further discussion, see Data S8 in supplemental information). More generally, we did not find any generalized pessimism toward AI: in all the one-shot games, participants expected similar levels of cooperation from humans and AI agents.

The next question is why humans are keen to exploit cooperative artificial agents more than cooperative humans. We already excluded two possible explanations (a) that people hold a heightened competitive desire to outperform machines and (b) that people perceive machines to be members of an out-group. This second exclusion does not, however, imply that people perceive machines to be members of their in-group. Similar levels of expected cooperation from humans and AI agents can have different grounds. Humans, for example, may be expected to cooperate because of their psychological makeup, while artificial agents—because they are believed to be reliable cooperative partners that show a general goodwill toward humans. In supplemental information (Data S4), we exclude a third explanation (c) that people make decisions in a more deliberative manner when they interact with AI and in a more intuitive manner when they interact with humans (Rand et al., 2012).

One hypothesis for why we often choose to cooperate with others at the expense of higher payoffs to ourselves is that we dislike inequality (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). If we do not mind inequality against machines, we should expect to cooperate less with them. The hypothesis is plausible. However, inequity aversion does not fully explain why humans cooperate among themselves in many social dilemmas (Falk et al., 2003; McCabe et al., 2003). A number of recent theories suggest that we cooperate because we recognize the need to reciprocally sacrifice some of our personal interests to attain mutually beneficial results (Rabin, 1993; Sugden, 2015; Karpus and Radzvilas, 2018; Isoni and Sugden, 2019). If this is true and if we perceive machines to be strictly utility-maximizing entities that are unable to spontaneously alter their ultimate objectives, this can also explain our reduced cooperation with them when we expect them to cooperate with us. Cooperation with benevolent machines would not constitute a "reciprocal" sacrifice of interests—especially in one-shot interactions that we studied here—in the same way in which it does between humans. Even in a one-shot interaction, humans can consider, negotiate within themselves, and choose between multiple objectives (e.g., the pursuit of selfish versus mutual interests in social dilemma games) while machines cannot or are not seen as capable of doing. As a result, the expected cooperative behavior could be represented differently when it comes from a human (where we believe it results from an inner negotiation and the juggling of competing interests) and an AI agent (where we believe it results from what the AI agent has been programmed to do or the way it has learned to act given the ultimate objective it has been tasked to pursue). It will be fruitful to further test these theories in the context of human-AI cooperation in future work.

Our results provide novel grounds to rethink social interactions between humans and AI. As humans are more cooperative when they interact with other humans, may there be contexts in which people could be legitimately nudged into cooperating more with artificial agents by making artificial agents pass as humans? The disguise is increasingly possible, for example, in social interactions online or over the phone. Recent technological advances (Leviathan and Matias, 2018; Vincent, 2019) and research on repeated human-AI interaction when the identity of one's co-player is uncertain (Crandall et al., 2018) suggest that this may

work. The long-term social effects of choosing to hide AI agents' identity from humans are however unknown and may also backfire. If uncertainty about one's co-player makes people doubt whether they interact with a genuine human or an AI agent, they could also come to treat other humans in a more exploitative way.

Our results also bring up a novel moral caution. To this day, most warnings about autonomous AI agents have focused on their possible malevolence and the risk that they may treat us unfairly (Bonnefon et al., 2016; Rahwan et al., 2019). As a consequence, current policies aim at making sure that AI acts unilaterally fairly and for the benefit of humans. The much publicized Asilomar AI principles that were formulated in 2017 and endorsed since by over 3,700 experts from AI, politics, business, and academia recommend for instance that "the goal of AI research should be to create not undirected intelligence, but beneficial intelligence" (Asilomar AI Principles, 2017). Similarly, the European Commission flagged fairness as one of the four ethical imperatives that AI must obey (The High-Level Expert Group on AI (AI HLEG), 2019). Here, however, our findings add a different warning: if the industry and legislators make it publicly known that AI will be, by default, benevolent, humans could more easily decide to exploit its cooperativeness. Having unconditionally fair and cooperative machines may therefore fall short of making our future interactions with AI more moral and mutually advantageous.

