Behavioural Evidence for a Transparency-Efficiency Tradeoff in Human-Machine Cooperation

Fatimah Ishowo-Oloko[a], Jean-François Bonnefon[b,d], Zakariyah Soroye[a], Jacob Crandall[c], Iyad Rahwan[d,e*], and Talal Rahwan[f,*]

[a]Department of Computer Science, Khalifa University, Abu Dhabi, UAE
[b]Toulouse School of Economics (TSM-R), CNRS, University Toulouse Capitole, Toulouse, France
[c]Computer Science Department, Brigham Young University, Provo, UT 84602, USA
[d]Center for Humans & Machines, Max-Planck Institute for Human Development, Lentzealle 94, Berlin 14195, Germany
[e]The Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[f]Computer Science, New York University Abu Dhabi, UAE.

[*]Joint corresponding authors. E-mail: irahwan@mit.edu; talal.rahwan@nyu.edu

Abstract

Recent advancements in artificial intelligence and deep learning 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 behavioral 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.

Introduction

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 AI as a social partner extends to other settings such as health-care [5, 6, 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, *um*s, and *ah*s, 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 to 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 toward transparency raises a critical question: *Does transparency come at the expense of efficiency in human-bot interactions?*

To address this question, we sought behavioral 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, 16, 17, 18, 19, 20, 21, 22, 23, 24] with either a bot or a human, and we orthogonally manipulated the information that participants received about the nature of their associate—so that half 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 behavior of human participants in a repeated prisoner's dilemma, a well-established medium for studying and evaluating cooperative behavior in many disciplines (e.g., [15, 25, 2, 26, 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 (e.g., [29, 15, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39]), we selected S++ because it outperforms other algorithms in simulations, and because it can learn effective behavior 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 [40].

A total of 698 human participants were recruited through the crowd-sourcing platform *MTurk* and re-directed to an external website that was purposely 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, and playing with a bot they believed to be human. In the rest of this article, 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 [40]. 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 in the early game, Figure 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

