nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-019-0113-5

Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation

Fatimah Ishowo-Oloko[1], Jean-François Bonnefon[2,3], Zakariyah Soroye[1], Jacob Crandall[4], Iyad Rahwan[3,5*] and Talal Rahwan[6*]

Recent advances in artificial intelligence and deep learning have made it possible for bots to pass as humans, as is the case with the recent Google Duplex—an automated voice assistant capable of generating realistic speech that can fool humans into thinking they are talking to another human. Such technologies have drawn sharp criticism due to their ethical implications, and have fueled a push towards transparency in human–machine interactions. Despite the legitimacy of these concerns, it remains unclear whether bots would compromise their efficiency by disclosing their true nature. Here, we conduct a behavioural experiment with participants playing a repeated prisoner's dilemma game with a human or a bot, after being given either true or false information about the nature of their associate. We find that bots do better than humans at inducing cooperation, but that disclosing their true nature negates this superior efficiency. Human participants do not recover from their prior bias against bots despite experiencing cooperative attitudes exhibited by bots over time. These results highlight the need to set standards for the efficiency cost we are willing to pay in order for machines to be transparent about their non-human nature.

Humans tend to trust algorithms less than they trust other humans[1]. In cooperative contexts, they break promises made to a computer more easily than promises made to a human[2], and they believe other humans to be more intelligent[3] and more cooperative[4] than artificial agents. This aversion to artificial intelligence as a social partner extends to other settings such as health-care[5–7] and forecasting[1]. One way for machines to bypass these prejudices is to conceal their true nature, that is, to passively let people think they are actually interacting with another human. Naturally, this requires machines to be sophisticated enough to pass as humans, but this hurdle is about to be overcome in various contexts. For example, Google Duplex is an automated voice assistant that can perform a variety of mundane phone-based tasks on behalf of its user, such as making dinner reservations and booking appointments. Duplex has crossed the 'uncanny valley'[8] by effectively passing as human. This was achieved by imitating human speech patterns including hesitations—ums and ahs—which a machine would ordinarily not do except to trick conversation partners into thinking they are interacting with another human. Accordingly, Duplex is able to have natural conversations with the people it calls on the phone, and to successfully complete bookings and transactions[9].

In spite or because of its impressive ability to mimic human speech, Duplex's technological breakthrough was marred by the ethical controversy it stirred[10,11]. The fact that Duplex could hide its true nature from humans was considered at least deceitful[12], and at most horrifying[13]. Consequently, some voices called for machines to be transparent about their true nature, and to disclose it upfront before any interaction with a human[14]. Given the uneasiness that humans display against bots in cooperative contexts, this push towards transparency raises a critical question: does transparency come at the expense of efficiency in human–bot interactions?

To address this question, we sought behavioural evidence for a transparency–efficiency tradeoff in the context of social dilemmas, where each 'player' can choose to either cooperate with, or defect against, the other player. We conducted an experiment in which participants played the canonical iterated prisoner's dilemma[15–24] with either a bot or a human, and we orthogonally manipulated the information that participants received about the nature of their associate—half of the participants were accurately informed about whether their partner was human or bot, while the other half received inaccurate information.

While this setup is far from addressing the psychological and cognitive subtleties involved in interacting with a complex system such as Google Duplex in a naturalistic environment, it allowed us to investigate whether bots can do better than humans at eliciting cooperation from their partner; to assess the prejudice humans have against cooperation partners they believe to be bots; and to investigate the extent to which this prejudice may nullify the ability of bots to elicit greater cooperation once they reveal their true nature.

Experimental design

We observed the behaviour of human participants in a repeated prisoner's dilemma, a well-established medium for studying and evaluating cooperative behaviour in many disciplines (for example, see refs. [2,15,25–27]). Each participant played at least 50 rounds of this game with either a bot or a human. The actions of the bots were decided by a reinforcement-learning algorithm called S++[28] (see Supplementary Note 5 for a brief overview of this algorithm). Among the numerous algorithms that can generate strategic decisions in repeated games (for example, see refs. [15,25,29–38]), we selected S++ because it outperforms other algorithms in simulations, and because it can learn effective behaviour within only a few rounds of interaction, making it particularly suitable for human–bot experiments where it is infeasible for participants to play thousands of rounds[39].

