COMMENTARY

Intelligent machines as social catalysts

Iyad Rahwan[a,1], Jacob W. Crandall[b,1], and Jean-François Bonnefon[c,1]

Some people excel at facilitating communication between other people. They are social catalysts: they promote engagement within a team, keep the team on point, defuse hostility, make sure that everyone contributes to the best of their capacity, and generally ensure that the team performs better than the sum of its members. Social catalysts need to be subtly aware of social dynamics, mood changes, face-saving strategies, and unspoken rules of communication. Given these requirements, it would seem that being a social catalyst is a quintessentially human role. The research of Traeger et al. (1), however, opens perspectives about the possibility of having machines rather than people perform the role of social catalysts in human groups.

Before we can discuss these perspectives, we need to ask the following questions: why and how would we want a machine to perform such a role? Sometimes, we want a machine to act as a social catalyst simply because there is no available human to perform that role or when we want the social catalyst to act on a nonhuman scale, simultaneously monitoring thousands of conversations. In these cases, we may adopt a better than nothing attitude. We program the machine to do as decent as possible an imitation of the behavior of a human social catalyst and accept that its performance will be somewhere below human level.

In other cases, we may want a machine to act as a social catalyst because the machine has decidedly nonhuman ways to perform the role, which work better than what humans might do. For example, the machine may be expert at learning patterns that maintain cooperation between humans or expert at detecting patterns that get humans stuck into suboptimal solutions, and it may act on this knowledge in a way that humans would not even consider—similar to the AlphaGo move with nonhumanness that stupefied grandmaster Lee Sedol (2).

As we will discuss shortly, these two approaches come with their own sets of technical and ethical challenges. However, the research by Traeger et al. (1) suggests that a third and unexpected approach is possible to using machines as social machines whose social actions are not as smooth as humans, but yet have more impact in changing human social behavior. In their research, Traeger et al. (1) show that a robot uttering "vulnerable" statements can promote engagement within a human team both in absolute terms (total speaking time increases) and in relative terms (speaking time is more equally distributed). The subjective experience of the team also improved and was rated as more positive and fun. The vulnerable statements made by the robot fell in three categories.

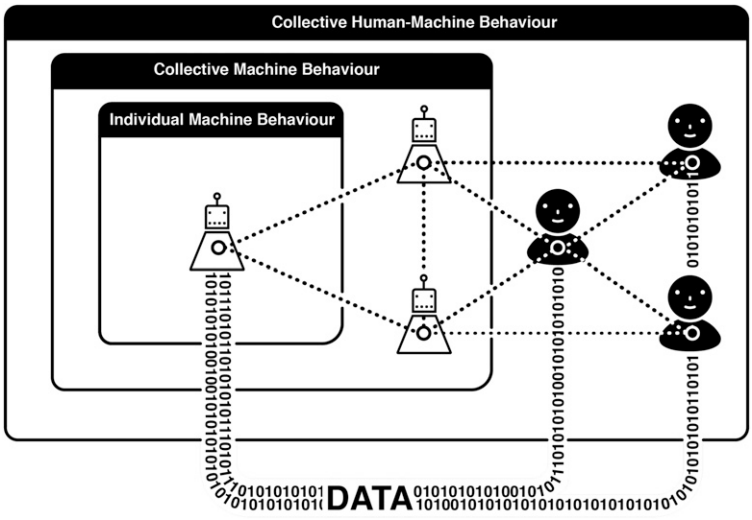

Fig. 1. Scales of modeling human and machine behavior.

- Self-disclosure, such as the following: "Great job, even though I sometimes doubt my abilities, I am glad I contributed to our team success this round." Such disclosures can feel strange when the robot mentions emotions or states of mind that it obviously cannot experience, such as doubt or happiness. Think, for example, of how a human would come across if they were to state these things in a totally deadpan, nonemotional voice.

[a]Center for Humans & Machines, Max Planck Institute for Human Development, Berlin 14195, Germany; [b]Computer Science Department, Brigham Young University, Provo, UT 84602; and [c]Toulouse School of Economics, CNRS, Université Toulouse Capitole, Toulouse 31000, France

Author contributions: I.R., J.W.C., and J.-F.B. wrote the paper.

The authors declare no competing interest.

See companion article, "Vulnerable robots positively shape human conversational dynamics in a human–robot team," 10.1073/pnas.1910402117.

[1]To whom correspondence may be addressed. Email: rahwan@mpib-berlin.mpg.de, crandall@cs.byu.edu, or jean-francois.bonnefon@tse-fr.eu.

First published March 19, 2020.