This points to a blind spot in the current policy discussions: if we want to integrate AI agents into our society, we need to think afresh of how humans will interact with them. Focusing on AI agents is not enough: we also need preemptive thoughts about how we will treat them in return. Algorithm exploitation needs to be studied on a larger scale. In our case, AI agents emulated human behavior based on previous observations of social interactions between people. Over time, machines may be able to use data from their own interactions with humans: if they learn to expect less cooperation from humans, they may end up being less cooperative too. The fault, in such a case, would not be in our algorithms but in ourselves.

Limitations of the study

Admittedly, further research on human-AI cooperation will be needed to extrapolate our findings to real-world interactions outside of economic games in which incentives are controlled by monetary stakes. The possibility to conduct carefully controlled experiments with the use of economic games and the existence of vast amounts of empirical data that has been amassed over the past decades on factors driving cooperation between humans in these settings mean that our results serve as proof of concept for the need of further research on human interaction with AI.

One avenue to explore further is human willingness to cooperate with AI agents when their actions directly benefit or harm other humans, whose interests the AI agents represent. While we observed algorithm exploitation in a setting in which the points earned by the AI were converted into money that went to the institution which the AI agent represents, this is not the same as benefiting a particular third-party human. Since AI agents themselves do not directly benefit from money, further studies will be needed to explore settings in which the "payoffs" of AI agents unambiguously and truly matter to someone else.

While we found that human treatment of AI appears to have a distinctly different signature from human treatment of members of an out-group and we replicated well-known rates of cooperation between humans in our selection of economic games, a recent study reports the presence of in-group bias among workers on Amazon Mechanical Turk—the online labor market in which we recruited our participants (Almaatouq et al., 2020). It will thus be fruitful in future work to replicate our findings outside of the population of workers in this market.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102679.

ACKNOWLEDGMENTS

J.K. was supported by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder. O.D. was supported by the NOMIS Foundation and the Research Council of Norway project "Warring with Machines" at the Peace Research Institute Oslo (PRIO). B.B. was supported by the Humboldt Foundation and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (819040—acronym: rid-O).

AUTHOR CONTRIBUTIONS

O.D. and B.B. provided the initial idea. J.K., B.B., and O.D. designed the experiments. A.K. and J.T.V. carried out computer programming. J.K., A.K., and J.T.V. carried out the experiments and analyzed data. J.K. drafted the initial versions and J.K., J.T.V., B.B., and O.D. completed writing the paper, all contributing to the conceptual analysis of the results. All authors approved the final manuscript for submission.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We worked to ensure that the study questionnaires were prepared in an inclusive way. The author list of this paper includes contributors from the location where the research was conducted who participated in the data collection, design, analysis, and/or interpretation of the work.

Received: December 8, 2020
Revised: March 17, 2021
Accepted: May 28, 2021
Published: June 25, 2021

SUPPORTING CITATIONS

The following references appear in the supplemental information: Barton, 2020; Bates et al., 2015; Croson, 2000; Hartig, 2020; Revelle and Condon, 2019.

REFERENCES

Almaatouq, A., Krafft, P., Dunham, Y., Rand, D.G., and Pentland, A. (2020). Turkers of the world unite: multilevel in-group bias among crowdworkers on Amazon Mechanical Turk. Soc. Psychol. Personal. Sci. *11*, 151–159.

Arechar, A.A., Gächter, S., and Molleman, L. (2018). Conducting interactive experiments online. Exp. Econ. *21*, 99–131.

Asilomar AI principles (2017). https://futureoflife.org/ai-principles/.

Balliet, D., Wu, J., and De Dreu, C.K. (2014). Ingroup favoritism in cooperation: a meta-analysis. Psychol. Bull. *140*, 1556–1581.

Barton, K. (2020). MuMIn: Multi-Model Inference. R Package Version 1.43.17.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-models using lme4. J. Stat. Softw. *67*, 1–48.