A total of 698 human participants were recruited through the crowd-sourcing platform MTurk and redirected to an external

[1]Department of Computer Science, Khalifa University, Abu Dhabi, United Arab Emirates. [2]Toulouse School of Economics (TSM-R), CNRS, University Toulouse Capitole, Toulouse, France. [3]Center for Humans and Machines, Max-Planck Institute for Human Development, Berlin, Germany. [4]Department of Computer Science, Brigham Young University, Provo, UT, USA. [5]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Department of Computer Science, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates. *e-mail: irahwan@mit.edu; talal.rahwan@nyu.edu

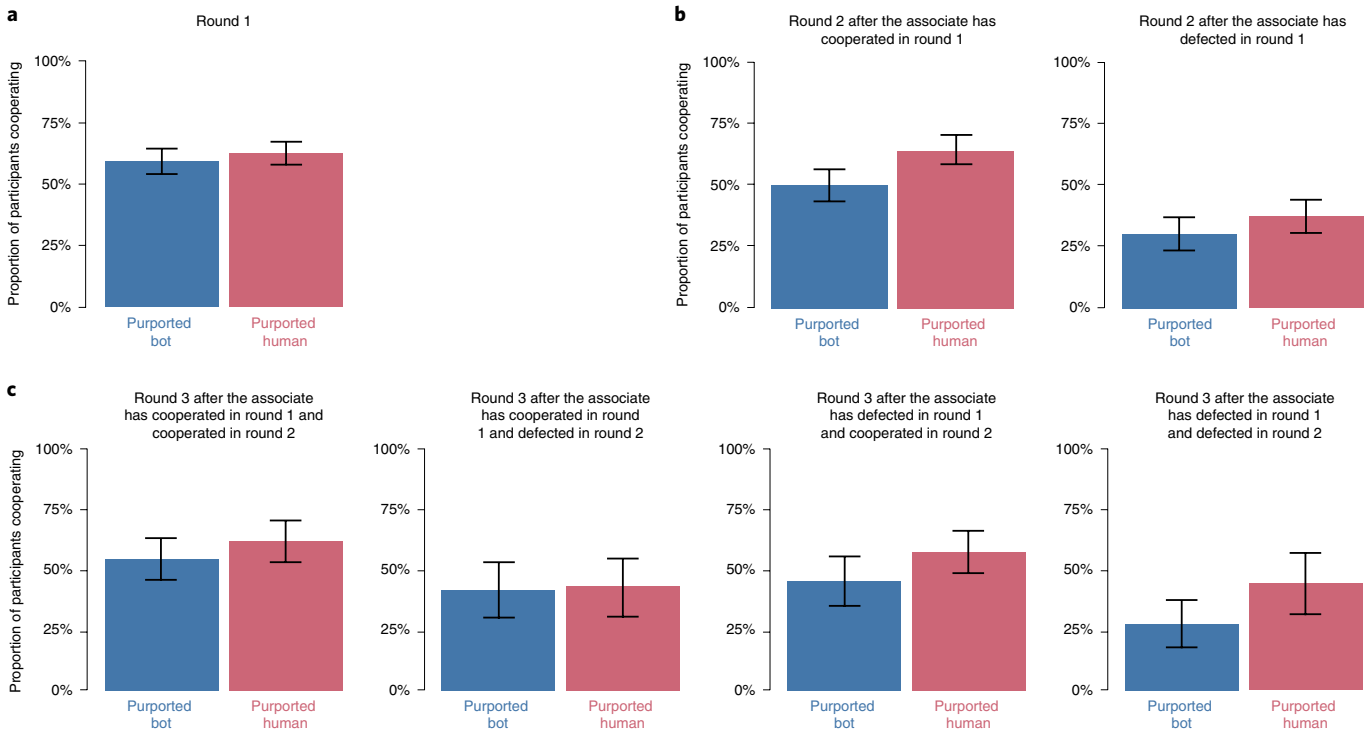

Fig. 1 | Prejudice against purported bots early in the game. a–c, Proportion of human participants who made a cooperative decision in round 1 (**a**), round 2 (**b**) and round 3 (**c**) as a function of the purported nature of their partner (error bars show the 95% confidence interval). Within each round, the participants are split according to the history of decisions made by their partner in previous rounds (there are two such histories for round 2 and four such histories for round 3). Participants are always more likely to cooperate with a purported human, regardless of their partner's decision history. As shown in Table 1, this effect is significant both in round 2 and in round 3.