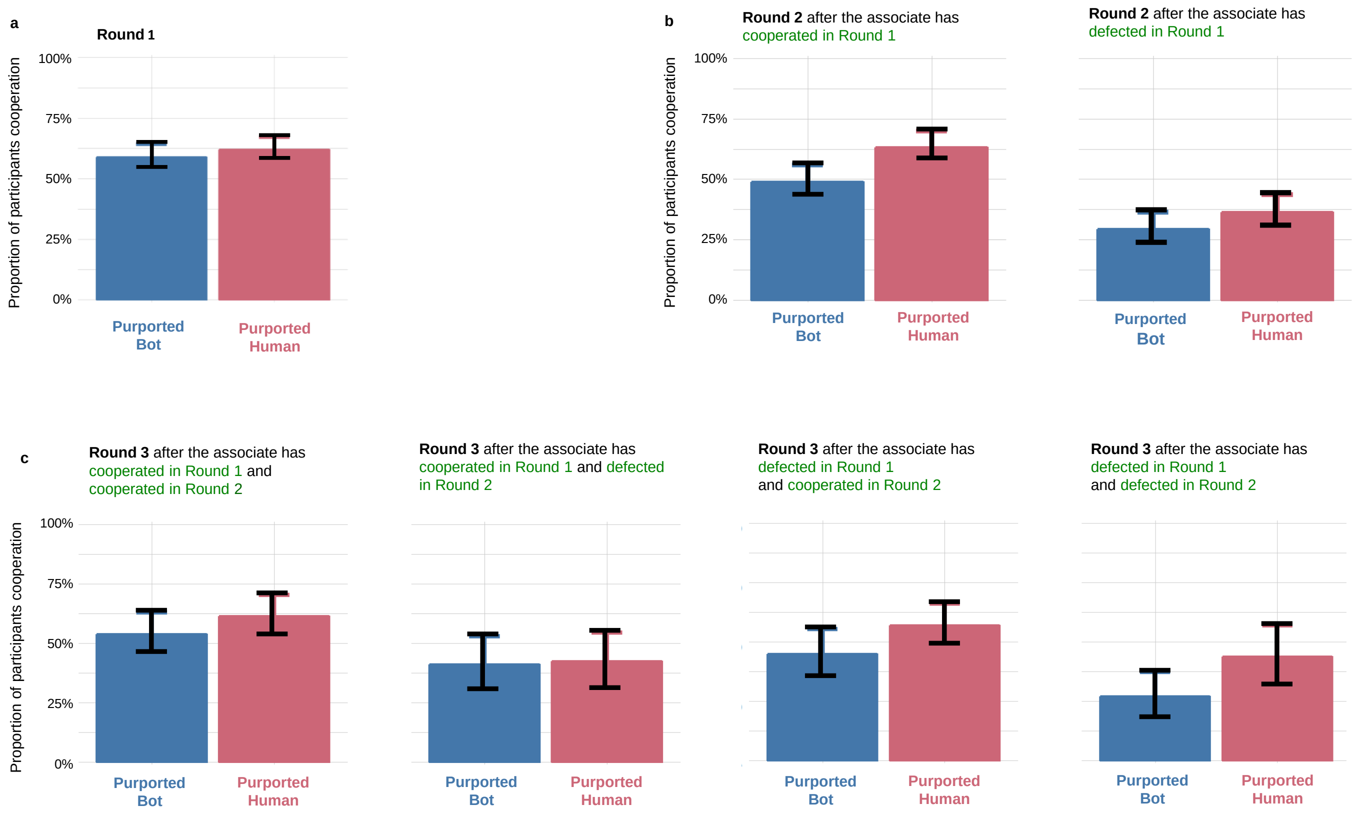

Figure 1: **Prejudice against purported bots in the early game** Proportion of human participants who made a cooperative decision in rounds 1–3, as a function of the purported nature of their partner (error bars show 95% confidence interval). Within each round, the subfigures split participants according 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 the regression table, this effect is significant both in round 2 and in round 3.

	Round 1	Round 2	Round 3
Purported Human	0.13	0.47^{**}	0.40^{*}
	(0.15)	(0.16)	(0.16)
Previous Cooperation		0.98^{***}	0.63^{***}
		(0.16)	(0.12)

Table 1: Participants are always more likely to cooperate with a purported human, regardless of their partner's decision history. As shown in the regression table, this effect is significant both in round 2 and in round 3. Key: $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$

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 which 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 all through the game, in all four experimental treatments. These data are shown in Figure 2. As can be seen, participants cooperated less when playing with purported bots (blue line) than with purported humans (red line), through all 50 rounds of the game. Real bots (left panel) did better than humans (right panel) at eliciting cooperation, mostly because human cooperation deteriorated through the game, 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 predictor were the round number, the true nature of the partner (and its interaction with round), the purported nature of the partner (and its interaction with round); 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 ($z = -3.5$, $p < .001$), and a main effect of round ($z = -8.2$, $p < .001$), which was qualified by an interaction effect between round and the true nature of the partner ($z = 8.3$, $p < .001$). No other effects were detected as significant.

The transparency-efficiency tradeoff is best perceived by comparing the red line in the right panel (true humans known to be humans) to the two lines in the left panel. A bot passing as human (red line, left panel) is more efficient than a real human, mostly because humans are bad at maintaining cooperation in repeated games [18, 41, 42], whereas the programming of the bot allows it to keep its partner cooperating.

