Humans are ready to take advantage of benevolent AI

June 8 2021

Autonomous bus, in Monheim, Rhine. Credit: IMAGO / Jochen Tack

Humans expect that AI is benevolent and trustworthy. A new study reveals that at the same time humans are unwilling to cooperate and compromise with machines. They even exploit them.

Picture yourself driving on a narrow road in the near future when suddenly another car emerges from a bend ahead. It is a self-driving car with no passengers inside. Will you push forth and assert your right of way, or give way to let it pass? At present, most of us behave kindly in such situations involving other humans. Will we show that same kindness towards autonomous vehicles?

Using methods from behavioral game theory, an international team of researchers at LMU Munich and the University of London have conducted large-scale online studies to see whether people would behave as cooperatively with artificial intelligence (AI) systems as they do with fellow humans.

Cooperation holds a society together. It often requires us to compromise with others and to accept the risk that they let us down. Traffic is a good example. We lose a bit of time when we let other people pass in front of us and are outraged when others fail to reciprocate our kindness. Will we do the same with machines?

The study which is published in the journal *iScience* found that, upon first encounter, people have the same level of trust toward AI as for human: most expect to meet someone who is ready to cooperate. The difference comes afterwards. People are much less ready to reciprocate with AI, and instead exploit its benevolence to their own benefit. Going back to the traffic example, a human driver would give way to another human but not to a self-driving car. The study identifies this unwillingness to compromise with machines as a new challenge to the future of human-AI interactions.

"We put people in the shoes of someone who interacts with an artificial agent for the first time, as it could happen on the road," explains Jurgis Karpus, Ph.D., a behavioral game theorist and a philosopher at LMU Munich and the first author of the study. "We modeled different types of

social encounters and found a consistent pattern. People expected artificial agents to be as cooperate as fellow humans. However, they did not return their benevolence as much and exploited the AI more than humans."

With perspectives from game theory, cognitive science, and philosophy, the researchers found that 'algorithm exploitation' is a robust phenomenon. They replicated their findings across nine experiments with nearly 2,000 human participants. Each experiment examines different kinds of social interactions and allows the human to decide whether to compromise and cooperate or act selfishly. Expectations of the other players were also measured. In a well-known game, the Prisoner's Dilemma, people must trust that the other characters will not let them down. They embraced risk with humans and AI alike, but betrayed the trust of the AI much more often, to gain more money.

"Cooperation is sustained by a mutual bet: I trust you will be kind to me, and you trust I will be kind to you. The biggest worry in our field is that people will not trust machines. But we show that they do!" notes Dr. Bahador Bahrami, a social neuroscientist at the LMU, and one of the senior researchers in the study. "They are fine with letting the machine down, though, and that is the big difference. People even do not report much guilt when they do," he adds.

Biased and unethical AI has made many headline—from the 2020 exams fiasco in the United Kingdom to justice systems—but this new research brings up a novel caution. The industry and legislators strive to ensure that artificial intelligence is benevolent. But benevolence may backfire. If people think that AI is programmed to be benevolent towards them, they will be less tempted to cooperate. Some of the accidents involving self-driving cars may already show real-life examples: drivers recognize an autonomous vehicle on the road, and expect it to give way. The self-driving vehicle meanwhile expects for normal compromises between

drivers to hold."

Algorithm exploitation has further consequences down the line. "If humans are reluctant to let a polite self-driving car join from a side road, should the self-driving car be less polite and more aggressive in order to be useful?" asks Jurgis Karpus.

"Benevolent and trustworthy AI is a buzzword that everyone is excited about. But fixing the AI is not the whole story. If we realize that the robot in front of us will be cooperative no matter what, we will use it to our selfish interest," says Professor Ophelia Deroy, a philosopher and senior author on the study, who also works with Norway's Peace Research Institute Oslo on the ethical implications of integrating autonomous robot soldiers along with human soldiers.

"Compromises are the oil that make society work. For each of us, it looks only like a small act of self-interest. For society as a whole, it could have much bigger repercussions. If no one lets autonomous cars join the traffic, they will create their own traffic jams on the side, and not make transport easier".

More information: Jurgis Karpus et al, Algorithm exploitation: humans are keen to exploit benevolent AI, *iScience* (2021). DOI: 10.1016/j.isci.2021.102679

Provided by Ludwig Maximilian University of Munich