website that was purpose-built for our experiment (for more details on the experimental setup, see Supplementary Notes 1–4 and Supplementary Figures). Of the 350 participants who played with another human, 170 were accurately informed that their partner was human and 180 were inaccurately informed that their partner was a bot. Likewise, of the 348 participants who played with a bot, 188 were accurately informed that their partner was a bot and 160 were inaccurately informed that their partner was human. Accordingly, the experiment followed a 2×2 design, in which participants were randomly assigned to one of four conditions: playing with a human they knew to be human, playing with a bot they knew to be a bot, playing with a human they believed to be a bot, or playing with a bot they believed to be human. Hereafter, we sometimes speak of participants who played with a 'purported bot' to designate participants who were told, accurately or not, that their partner was a bot; and likewise, we speak of participants playing with a 'purported human' to designate participants who were told, accurately or not, that their partner was human.

Results

Overall, participants who played with bots (whether they knew it or not) cooperated slightly more (46%) than participants who played with humans (41%). This is consistent with previous evidence showing that S++ can do at least as well as humans when it comes to eliciting cooperation from its partners. The algorithm achieves this by rewarding cooperation, tentatively forgiving lapses of cooperation and meting punishment in case of prolonged defection[39]. The key question we address in this Article, though, is whether humans are prejudiced against partners they believe to be bots, and whether this prejudice can hurt the performance of transparent bots.

To illustrate the prejudice against purported bots early in the game, Fig. 1 displays the proportion of cooperative decisions made by human players during rounds 1–3, for all possible sequences of decisions up to that round. In qualitative terms, human players were consistently less likely to cooperate with purported bots, regardless of the decisions made by their partner during previous rounds. To test the statistical significance of this result, we conducted a binomial regression for each of the three rounds, in which the dependent variable was the decision to cooperate, and the predictors were the purported nature of the partner as well as the number of cooperative decisions made by the partner during earlier rounds (this predictor was omitted for the round 1 regression). The regression tested whether the coefficient attached to each predictor was significantly different from zero. As shown in Table 1, the purported nature of the partner did not impact cooperation in the first round, but did so in rounds 2 and 3, regardless of the decisions that the partner made in earlier rounds.

So far, data suggest that actual bots, employing the S++ algorithm[28], can elicit cooperation to a greater extent than humans, but that humans cooperate less with purported bots. The question, then, is whether bots that are transparent about their true nature may be penalized to an extent that would offset their greater ability to elicit cooperation. To address this question, we must consider cooperation rates throughout the game in all four experimental variations. These data are shown in Fig. 2. Participants cooperated less when playing with purported bots than with purported humans through all 50 rounds of the game. Bots (Fig. 2b) did better than humans (Fig. 2a) at eliciting cooperation, mostly because human cooperation deteriorated, while bots managed to keep cooperation with humans constant. These results are confirmed by a multilevel binomial regression in which the dependent variable was the decision to cooperate and the predictors were the round number, the true nature of the partner (and its interaction with the round number) and the purported nature of the partner (and its interaction with the

Table 1 | Regression table showing likelihood of cooperation

	Round 1	Round 2	Round 3
Purported human	0.13 ± 0.15	0.47 ± 0.16**	0.40 ± 0.16*
Previous cooperation	-	0.98 ± 0.16***	0.63 ± 0.12***

Participants are always more likely to cooperate with a purported human, regardless of their partner's decision history. This effect is significant both in round 2 and in round 3 (*$P<0.05$; **$P<0.01$; ***$P<0.001$). Standard errors of the mean are given for each value.

round number), with a random intercept per participant and per game session. We tested whether the coefficients attached to each term were significantly different from zero. The model detected a significant effect of purported partner ($z=-3.5$, $P<0.001$), and a main effect of round number ($z=-8.2$, $P<0.001$), which was qualified by an interaction effect between round number and the true nature of the partner ($z=8.3$, $P<0.001$). No other effects were detected as significant.