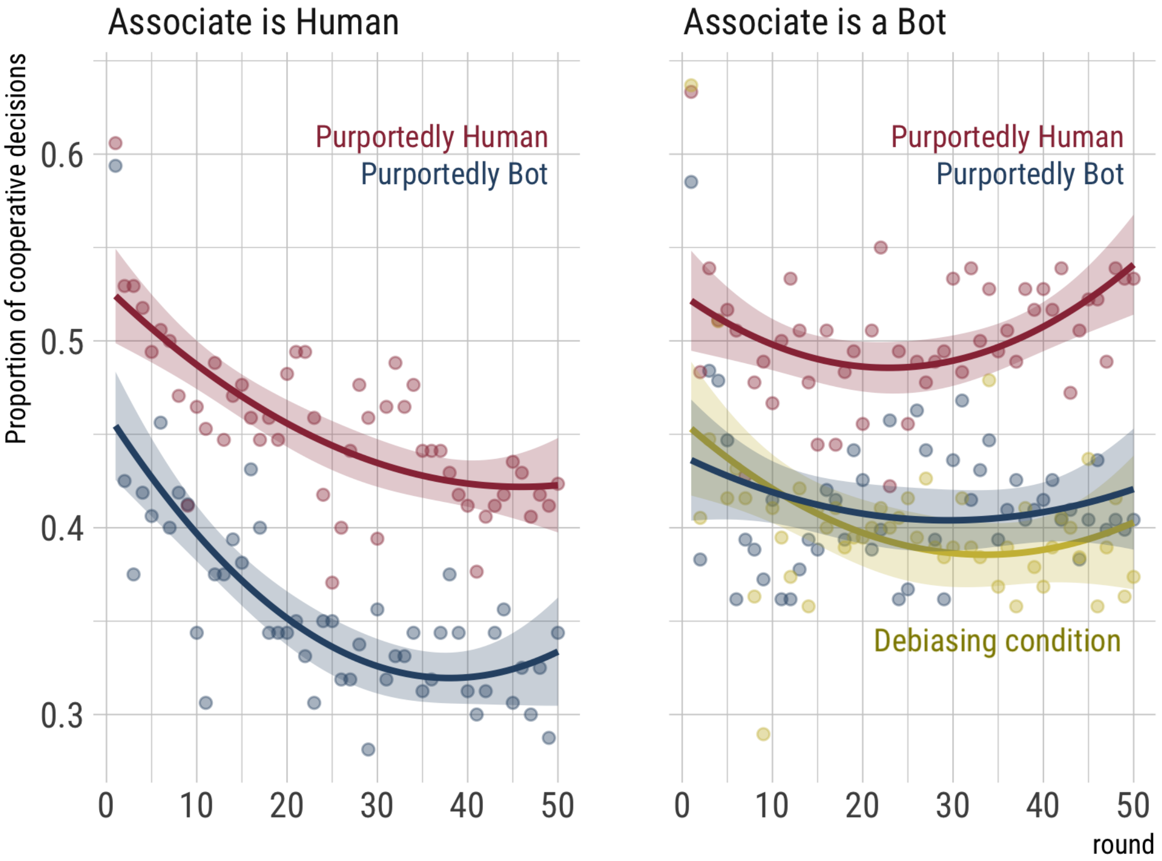

Figure 2: **The tradeoff between efficiency and transparency** Proportion of cooperative decisions made by human participants, as a function of the true and purported nature of their partner, across the 50 rounds of the game. For better visualization, fitted lines display a quadratic model of the data, with the shaded area representing the 95% confidence interval.

But as soon as the bot reveals its true nature (blue line, left panel), 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 SI for 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 Figure 3, on average, this aspiration level decreases over time as S++ interacts with people, and to even lower levels when S++ is being transparent about its nature. A linear regression of aspiration level on partner (human vs. bot) and round detected

