Deceiving possibilities

Robots are making a transition into human environments, where they can directly interact with us, in shops, hospitals, schools and more. Transparency about robots' capabilities and level of autonomy should be integrated into the design from the start.

In early September, a collection of robots took to the floor in the Milton Keynes shopping centre and wheeled around ordering coffee, taking the elevators and making deliveries. The robots — or the human teams operating them — competed in a new robotics challenge organized by the European Robotics League to test robots' abilities in smart city environments. Regular shoppers were excited and bemused by the robots' activities, some slightly wary that robots might be taking over jobs, while others at the same time were surprised to see that robots have so much difficulty with mundane tasks like opening doors. One component of the challenge was to test the robots' social abilities in specific scenarios such as asking for help when taking the elevator, and public responses were collected in a survey[1].

Researchers from the field of human–robot interaction have an important role to play to make sure we understand the complexities of human responses to robots and willingness to cooperate with robots in the short and long term. In mulling over findings from studies on human reactions to robots in our daily environments, a focus should be the importance of transparency. In particular, when encountering a robot, users should be able to quickly get a realistic idea of its purpose, capabilities and level of autonomy or teleoperation. Whereas the robots tested in Milton Keynes are intended to function autonomously, in practice robots in public environments are often at least partly controlled remotely by human operators, which is not always declared[2]. Humans tend to anthropomorphize and may find themselves assigning the robot its own goals and even a personality. To avoid damaging trust and ultimately our wilingness to cooperate with robots, transparency and clarity about robots' capabilities, level of autonomy and pre-programmed behaviour needs to be in place.

The issue of transparency and explicability is a main theme in the current ongoing debate about AI ethics[3]. Away from the field of robotics, disembodied AI is already making a substantial impact on our lives. For example, we interact regularly with human-like text- and speech-generating AI systems. Our need to anthropomorphize is so strong that we are even inclined to engage with these systems on a social level — "Alexa, tell me a joke" — but until recently nobody was really fooled into thinking that they are talking to a human. The demonstration of Google Duplex last year[4], a hyper-realistic voice assistant, opened up the possibility that this is exactly what could happen. Google quickly clarified that the system would identify itself to humans, although this may not rule out the potential for confusion.

This year GPT-2, a deep learning model for language generation developed by OpenAI, surprised the world with its remarkable ability to produce coherent passages of text that are difficult to distinguish from human written text. A recent *New Yorker* article about GPT-2 contains passages written by the programme[5]. The author mentions feeling 'spooked' about the experience, as GPT-2 began to make up quotes from OpenAI's Ilya Sutskever, who was interviewed for the article: "I worried that I'd forget what he really said, because the A.I. sounded so much like him, and that I'd inadvertently use in my article the machine's fake reporting, generated from my notes", he writes.

Most of us would like to know when we are dealing with a system or content in which AI is involved. But what kind of transparency do we want? Do we need to know what part of the AI is automated, what technology is incorporated into the product, what its learning capabilities are and what data have been used to train the system?

An Article in this issue explores an intriguing related question: in situations where humans and AI systems cooperate, does transparency come at a price? The authors invited humans to play cooperative games with opponents that were either human or an algorithm. They show that humans don't trust their opponent when they find out it's an algorithm, even if it plays more cooperatively than human players. This result is worth pondering, though naturally the conclusion is not that it's fine to conceal the fact we're interacting with an AI system just to improve efficiency. In a News & Views on the research article, Michael Rovatsos points out that transparency could mean more than revealing whether or not AI is involved: the participants might be given further information about the AI's learning capabilities and their ability to cooperate. It would be interesting to explore whether it is possible to regain humans' trust if they knew more about the design of the algorithm.

Human interaction and cooperation with AI and robots is likely to be beneficial when users are offered a better and more realistic idea of the systems they are dealing with: their autonomy, purpose and limitations. And perhaps even their capability to deceive us. ❐

Published online: 12 November 2019
https://doi.org/10.1038/s42256-019-0121-5

References

1. Wang, L., Iocchi, L., Marrella, A. & Nardi, D. in 28th *IEEE International Conference on Robot and Human Interactive Communication* (2019).
2. Davies, A. *Wired* https://www.wired.com/story/designated-driver-teleoperations-self-driving-cars/ (2019).
3. Floridi, F. et al. *Minds Mach* **28**, 689–707 (2018).
4. Metz, R. *MIT Technology Review* https://www.technologyreview.com/s/611539/google-demos-duplex-its-ai-that-sounds-exactly-like-a-very-weird-nice-human/ (2018).
5. Seabrook, J. *New Yorker* https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker (2019).