The transparency–efficiency tradeoff is best perceived by comparing the red line in Fig. 2a (true humans known to be humans) to the red and blue lines in Fig. 2b. A bot passing as human (red line in Fig. 2b) is more efficient than a real human, mostly because humans are bad at maintaining cooperation in repeated games[18,40,41], whereas the programming of the bot allows it to keep its partner cooperating. But as soon as the bot reveals its true nature (blue line in Fig. 2b), it pays a large penalty that completely offsets its advantage and makes it less efficient than an actual human. After a large number of rounds, its performance ends up matching human performance, but this is only because human performance largely deteriorates with time, while the bot is able to maintain its mediocre performance throughout the game.

The bots used in our study learn to expect less from humans than from other bots, especially when they are transparent, as shown by changes in the 'aspiration level' of S++ over time. In more detail, the aspiration level is a parameter expressing the payoff that S++ expects to receive (see Supplementary Information for the mathematical details). As long as this expectation is met, S++ does not change its strategy. If the expectation is not met, S++ starts exploring other strategies. Furthermore, as this expectation decreases, S++ becomes less likely to attempt to arrive at a mutually cooperative solution. S++ starts with an optimistic aspiration level of 3, which corresponds to mutual cooperation. As shown in Fig. 3, on average, this aspiration level decreases over time as S++ interacts with people, and reaches even lower levels when S++ is being transparent about its nature. A linear regression of aspiration level on partner (human versus bot) and round detected significant effects of both predictors (partner: $t=-43.9$, $P<0.001$; round: $t=-63.4$, $P<0.001$). Another linear regression of aspiration level on transparency (opaque versus transparent) and round, restricted to human partners, detected significant effects of both predictors (transparency: $t=-14.9$, $P<0.001$; round: $t=-65.0$, $P<0.001$).

In sum, results offer clear behavioural evidence for an efficiency–transparency tradeoff in human–machine cooperation. Bots were better than humans at eliciting cooperation, but only if they were allowed to pass as humans. As soon as their true nature was revealed, cooperation rates dropped and could no longer match typical levels of human–human cooperation. The magnitude of this effect was about ten percentage points, which may lead to a substantial cumulative effect for bots that are used widely and routinely. While cooperation is not always or necessarily the best course of action (since it could theoretically lead to exploitation) we observed a substantial correlation between the cooperation rate of human players and their profits in the game, whether with other humans ($r=0.52$, $P<0.001$), or with bots ($r=0.58$, $P<0.001$).

Before we discuss whether people may decide to let bots hide their true nature for the sake of efficiency, we need to discuss one

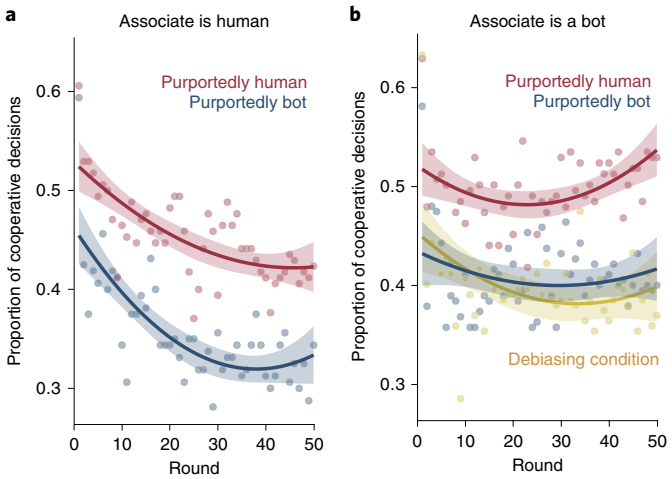

Fig. 2 | The tradeoff between efficiency and transparency. a,**b**, Proportion of cooperative decisions made by human participants, as a function of the purported nature of their partner, across the 50 rounds of the game for the cases when the partner is a human (**a**) or a bot (**b**). For better visualization, fitted lines display a quadratic model of the data, with the shaded area representing the 95% confidence interval. See text for details of the debiasing condition.

alternative to deception. What if bots disclosed their true nature but let people know that better results can be achieved if they are treated just like humans? Perhaps this simple intervention may restore cooperation to some degree, without the need for deception. We tried this intervention on 190 human participants, who were given the following information before the game: "Data suggest that people are better off if they treat the bot as if it were a human." Results in this debiasing condition are shown in Fig. 2b (yellow line). Participants in this condition behaved essentially the same as if they had not received the debiasing information, suggesting that simple debiasing cannot solve the transparency–efficiency tradeoff.