- Personal stories, such as "my soccer team went undefeated in the 2014 season." These stories are quite clearly made up, which is interesting because sharing (obviously) fabricated anecdotes would not be a great strategy for a human.
- Jokes, such as "why is the railroad angry? Because people are always crossing it!" As the authors mention, this joke is meant to be corny—in other words, the robot is intentionally programmed to make bad jokes or at least jokes that do not compare favorably with the ones a funny human would make.

In sum, a striking aspect of the results of Traeger et al. (1) is that the vulnerable robot is very successful at improving the conversational dynamics of the group, even though its techniques would be found wanting if they were used by a human. In fact, what we speculate is that the robot is successful because its techniques are intentionally not as good as that a human would employ.

Imagine a human who would systematically try to tell bad jokes and share fabricated personal anecdotes. This person would probably be perceived as phony and trying too hard and would make everyone uncomfortable. However, a robot doing the same things gets different results. First, it is not making people too uncomfortable because it is not a person whose face the group has to save. Second, it intentionally sets a low bar for the sophistication of one's contribution to the conversation, which may encourage the most self-conscious, shy humans to engage in this conversation. This would help explain both the absolute and relative effects observed in the experiment of Traeger et al. (1). If the humans who would have refrained from speaking are now encouraged to speak, human speaking time increases, and speaking time is more equally distributed among humans.

Technical and Ethical Challenges

A robot that is a better social catalyst by being less socially suave than a human offers new and fascinating perspectives. Consider the technical and ethical challenges raised by robots that attempt at being social catalysts by imitating human social behavior as closely as possible. First, this strategy requires the robot to be endowed with sophisticated abilities to navigate social contexts, emotional states, and face-saving strategies and more generally, to approximate the theory of mind-related reasoning that humans are so very good at. Although social robotics is certainly making strides in that direction, we know that this strategy comes with formidable challenges (3): for example, overcoming the uncanny valley (4). A robot that attempts to be a good social catalyst by being intentionally less socially sophisticated than a human bypasses at least some of these challenges.

Second, a robot with abilities and behavior that are designed to make it as human like as possible raises specific ethical concerns. As some ethicists argued, this humanizing strategy may amount to unethical deception (5) and may muddle our perception of who is really accountable for the robot's behavior (6). Similarly, some privacy concerns may arise when social robots are humanized to the extent that people form emotional bonds with them. In particular, these emotional bonds may prompt humans to interact in a more open or intimate way and can encourage them to share sensitive information that they would have otherwise kept private (7). In all of these cases, a robot with behavior that is a clear reminder that it is not trying to display human-level social skills can mitigate a variety of risks.

Consider now the technical and ethical challenges raised by robots that try to be social catalysts by behaving in a nonhuman way. First, they may succeed by using unconventional means at an unconventional scale. For example, robots may have access to vast amounts of data about us (beyond what would be required by a human-like approach), enabling it to personalize interventions to push people's right buttons. This is not only a technical challenge, but also a privacy challenge, echoing current debates in the field of human resources analytics (8) and Surveillance Capitalism more broadly (9). In other words, being nonhuman may entail having dangerous (superhuman) influence capabilities.

Traeger et al. show that a robot uttering "vulnerable" statements can promote engagement within a human team both in absolute terms (total speaking time increases) and in relative terms (speaking time is more equally distributed).

Second, the efficacy of this approach may require (when possible) hiding the fact that a machine is acting as a social catalyst to avoid defensive reactions from humans who may be concerned about being manipulated by an opaque algorithm. This creates a tension between the pursuit of performance and the ethical pursuit of transparency and personal autonomy (10). Third, to seek social catalyzation through uncharted, nonhuman means creates the risk of adverse indirect effects, which are hard to anticipate precisely because there might be little precedent for what the machine is attempting.

Here too, programming a robot to use techniques that are common to human social catalysts but intentionally not as smooth may mitigate all of these risks. This strategy requires less sensitive data when the robot performs its role (since it does not need a subtle understanding of its teammates), it is by design transparent (since it works precisely because it is performed by a robot rather than a human), and using well-charted human techniques narrows down the possibility of totally unexpected indirect effects.

Future Directions

There is much to be done in the wake of the research of Traeger et al. (1). Indeed, our interpretation of their findings is entirely speculative as are most of the perspectives we sketched in this comment. An important step would be to identify other contexts in which a robot, merely by being a robot, can put to good use techniques that would not be optimal for a human social catalyst. For example, there is research showing that a robot can helpfully call out hostile or offensive behavior directed by a human teammate to another (11). This is a delicate role when performed by a human because calling out the behavior of another can be perceived as judgmental and increase tension and hostility as a result, adversely affecting the social dynamics of the team. Accordingly, it requires subtle attention to emotional shifts and a sophisticated mastery of politeness protocols. It is fascinating to think that a robot may be able to bluntly perform that role without the need for subtle conversational moves precisely because a blunt call out can emphasize that the robot is a nonjudgmental machine.