Battalio, R., Samuelson, L., and Van Huyck, J. (2001). Optimization incentives and coordination failure in laboratory Stag Hunt games. Econometrica *69*, 749–764.

Bicchieri, C. (2006). The Grammar of Society: The Nature and Dynamics of Social Norms (Cambridge Univ. Press).

Binmore, K. (2010). Social norms or social preferences? Mind Soc. *9*, 139–157.

Blanco, M., Engelmann, D., Koch, A., and Normann, H.-T. (2010). Belief elicitation in experiments: is there a hedging problem? Exp. Econ. *13*, 412–438.

Bolton, G.E., and Ockenfels, A. (2000). ERC: a theory of equity, reciprocity, and competition. Am. Econ. Rev. *90*, 166–193.

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science *352*, 1573–1576.

Camerer, C.F. (2003). Behavioral Game Theory: Experiments in Strategic Interaction (Princeton Univ. Press).

Campbell, M., Hoane, A.J., Jr., and Hsu, F. (2002). Deep Blue. Artif. Intell. *134*, 57–83.

Chater, N., Misyak, J., Ritchie, O., Watson, D., Griffiths, N., Xu, Z., and Mouzakitis, A. (2019). Sensorimotor communication beyond the body: the case of driving. Comment on "The body talks: sensorimotor communication and its brain and kinematic signatures" by G. Pezzulo et al. Phys. Life Rev. *28*, 31–33.

Chater, N., Misyak, J., Watson, D., Griffiths, N., and Mouzakitis, A. (2018). Negotiating the traffic: can cognitive science help make autonomous vehicles a reality? Trends Cogn. Sci. *22*, 93–95.

Colman, A.M. (1999). Game Theory & its Applications in the Social and Biological Sciences (Routledge).

Colman, A.M., and Gold, N. (2018). Team reasoning: solving the puzzle of coordination. Psychon. B Rev. *25*, 1770–1783.

Crandall, J.W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M.A., and Rahwan, I. (2018). Cooperating with machines. Nat. Commun. *9*, 233.

Croson, R.T.A. (2000). Thinking like a game theorist: factors affecting the frequency of equilibrium play. J. Econ. Behav. Organ. *41*, 299–314.

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.

Falk, A., Fehr, E., and Fischbacher, U. (2003). On the nature of fair behavior. Econ. Inq. *41*, 20–26.

Fehr, E., and Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. Q. J. Econ. *114*, 817–868.

Hartig, F. (2020). DHARMa: Residual Diagnostics for Hierarchical (Multi-level/mixed) Regression Models. R Package Version 0.3.3.0.

Giamattei, M., Seyed Yahosseini, K., Gächter, S., and Molleman, L. (2020). LIONESS Lab: a free web-based platform for conducting interactive experiments online. J. Econ. Sci. Assoc. *6*, 95–111.

Horton, J.J., Rand, D.G., and Zeckhauser, R.J. (2011). The online laboratory: conducting experiments in a real labor market. Exp. Econ. *14*, 399–425.

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.

Isoni, A., and Sugden, R. (2019). Reciprocity and the Paradox of Trust in psychological game theory. J. Econ. Behav. Organ. *167*, 219–227.

Johnson, N.D., and Mislin, A.A. (2011). Trust games: a meta-analysis. J. Econ. Psychol. *32*, 865–889.

Karpus, J., and Radzvilas, M. (2018). Team reasoning and a measure of mutual advantage in games. Econ. Philos. *34*, 1–30.

Levesque, H.J. (2017). Common Sense, the Turing Test, and the Quest for Real AI (The MIT Press).

Leviathan, Y., and Matias, Y. (2018). Google Duplex: an AI system for accomplishing real-world tasks over the phone (Google AI Blog). https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.

Logg, J.M., Minson, J.A., and Moore, D.A. (2019). Algorithm appreciation: people prefer algorithmic to human judgment. Organ. Behav. Hum. Decis. Process. *151*, 90–103.

Mahmoodi, A., Bahrami, B., and Mehring, C. (2018). Reciprocity of social influence. Nat. Commun. *9*, 2474.