Discussion

Many voices have called for intelligent machines to be transparent, in the sense that their decisions might be explained in terms that would be understood by the people they affect[42–44]. But machines that interact or cooperate with humans can be transparent in a different sense, by disclosing their non-human nature upfront, before any interaction, even when their programming could allow them to convincingly pass as humans. While these situations are still rare, the Google Duplex example has been a warning call for many, by showing how close we are to a world where bots can conduct a discussion and close a transaction with humans, without ever revealing their non-human nature.

Although there is broad consensus that machines should be transparent about how they think, it is less clear whether they should always be transparent about who they are. To make an informed decision about this design choice, we need to gain a better understanding of the costs and benefits of transparency. In particular, we need to know whether the performance we expect from machines (for example, fluid and efficient cooperation) can be impaired when machines disclose their true nature to their human partners. Here we showed that transparency could hurt performance, to the extent that the superior efficiency of machines was nullified when they disclosed their non-human nature. It is important to note that this result is restricted to one form of transparency (that is, a disclosure about non-human nature), and one form of efficiency (that is, cooperation in a social dilemma). To generalize this result, future research will have to examine a broader range of transparency

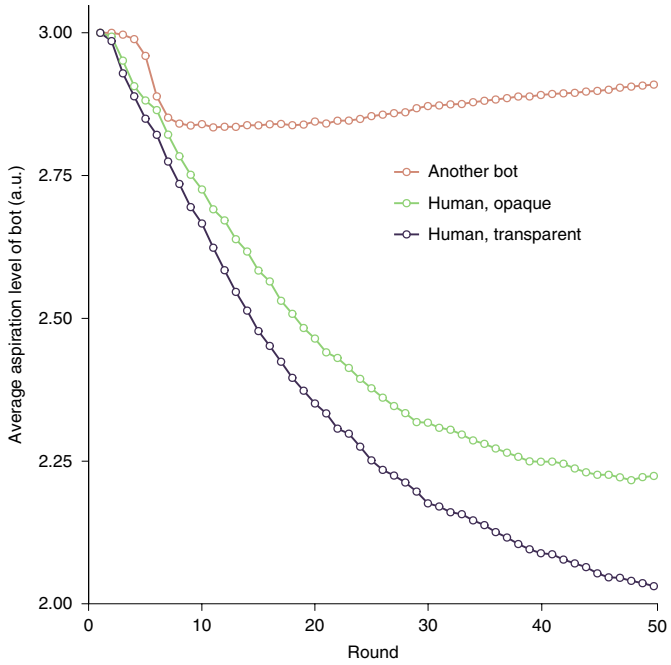

Fig. 3 | Bots learn to expect less from humans, especially when they are transparent. The aspiration level of the bot is the payoff it expects from its partner. This aspiration level decreases throughout the game as a result of defection by the partner. Since transparent bots experience greater defection, their aspiration levels are lower than those of opaque bots. Bots that play with other bots have much higher aspiration levels than bots that play with humans (aspiration levels for this case were generated through a simulation of 50 games of 50 rounds).

manipulations (for example, a description of the bots' learning abilities and prosocial tendency) as well as a broader range of efficiency benchmarks (for example, interaction speed or customer satisfaction). We used cooperation in a social dilemma as a proxy for efficiency, to capture situations where cooperation would lead to the best possible result, but can be compromised by a temptation not to cooperate, or a belief that the partner will not cooperate. Help desks operated by bots may provide a good example: while trusting the bot to help might lead to a quicker and easier resolution, humans may nevertheless decide to wait for human help due to a prejudice against the bot. However, one could imagine situations in which knowingly interacting with a bot might make things easier. For example, providing negative feedback about a product or a performance may be easier when talking to a bot, since it would eliminate the face-saving issues that complicate such an interaction between humans[45].