Another area of great potential is interaction with children. Recent work showed that robots that interact with children as peers could promote growth mindsets in children, thus facilitating learning (12). Simple robots were also used in therapy for children with Autism Spectrum Disorder (13). There is evidence that, when interacting with such robots, the children directed more speech at the robot than humans (14). Robots may offer a new role that

facilitates all manners of learning and therapy in children precisely because they are not human like and thus, may not evoke negative feelings, such as fear of competition or disappointment.

Even more broadly, humans increasingly interact with machines in an ecology that contains other humans and other machines (Fig. 1). This means that we might not only engineer individual robots, but also machine collectives that are able to influence humans (and each other) on vast scales. This raises many questions about understanding the emergent behavior of these machines and the emergent hybrid human–machine system (15). The vulnerable robot of Traeger et al. (1), with its corny jokes and bogus anecdotes, puts a friendly face on this social order; however, even that robot may face its own ethical issues: for example, fairness. Inclusive and culturally sensitive social catalyzation arguably requires greater social sophistication. If we intentionally simplify the social capabilities of a machine for the purpose of greater average performance, we may inadvertently reduce its ability to recognize that different people may face different challenges or have different expectations when contributing to a team effort. This is not a criticism of the research of Traeger et al. (1), who never suggested downgrading the social capabilities of the robot teammate. It is merely a pointer for those who, like us, believe that this research is opening fascinating perspectives about when and why robots could paradoxically end up being better social catalysts by doing things that would not work well for humans.

Acknowledgments

J.-F.B. acknowledges support from the Agence National de la Recherche (ANR)-Labex Institute for Advanced Study in Toulouse and from the ANR Institut Interdisciplinaire d'Intelligence Artificielle Artificial and Natural Intelligence Toulouse Institute.

1 M. L. Traeger, S. S. Sebo, M. Jung, B. Scassellati, N. A. Christakis, Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 6370–6375 (2020).
2 C. Metz, In two moves, AlphaGo and Lee Sedol redefined the future. *Wired*, 16 March 2016. https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/. Accessed 10 February 2020.
3 E. Broadbent, Interactions with robots: The truths we reveal about ourselves. *Annu. Rev. Psychol.* **68**, 627–652 (2017).
4 M. Mori, The uncanny valley. *Energy* **7**, 33–35 (1970).
5 R. Sparrow, L. Sparrow, In the hands of machines? The future of aged care. *Minds Mach.* **16**, 141–161 (2006).
6 J. Bryson, "The moral, legal, and economic hazard of anthropomorphizing robots and AI" in *Envisioning Robots in Society–Power, Politics, and Public Space*, M. Coeckelbergh, J. Loh, M. Funk, J. Seibt, M. Nørskov, Eds. (Frontiers in Artificial Intelligence and Applications, IOS Press, Lansdale, PA, 2018), p. 11.
7 C. Lutz, M. Schöttler, C. P. Hoffmann, The privacy implications of social robots: Scoping review and expert interviews. *Mob. Media Commun.* **7**, 412–434 (2019).
8 J. H. Marler, J. W. Boudreau, An evidence-based review of HR Analytics. *Int. J. Hum. Resour. Manage.* **28**, 3–26 (2017).
9 S. Zuboff, Big other: Surveillance capitalism and the prospects of an information civilization. *J. Inf. Technol.* **30**, 75–89 (2015).
10 F. Ishowo-Oloko *et al.*, Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nat. Mach. Intell.* **1**, 517–521 (2019).
11 M. F. Jung, N. Martelaro, P. J. Hinds, "Using robots to moderate team conflict: The case of repairing violations" in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (ACM)* (Association for Computing Machinery, New York, NY, 2015), pp. 229–236.
12 H. W. Park, R. Rosenberg-Kima, M. Rosenberg, G. Gordon, C. Breazeal, "Growing growth mindset with a social robot peer" in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Association for Computing Machinery, New York, NY, 2017), pp. 137–145.
13 P. Pennisi *et al.*, Autism and social robotics: A systematic review. *Autism Res.* **9**, 165–183 (2016).
14 E. S. Kim *et al.*, Social robots as embedded reinforcers of social behavior in children with autism. *J. Autism Dev. Disord.* **43**, 1038–1049 (2013).
15 I. Rahwan *et al.*, Machine behaviour. *Nature* **568**, 477–486 (2019).