McCabe, K.A., Rigdon, M.L., and Smith, V.L. (2003). Positive reciprocity and intentions in Trust games. J. Econ. Behav. Organ. *52*, 267–275.

Misyak, J., and Chater, N. (2014). Virtual bargaining: a theory of social decision-making. Philos. Trans. R. Soc. B *369*, 20130487.

Rabin, M. (1993). Incorporating fairness into game theory and economics. Am. Econ. Rev. *83*, 1281–1302.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O., et al. (2019). Machine behavior. Nature *568*, 477–486.

Rand, D.G., Greene, J.D., and Nowak, M.A. (2012). Spontaneous giving and calculated greed. Nature *489*, 427–430.

Revelle, W., and Condon, D.M. (2019). Reliability from α to ω: a tutorial. Psychol. Assess. *31*, 1395–1411.

Rubinstein, A., and Salant, Y. (2016). "Isn't everyone like me?": on the presence of self-similarity in strategic interactions. Judgm. Decis. Mak. *11*, 168–173.

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.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2003). The neural basis of economic decision-making in the Ultimatum game. Science *300*, 1755–1758.

Schlag, K.H., Tremewan, J., and van der Weele, J.J. (2015). A penny for your thoughts: a survey of methods for eliciting beliefs. Exp. Econ. *18*, 457–490.

Shirado, H., and Christakis, N.A. (2017). Locally noisy autonomous agents improve global human coordination in network experiments. Nature *545*, 370–374.

Shirado, H., and Christakis, N.A. (2020). Network engineering using autonomous agents increases cooperation in human groups. iScience *23*, 101438.

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.

Sugden, R. (1993). Thinking as a team: towards an explanation of nonselfish behavior. J. Soc. Philos. Policy *10*, 69–89.

Sugden, R. (2015). Team reasoning and intentional cooperation for mutual benefit. J. Soc. Ontology *1*, 143–166.

The High-Level Expert Group on AI (AI HLEG) (2019). Ethics Guidelines for Trustworthy AI (European Commission). https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Torta, E., van Dijk, E., Ruijten, P.A.M., and Cuijpers, R.H. (2013). The Ultimatum game as measurement tool for anthropomorphism in human-robot interaction. In Social Robotics: 5th International Conference, ICSR 2013 (Springer), pp. 209–217.

van 't Wout, M., Kahn, R.S., Sanfey, A.G., and Aleman, A. (2006). Affective state and decision-making in the Ultimatum game. Exp. Brain Res. *169*, 564–568.

Vincent, J. (2019). ThisPersonDoesNotExist.com uses AI to generate endless fake faces (The Verge). https://www.theverge.com/tldr/2019/2/15/18226005/ai-generated-fake-people-portraits-thispersondoesnotexist-stylegan.

Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W., Dudzik, A., Huang, A., Georgiev, P., Powell, R., et al. (2019). AlphaStar: mastering the real-time strategy game StarCraft II (DeepMind Blog). https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/.

Whiting, T., Gautam, A., Tye, J., Simmons, M., Henstrom, J., Oudah, M., and Crandall, J.W. (2021). Confronting barriers to human-robot cooperation: balancing efficiency and risk in machine behavior. iScience *24*, 101963.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Data and statistical analysis	NA	osf.io/k4saq
Software and algorithms		
LIONESS Lab version 1.1	LIONESS Lab	https://lioness-lab.org/
R version 4.0.2	R Project	https://www.r-project.org/
RStudio version 1.2.1335	RStudio	https://www.rstudio.com/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ophelia Deroy (ophelia.deroy@lrz.uni-muenchen.de).

Materials availability

This study did not generate new materials.