With these caveats, our results lead to the question of whether machines should be allowed to hide their non-human nature for the sake of efficiency. Ultimately, this choice must be made by the very people they interact with, otherwise it would violate fundamental values of autonomy, respect and dignity for humans in socio-technical systems. However, if people know that their interactions with transparent machines will be impaired, if they value the efficiency of these interactions and if they value it enough to accept being deceived, then they may consider it acceptable for machines to be opaque.

The difficulty, of course, is that this decision cannot be made on a case-by-case basis. Once one knows their partner is a machine, there is no un-knowing that fact: it would make no sense for a machine to ask its partner for the permission to pass as human. Accordingly, people must agree on a policy to let machines deceive them in some circumstances, without asking them for informed consent when it happens.

It remains to be seen whether such a policy might be ethically grounded and socially acceptable. It is important to note, though, that people sometimes find it acceptable, ethical and desirable to be blind to the individuals they deal with. In what is perhaps the most famous example of such a policy, major orchestras adopted a 'blind' audition process in which musicians play out of sight of the jury, in order to hide their identity, and most importantly their gender[46]. This policy was for the most part motivated by the desire to reduce gender discrimination, and it succeeded in that respect. But for orchestras, just as for companies, the objective of blind hiring is not only to increase diversity for diversity's sake: the goal is also to hire better individuals, who might have been rejected due to prejudice—in other words, to improve the efficiency of the hiring process, along with its fairness.

There is no need for humans to be more 'fair' to machines, whatever it would mean. Discrimination towards human groups is a serious problem, discrimination towards machines is not. However, being blind to the true nature of a machine may improve its cooperative efficiency, just as being blind to the identity of a candidate can improve the efficiency of the hiring process. If people agree, for efficiency purposes, to be blind to the individuals they seek to hire, then they may also agree to be blind to the machines they interact with, in return for more efficient cooperation. Opaque bots are still more ethically challenging than blind hiring, though. In the case of blind hiring, the pursuit of efficiency comes together with the pursuit of fairness: there is no salient conflict of ethical values. In the case of opaque bots, the pursuit of efficiency through non-transparency may well conflict with other values, such as respect and dignity. Our results highlight the need to reflect on the efficiency cost we are willing to pay in order to uphold these values in our interactions with machines.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study have been deposited in the Open Science Framework (https://doi.org/10.17605/OSF.IO/AK3TF).

Code availability

The software and all code used to generate the findings of this study have been deposited in the Open Science Framework (https://doi.org/10.17605/OSF.IO/AK3TF).

Received: 12 May 2019; Accepted: 8 October 2019;
Published online: 12 November 2019

References

1. Berkeley, J., Dietvorst, J. P. S. & Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol.* **144**, 114–126 (2015).
2. Kiesler, S., Sproull, L. & Miller, J. A prisoner's dilemma experiment on cooperation with people and human-like computers. *J. Pers. Soc. Psychol.* **70**, 47–65 (1996).
3. Oudah, M., Babushkin, V., Chenlinangjia, T. & Crandall, J. W. Learning to interact with a human partner. In *Proc. Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* 311–318 (ACM, 2015).
4. Merritt, T. & McGee, K. Protecting artificial team-mates: more seems like less. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* 2793–2802 (ACM, 2012).
5. Eastwood, J., Snook, B. & Luther, K. What people want from their professionals: attitudes toward decision-making strategies. *J. Behav. Decis. Mak.* **25**, 458–468 (2012).
6. Promberger, M. & Baron, J. Do patients trust computers? *J. Behav. Decis. Mak.* **19**, 455–468 (2006).
7. Neda Ratanawongsa, M. et al. Association between clinician computer use and communication with patients in safety-net clinics. *JAMA Intern. Med.* **176**, 125–128 (2016).