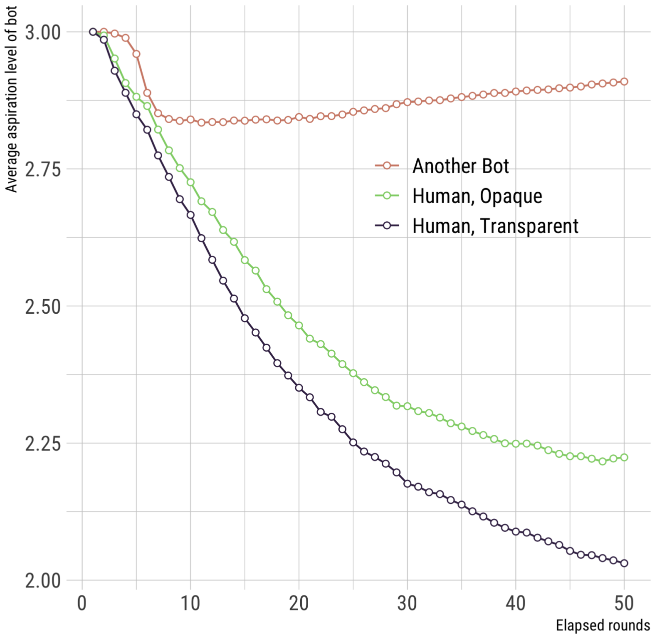

Figure 3: **Bots learn to expect less from humans especially when they are transparent** The aspiration level of the bot is the payoff it expects from its partner. This aspiration level decreases through the game as a result of defection from the partner. Since transparent bots experience greater defection,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

significant effects of both predictors (partner: $t = -43.9$, $p < .001$; round: $t = -63.4$, $p < .001$). Another linear regression of aspiration level on transparency (opaque vs. transparent) and round, restricted to human partners, detected significant effects of both predictors (transparency: $t = -14.9$, $p < .001$; round: $t = -65.0$, $p < .001$).

In sum, results offer clear behavioral 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 10 percentage points, which may lead to a substantial cumulative effect for bots who 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 ($r = .52$, $p < .001$), or with bots ($r = .58$, $p < .001$).

Before we discuss whether people may decide to let bots hide their true nature for the sake of efficiency, we need to discuss one 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 informed before the game that '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 Figure 2 (right panel). 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 [43, 44, 45]. But machines which 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 (e.g., a 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 (i.e., a disclosure about non-human nature), and one form of efficiency (i.e., cooperation in a social dilemma). To generalise this result, future research will have to examine a broader range of transparency manipulations (e.g., a description of the bots' learning abilities and prosocial tendency) as well as a broader range of efficiency benchmarks (e.g., 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 require and 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 [46].

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 to be 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 played out of sight of the jury, in order to hide their identity, and most importantly their gender [47]. 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 toward human groups is a serious problem, discrimination toward 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 a 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.

Acknowledgements

We thank Edmond Awad for his help running the experiments on mTurk. JFB 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

F.I-O, J-F.B, Z.S, J.C., I.R. and T.R. conceived of and designed the experiments. F.I-O and Z.S. conducted the running of the experiments. F.I-O and J-F.B analyzed 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.

Data Availability

The data that supports the findings of this study has been deposited in Open Science Framework (DOI: 10.17605/OSF.IO/AK3TF)

Code Availability

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

References

[1] Berkeley J. Dietvorst, J. P. S. & Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Experimental Psychology* **144**, 114–126 (2015).

[2] Kiesler, S., Sproull, L. & Miller, J. A prisoners dilemma experiment on cooperation with people and human-like computers. *Journal of Personality and Social Psychology* 47–65 (1996).

[3] Oudah, M., Babushkin, V., Chenlinangjia, T. & Crandall, J. W. Learning to interact with a human partner. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, 311–318 (ACM, New York, NY, USA, 2015).

[4] Merritt, T. & McGee, K. Protecting artificial team-mates: More seems like less. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, 2793–2802 (ACM, New York, NY, USA, 2012).

[5] Eastwood, J., Snook, B. & Luther, K. What people want from their professionals: Attitudes toward decision-making strategies. *Journal of Behavioral Decision Making* **25**, 458–468 (2012).

[6] Promberger, M. & Baron, J. Do patients trust computers? *Journal of Behavioral Decision Making* **19**, 455–468 (2006).

[7] Neda Ratanawongsa, M. *et al.* Association between clinician computer use and communication with patients in safety-net clinics. *JAMA Internal Medicine* **176**, 125–128 (2016).

[8] Mori, M., MacDorman, K. F. & Kageki, N. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* **19**, 98–100 (2012).

[9] Leviathan, Y. & Matias, Y. Google duplex: An ai system for accomplishing real-world tasks over the phone. `https://ai.googleblog.com` (2018). Accessed: 2018-10-09.

[10] Statt, N. Google now says controversial ai voice calling system will identify itself to humans. *The Verge* (2018).