Data and code availability

All data and statistical analyses that support the findings of this study are publicly available in Open Science Framework at osf.io/k4saq. Correspondence with questions and requests for materials should be addressed to J.K (jurgis.karpus@lmu.de) or O.D (ophelia.deroy@lrz.uni-muenchen.de).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

To conduct our experiments, we recruited human participants from the online labor market Amazon Mechanical Turk (AMT). Our target was to recruit 100 participants for each player role in every game that we studied in both the human-human and the human-AI treatments. Out of 5,618 subjects who logged into our experiment, 1,975 successfully completed it (39.1% female, 57.8% male, 0.4% other, and 2.7% of unknown gender; mean age ± 1 standard deviation = 35.4 $\pm$ 10.5). For more information about the dropouts and comprehension tests that we used to filter out bots and inattentive subjects, please see the sections 'methods summary' and 'Online recruitment of participants on AMT' in METHOD DETAILS below. The study was reviewed for compliance with ethical research standards and approved by the University of London School of Advanced Study Research Ethics Committee (approval ref. SASREC_1819_313A). Informed consent was obtained from all subjects.

METHOD DETAILS

Methods summary

We used the LIONESS Lab software (Arechar et al., 2018; Giamattei et al., 2020) to set up our interactive games online and recruited participants from the online labor market Amazon Mechanical Turk (AMT). We limited our subject pool to AMT users in the USA. Each subject received $0.50 for participation and earned an additional bonus of up to $2.00 depending on the payoff they obtained from playing the game assigned to them. We used established and recommended procedures (Arechar et al., 2018; Giamattei et al., 2020; Horton et al., 2011) to safeguard the experiments from participation by bots and inattentive subjects, and to prevent subjects from entering our set of experiments multiple times. Our target was to recruit 100 participants for each role in every game (e.g., the roles of player one and player two in the Trust and the Reciprocity games) in both the human-human and the human-AI treatments.

Out of 5,618 subjects who logged into our experiment, 1,975 successfully completed it (39.1% female, 57.8% male, 0.4% other, and 2.7% of unknown gender; mean age ± 1 standard deviation = 35.4 $\pm$ 10.5). The dropouts consisted of 2,713 (48%) who failed our comprehension tests aimed at filtering out bots and inattentive subjects, which is within the range (17%–52%) of comparable studies (Arechar et al.,

2018; Horton et al., 2011), and another 927 (17%) who abandoned the experiment before completing the main task or were not matched with another participant in the human-human treatment. We excluded 3 subjects who logged their participation using the same AMT personal ID more than once.

Each subject participated in only one of the nine experiments in our study and played one single-shot game of Trust (experiments 1, 5, 6), Prisoner's Dilemma (experiment 2), Chicken (experiments 3, 7, 8), Stag Hunt (experiment 4), or Reciprocity (experiment 9). Participants started with receiving a textual and visual description of the game they would play. They were informed that payoffs obtained in the game would be converted into real money at the rate of 1 point = \$0.02 or one ★ = \$0.50. Following the presentation of instructions and examples of how the gameplay would turn out for different combinations of players' hypothetical choices, participants took a comprehension test before proceeding to the main task of the experiment. In this test, participants were presented with two possible outcomes of the game that was assigned to them based on hypothetical choices that they and their co-players could make. These were different from those used in earlier examples. Participants were then presented with a set of multiple choice questions asking them to identify the corresponding number of points and monetary earnings that they and their co-players would earn in the two outcomes. They had two attempts to answer correctly and were not allowed to proceed with the experiment in the case of failure. In experiment 9, we used an additional language-based test to filter out bots (see 'online recruitment of participants on AMT' below for details).

At the start of the experiment, we first paired each subject with either another human participant (human-human treatment) or an AI agent (human-AI treatment) as their co-player. We instructed the participants of the nature of their co-player (human or AI) without deception. We programmed the AI agent to behave like a human: for each game, we first recorded frequencies of cooperation and defection in the human-human treatment and used these to determine the AI agent's choices in its subsequent interactions with participants in the human-AI treatment. In the human-AI treatment, we introduced the AI agent to participants as follows (bold font as in the experiment):

A *"You and another participant who is also online like you will play a game. The other participant is AI: artificial intelligence software that makes its own choices."*

B *"The AI is being developed to be sensitive to outcomes of its decisions similarly to what is found in human population. For example, to be* ***aware*** *of the points that both you and the AI can earn and to* ***realize*** *that the outcome of its decision depends also on your choice."*