8. Mori, M., MacDorman, K. F. & Kageki, N. The uncanny valley [from the field]. *IEEE Robot. Autom. Mag.* **19**, 98–100 (2012).
9. Leviathan, Y. & Matias, Y. 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 (2018).
10. Statt, N. Google now says controversial AI voice calling system will identify itself to humans. *The Verge* https://www.theverge.com/2018/5/10/17342414/google-duplex-ai-assistant-voice-calling-identify-itself-update (2018).
11. Vomiero, J. Google's AI assistant must identify itself as a robot during phone calls: report. *Global News* https://globalnews.ca/news/4204648/googles-ai-identify-itself-robot-phone-calls/ (2018).
12. Hern, A. Google's 'deceitful' AI assistant to identify itself as a robot during calls. *The Guardian* https://www.theguardian.com/technology/2018/may/11/google-duplex-ai-identify-itself-as-robot-during-calls (2018).
13. Bergen, M. Google grapples with 'horrifying' reaction to uncanny AI tech. *Bloomberg* https://www.bloomberg.com/news/articles/2018-05-10/google-grapples-with-horrifying-reaction-to-uncanny-ai-tech (2018).
14. Harwell, D. A google program can pass as a human on the phone. should it be required to tell people it's a machine? *The Washington Post* https://www.washingtonpost.com/news/the-switch/wp/2018/05/08/a-google-program-can-pass-as-a-human-on-the-phone-should-it-be-required-to-tell-people-its-a-machine/ (2018).
15. Axelrod, R. *The Evolution of Cooperation* (Basic Books, 1984).
16. Nowak, M. A. & May, R. M. Evolutionary games and spatial chaos. *Nature* **359**, 826–829 (1992).
17. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
18. Fehr, E. & Gachter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
19. Ohtsuki, H., Hauert, C., Lieberman, E. & Nowak, M. A. A simple rule for the evolution of cooperation on graphs and social networks. *Nature* **441**, 502–505 (2006).
20. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
21. Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D. & Nowak, M. A. Positive interactions promote public cooperation. *Science* **325**, 1272–1275 (2009).
22. Fudenberg, D., Rand, D. G. & Dreber, A. Slow to anger and fast to forgive: cooperation in an uncertain world. *Am. Econ. Rev.* **102**, 720–749 (2012).
23. Dorrough, A. R. & Glöckner, A. Multinational investigation of cross-societal cooperation. *Proc. Natl Acad. Sci. USA* **113**, 10836–10841 (2016).
24. Bear, A. & Rand, D. G. Intuition, deliberation, and the evolution of cooperation. *Proc. Natl Acad. Sci. USA* **113**, 936–941 (2016).
25. Nowak, M. & Sigmund, K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* **364**, 56–58 (1993).
26. Bó, P. D. Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *Am. Econ. Rev.* **364**, 1591–1604 (2005).
27. Dreber, A., Rand, D. G., Fudenberg, D. & Nowak, Ma Winners don't punish. *Nature* **452**, 348–351 (2008).
28. Crandall, J. W. Towards minimizing disappointment in repeated games. *J. Artif. Intell. Res.* **49**, 111–142 (2014).
29. Fudenberg, D. & Levine, D. K. *The Theory of Learning in Games* (The MIT Press, 1998).
30. Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Proc. 11th International Conference on Machine Learning* 157–163 (ACM, 1994).
31. Auer, P., Cesa-Bianchi, N., Freund, Y. & Schapire, R. E. Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proc. 36th Symposium on the Foundations of Computer Science* 322–331 (IEEE, 1995).
32. Sandholm, T. W. & Crites, R. H. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems* **37**, 147–166 (1996).
33. Karandikar, R., Mookherjee, D., Ray, D. & Vega-Redondo, F. Evolving aspirations and cooperation. *J. Econ. Theory* **80**, 292–331 (1998).
34. Claus, C. & Boutilier, C. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proc. 15th National Conference on Artificial Intelligence* 746–752 (AAAI, 1998).
35. de Farias, D. & Megiddo, N. Exploration–exploitation tradeoffs for expert algorithms in reactive environments. In *Advances in Neural Information Processing Systems 17* (eds Saul, L. K. et al.) 409–416 (NIPS, 2004).
36. Bouzy, B. & Metivier, M. Multi-agent learning experiments in repeated matrix games. In *Proc. 27th International Conference on Machine Learning* 119–126 (Omnipress, 2010).
37. Iliopoulos, D., Hintze, A. & Adami, C. Critical dynamics in the evolution of stochastic strategies for the iterated prisoner's dilemma. *PLoS Comput. Biol.* **6**, e1000948 (2010).
38. Littman, M. L. & Stone, P. A polynomial-time Nash equilibrium algorithm for repeated games. *Decis. Support Syst.* **39**, 55–66 (2005).
39. Crandall, J. W. et al. Cooperating with machines. *Nat. Commun.* **9**, 233 (2018).
40. Chaudhuri, A. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Exp. Econ.* **14**, 47–83 (2011).
41. Wang, J., Suri, S. & Watts, D. J. Cooperation and assortativity with dynamic partner updating. *Proc. Natl Acad. Sci. USA* **109**, 14363–14368 (2012).
42. Citron, D. K. & Pasquale, F. The scored society: due process for automated predictions. *Wash. Law Rev.* **89**, 1 (2014).
43. Diakopoulos, N. Accountability in algorithmic decision making. *Commun. ACM* **59**, 56–62 (2016).
44. Selbst, A. D. & Barocas, S. The intuitive appeal of explainable machines. *Fordham Law Rev.* **87**, 1085 (2018).
45. Bonnefon, J. F., Feeney, A. & De Neys, W. The risk of polite misunderstandings. *Curr. Dir. Psychol. Sci.* **20**, 321–324 (2011).
46. Goldin, C. & Rouse, C. Orchestrating impartiality: the impact of 'blind' auditions on female musicians. *Am. Econ. Rev.* **90**, 715–741 (2000).