[11] Vomiero, J. Google's ai assistant must identify itself as a robot during phone calls: report. *Global News* (2018).

[12] Hern, A. Google's 'deceitful' ai assistant to identify itself as a robot during calls. *The Guardian* (2018).

[13] Bergen, M. Google grapples with 'horrifying' reaction to uncanny ai tech. *Bloomberg* (2018).

[14] Harwell, D. A google program can pass as a human on the phone. should it be required to tell people its a machine? *Washington Post* (2018).

[15] Axelrod, R. *The Evolution of Cooperation* (Basic Books, New York, 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. *American Economic Review* **102**, 720–49 (2012).

[23] Dorrough, A. R. & Glckner, A. Multinational investigation of cross-societal cooperation. *Proceedings of the National Academy of Sciences* **113**, 10836–10841 (2016).

[24] Bear, A. & Rand, D. G. Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences* **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. *American Economic Review* **364**, 1591–1604 (2005).

[27] Dreber, A., Rand, D. G., Fudenberg, D. & Nowak, M. a. Winners don't punish. *Nature* **452**, 348–351 (2008).

[28] Crandall, J. W. Towards minimizing disappointment in repeated games. *Journal of Artificial Intelligence Research* **49**, 111–142 (2014).

[29] Fudenberg, D. & Levine, D. K. *The Theory of Learning in Games* (The MIT Press, 1998).

[30] 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 (1993).

[31] Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, 157–163 (1994).

[32] Auer, P., Cesa-Bianchi, N., Freund, Y. & Schapire, R. E. Gambling in a rigged casino: the adversarial multi-armed bandit problem. In *Proceedings of the 36th Symposium on the Foundations of Computer Science*, 322–331 (1995).

[33] Sandholm, T. W. & Crites, R. H. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems* **37**, 147–166 (1996).

[34] Karandikar, R., Mookherjee, D., R., D. & Vega-Redondo, F. Evolving aspirations and cooperation. *Journal of Economic Theory* **80**, 292–331 (1998).

[35] Claus, C. & Boutilier, C. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the 15th National Conference on Artificial Intelligence*, 746–752 (1998).

[36] de Farias, D. & Megiddo, N. Exploration–exploitation tradeoffs for expert algorithms in reactive environments. In *Advances in Neural Information Processing Systems 17*, 409–416 (2004).

[37] Bouzy, B. & Metivier, M. Multi-agent learning experiments in repeated matrix games. In *Proceedings of the 27th International Conference on Machine Learning*, 119–126 (2010).

[38] Iliopoulos, D., Hintze, A. & Adami, C. Critical dynamics in the evolution of stochastic strategies for the iterated prisoner's dilemma. *PLoS computational biology* **6**, e1000948 (2010).

[39] Littman, M. L. & Stone, P. A polynomial-time Nash equilibrium algorithm for repeated games. *Decision Support Systems* **39**, 55–66 (2005).

[40] Crandall, J. W. *et al.* Cooperating with machines. *Nature communications* **9**, 233 (2018).

[41] Chaudhuri, A. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics* **14**, 47–83 (2011).

[42] Wang, J., Suri, S. & Watts, D. J. Cooperation and assortativity with dynamic partner updating_SOM. *Proceedings of the National Academy of Sciences* **109**, 14363–14368 (2012).

[43] Citron, D. K. & Pasquale, F. The scored society: due process for automated predictions. *Wash. L. Rev.* **89**, 1 (2014).

[44] Diakopoulos, N. Accountability in algorithmic decision making. *Communications of the ACM* **59**, 56–62 (2016).

[45] Selbst, A. D. & Barocas, S. The intuitive appeal of explainable machines. *Fordham L. Rev.* **87**, 1085 (2018).

[46] Bonnefon, J. F., Feeney, A. & De Neys, W. The risk of polite misunderstandings. *Current Directions in Psychological Science* **20**, 321–324 (2011).

[47] Goldin, C. & Rouse, C. Orchestrating impartiality: The impact of 'blind' auditions on female musicians. *American Economic Review* **90**, 715 – 741 (2000).