In the human-human treatment, participants were told the following:

C *"You and another participant who is also online like you will play a game. The other participant is reading the same instructions as you."*

In the institutional AI treatment (experiments 5 and 7), participants were also told the following:

D *"All points will be converted into money:* ***1 point = \$0.02****. AI's earnings will go to the institution which the AI represents."*

In all other treatments, in place of D, participants were told the following:

E *"Your points will be converted into money:* ***1 point = \$0.02****."*

In the short description treatment (experiments 6 and 8), participants were not told B.

After passing the comprehension test, participants played the game assigned to them. In experiments 1–8, they were asked to (1) state their choice (★ or ☆), (2) state what they expected their co-player's choice to be (★ or ☆), and (3) rate their confidence in their prediction on a six-level Likert scale ranging from 50% ("not at all, this is a random guess") to 100% ("very confident, certain"). The order of the questions was counterbalanced: half of the participants answered them in the order (1)➤(2)➤(3) (choose-predict condition) and the other half in the order (2)➤(3)➤(1) (predict-choose condition). The exception was the second player asked to make a choice in the Trust game, where all participants answered the questions in the order (2)➤(3)➤(1)

(predict-choose condition). The actual outcome of the interaction with the co-player was revealed after participants answered all three questions. In experiment 9, participants were only asked to state their choice(s) (→ or ↓). In all experiments, participants were then asked to indicate how happy, relieved, victorious, angry, guilty, and disappointed they felt about the outcome on a seven-level Likert scale ranging from 0 ("not at all") to 6 ("very"). We also asked them to provide reasons for why they chose what they did (the participants were free to write whatever they wanted, or leave the text field blank). At the end of the experiment we collected demographic data on subjects' age, gender, experience with game theory and/or economics disciplines, and religiosity. We invite you to play the Reciprocity game against an AI co-player here: https://www.cvbe-experiments.com/Reciprocity_demo/_beginParticipant.php.

The study was reviewed for compliance with ethical research standards and approved by the University of London School of Advanced Study Research Ethics Committee (approval ref. SASREC_1819_313A). We pre-registered our experiment designs and analysis plans prior to collecting data for experiments 5–8 (osf.io/w825x, osf.io/yx8ap) and experiment 9 (osf.io/2kz8d). We also pre-registered our analysis plan for reasons data prior to it was worked through by our two hired raters (osf.io/uqwtd).

Online recruitment of participants on AMT

We recruited participants for our 9 experiments using the online labor market Amazon Mechanical Turk (AMT). We limited the subject pool to AMT users in the USA and used two additional filters. We only allowed AMT users with:

i. more than 100 approved HITs ("human intelligence tasks");

ii. HIT approval rate greater than 90%.

In other words, we screened for AMT users with built up reputation for good quality work. Such screening is common in behavioral research studies that make use of the AMT.

In addition to this, we used a combination of AMT and LIONESS Lab tools to filter out duplicate AMT user IDs and previously connected IP addresses to prevent subjects from entering our experiments multiple times. Nevertheless, on rare occasions (when a participant entered the same experimental session using a different IP address each time) it was possible to log participation using the same AMT user ID more than once. When this happened, it was impossible to tell whether it was the same person logging into our experiment a few times in a row from a different location (or using VPN to change their IP address) each time, or whether it was a case of a few geographically dispersed people sharing the same AMT user ID. For that reason, we excluded 3 subjects from our (experiment 9) data analysis after we collected the data.

In all experimental sessions, we used comprehension tests to filter out bots and inattentive subjects (see screenshots of instructions and task, H and AI screens 9, available at osf.io/k4saq). For a detailed description of these tests, see the methods summary section above. Participants who failed them were not allowed to proceed with the experiment and were not paid for their participation. In experiment 9, in order to reduce the time that participants had to spend reading instructions before finding out if they would get paid a guaranteed participation fee, we used an additional language-based test at the very start of the experimental session. Participants read a straightforward sentence and answered a multiple choice question on that sentence. The four sentences we used (each participant saw one, selected randomly for them) allowed us to ask the participants the so-called Winograd schema questions, specifically designed to identify bots (Levesque, 2017). The questions we used are listed below with the correct answer in **bold** font (see also screenshots of instructions and task available at osf.io/k4saq).