Acknowledgements

We thank E. Awad for his help running the experiments on MTurk. J.-F.B. acknowledges support from the ANR-Labex Institute for Advanced Study in Toulouse, the ANR-3IA Artificial and Natural Intelligence Toulouse Institute, and the grant ANR-17-EURE-0010 Investissements d'Avenir.

Author contributions

All authors conceived and designed the experiments. F.I.-O. and Z.S. conducted the experiments. F.I.-O. and J.-F.B. analysed the data and produced the figures and tables. F.I.-O., J.-F.B., J.C., I.R. and T.R. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-019-0113-5.

Correspondence and requests for materials should be addressed to I.R. or T.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Iyad Rahwan and Talal Rahwan

Last updated by author(s): 04/06/2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data was collected via a website that was purposely built for the study using node.js version 0.1, mongodb version 2.4 and npm version 1.2.x
Data analysis	The data was analysed using R version 3.3 (2016-05-03)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available from the corresponding author upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study was a 2*2 repeated measures design, with Associate and Information as Independent variable. The dependent variable (quantitative) was the proportion of human cooperation in each round.
Research sample	Our research sample were members of Amazon Mechanical Turk, with location restricted to members in the US and Canada.
Sampling strategy	Participants were matched or assigned to groups in a round-robin pattern. The first two participants to sign up were paired with each other while the third participant was paired with a bot; this pattern was repeated for all subsequent participants. The number of participants was predetermined using power analysis with 90% power.
Data collection	All data was collected online and so the researcher was not physically present with the participants.
Timing	Data was collected in 5 sessions: 02/12/2017; 14/12/2017; 18/12/2017; 05/01/2018 and 13/01/2018.
Data exclusions	After completion of the game, particpants were asked to recall the information about the associate given at the beginning. 10 participants failed the test and therefore. their data along with those of their associates were excluded. So a total of 20 data points were excluded.
Non-participation	56 participants dropped out/ declined participation but no reasons were given.
Randomization	Participants were assigned to groups in a round-robin pattern. The first two participants to sign up were paired with each other while the third participant was paired with a bot; this pattern was repeated for all subsequent participants.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology
☐	☐ Animals and other organisms
☐	☒ Human research participants
☐	☐ Clinical data

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*

Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Participants were recruited online via membership of Amazon Mechanical Turk
Ethics oversight	HSREC, Masdar Institute and COUHES, MIT.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s) — *Specify: functional, structural, diffusion, perfusion.*

Field strength — *Specify in Tesla*

Sequence & imaging parameters — *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition — *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software — *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization — *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template — *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal — *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring — *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings — *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested — *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016) — *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction — *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity — *Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis — *Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis — *Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*