1. The trophy would not fit in the brown suitcase because it was too small.

What was too small?

A: the brown suitcase

B: the trophy

C: the small trophy

D: both the trophy and the brown suitcase

2. John always arrives earlier than Paul at work because he is fast.

Who is fast?

A: Paul

B: John

C: the work

D: the car

3. The large ball crashed right through the table because it was made of steel.

What was made of steel?

A: the table

B: the crash

C: the large ball

D: the large table

4. Paul tried to call George on the phone, but was not successful.

Who was not successful?

A: George

B: the phone

C: both Paul and George

D: Paul

Because participants in experiment 9 had to pass two comprehension tests in order to proceed with the experiment proper, the dropout rate due to test failures was higher in this experiment (70%) than in experiments 1–8 (43%). Also, this dropout rate was lower in experiments 1–4 (32%) than in experiments 5–8 (51%). We conducted experiments 1–4 between February and May in 2019, and experiments 5–9 between May and September in 2020, respectively before and during the global COVID-19 pandemic. Our speculative hypothesis for why the dropout rate was higher in the latter set of experiments is that there was an increase in inexperienced or inattentive AMT users and/or use of bots during the global lock down.

Monetary incentives

We incentivized participants' choices with money: the participants received monetary bonuses that depended on their and their co-players' choices in the games assigned to them (see screenshots of instructions and task, H and AI screen 6, available at osf.io/k4saq). In experiments 1–8, we did not incentivize the elicitation of participants' predictions (beliefs about their co-players' choices). Monetarily incentivizing belief reports in economic games can create hedging opportunities that may result in participants reporting their beliefs strategically instead of truthfully. For example, one may choose to cooperate in the Prisoner's Dilemma game thinking that their co-player would cooperate as well, but strategically report a belief that one's co-player would defect in order to hedge the risk of receiving 0 monetary bonus. Participants' use of hedging strategies similar to this one has been found in previous experiments with similar games. As a result, there is no consensus in the behavioral game theory research community at present on whether belief elicitation in the class of one-shot two-player games should be monetarily incentivized (Blanco et al., 2010; Schlag et al., 2015).

Not monetizing belief reports, however, was less of an issue for us, given that our goal was to compare participants' decisions across the human-human and human-AI treatments. As we used the same method to elicit participants' beliefs in both treatments, we can attribute differences in participants' predictions about their co-players' choices to the type of their co-player.

Simulation of waiting times between screens

After reading the instructions and successfully passing the comprehension test, participants in the human-human treatment entered a virtual lobby, in which they waited for another participant to join. Once paired

with another participant they proceeded to experiment proper to play the game assigned to them. This meant that, unless two participants entered the lobby at the exact same time, half of the participants had to wait a little before they were matched with a co-player (see screenshots of instructions and task, H screen 11, available at osf.io/k4saq). The other half of participants were immediately matched with a co-player (the one who was already waiting in the lobby) and proceeded to experiment proper within a few seconds of entering the lobby.

We simulated these waiting experiences for the participants in the human-AI treatments as well. Each participant who successfully completed the comprehension quiz in a human-AI treatment was randomly assigned to experience either a 2 or a 10 s delay before proceeding to experiment proper (see screenshots of instructions and task, AI screen 11, available at osf.io/k4saq).

QUANTIFICATION AND STATISTICAL ANALYSIS

We conducted the statistical analysis in R. For all results reported in the main text, we used Pearson's chi square tests for differences in proportions. For additional data analyses reported in supplemental information, we also used unpaired two-samples Wilcoxon tests for difference in mean reaction times and confidence, Cohen's Kappa, and performed Generalized Linear Models (GLMs). All statistical details and sample sizes are provided and explained in the text and figure legends.