Opinion: Fearing emotionally manipulative robots

By Colin Allen and Fritz Breithaupt

“Keep going straight here!”

“Err, that’s not what the app is telling me to do.”

“Yes, but it’s faster this way. The app is taking you to the beltway. Traffic is terrible there!”

“OK. I don’t know these roads.”

So went a conversation with an Uber driver in northern Virginia recently. But imagine it was a self-driving Uber. Would you even have that conversation, or would you be doomed to a frustrating 25 minutes on the beltway when you could have been home in 15?

And as your frustration mounts, will the AI driving the car recognize this—or appear to—and respond accordingly? Will customers prefer cars that seem to empathize?

Or imagine instead that you and your partner are arguing in the back seat over which route to take. How will you feel when your partner seems to be siding with the machine? Or the machine is siding with your partner?

Empathy is widely praised as a good thing. But it also has its dark sides: empathy can be manipulated and it leads people to unthinkingly take sides in conflicts. Add robots to this mix, and the potential for things to go wrong multiplies. Give robots the capacity to appear empathetic, and the potential for trouble is even greater.

To know why this is a problem, it helps to understand how empathy works in our daily lives. Many of our interactions

involve seeking empathy from others. People aim to elicit empathy because it’s taken as a proxy for rational support. For example, the guy in front of you at an auto repair shop tells the agent that he wants his money back: “The repair you did last month didn’t work out.” The agent replies: “I’m sorry, but this brake issue is an unrelated and new repair.” The argument continues, and the customer is getting angry. It seems like he might even punch the agent.

But instead, at this point, the customer and the agent might both look to you. Humans constantly recruit bystanders. Taking sides helps to settle things before they escalate. If it’s two against one, the one usually backs down. A lot of conflicts thereby get resolved without violence. (Compare chimpanzees, where fights often lead to serious injury.) Our tendency to make quick judgments and to take sides in conflicts among strangers is one of the key features of our species.

When we take sides, we assume the perspective of our chosen side—and from here it is a short step to develop emotional empathy. According to the three-person model of empathy introduced by Breithaupt, this is not entirely positive, because the dynamic of side-taking makes the first side we take stick, and we therefore assume that our side is right, and the other side is wrong. In this way, empathy accelerates divisions. Further, we typically view this empathy as an act of approval that extends to our consequent actions, including, for example, lashing back at the other side.

Now let’s imagine that the agent at the repair shop is a robot. The robot may appeal to you, a supposedly neutral third party, to help it to persuade the frustrated customer to accept the charge. It might say: “Please trust me, sir. I am a robot and programmed not to lie.”

Sounds harmless enough, does it? But suppose the robot has been programmed to learn about human interactions. It will pick up on social strategies that work for its purposes. It

may become very good at bystander recruitment. It knows how to get you to agree with its perspective and against the other customer's. The robot could even provide perfect cover for an unscrupulous garage owner who stands to make some extra money with unnecessary repairs.

You might be skeptical that humans would empathize with a robot. Social robotics has already begun to explore this question. And experiments suggest that children will side with robots against people when they perceive that the robots are being mistreated. In one study, a team of American and Japanese researchers carried out an experiment in which children played several rounds of a game with a robot. Later the game was interrupted by an overzealous confederate of the experimenters, who ordered the robot into a closet before the game was over. The robot complained and pleaded not to be sent into the closet before the game could be completed. The children indicated that they identified socially with the robot and against the experimenter.

We also know that when bystanders watch a robot and a person arguing, they may take the side of the robot and may start to develop something like empathy for the machine. We already have some anecdotal evidence for this effect from traffic-directing robots in Kinshasa. According to photojournalist Brian Sokol in the Guardian newspaper, "People on the streets apparently respect the robots … they don't follow directions from human traffic cops." Similarly, a study conducted at Harvard demonstrated that students were willing to help a robot enter secured residential areas simply because it asked to be let in, raising questions about the potential dangers posed by the human tendency to respect a request from a machine that needs help.

It is a relatively short step from robots that passively engage human empathy to robots that actively recruit bystanders. Robots will provoke empathy in situations of conflict. They will draw humans to their side and will learn

to pick up on the signals that work. Bystander support will then mean that robots can accomplish what they are programmed to accomplish—whether that is calming down customers, or redirecting attention, or marketing products, or isolating competitors. Or selling propaganda and manipulating opinions.

It would be naive to think that AI corporations will not make us guinea pigs in their experiments with developing human empathy for robots. (Humans are already guinea pigs in experiments being run by the manufacturers of self-driving cars.) The robots will not shed tears, but may use various strategies to make the other (human) side appear overtly emotional and irrational. This may also include deliberately infuriating the other side. Humans will become unwitting participants in an apparatus increasingly controlled by AI with the capacity to manipulate empathy. And suddenly, we will have empathy with robots, and find ourselves taking their sides against fellow human beings.

When people imagine empathy by machines, they often think about selfless robot-nurses and robot suicide helplines, or perhaps also robot sex. In all of these, machines seem to be in the service of the human. However, the hidden aspects of robot empathy are the commercial interests that will drive its development. Whose interests will dominate when learning machines can outwit not only their customers but also their owners?

Researchers now speculate about whether machines will learn genuine empathy. But that question is a distraction from the more immediate issue, which is that machines will not “feel” what humans feel, even if they get good at naming human emotions and responding to them. (At least for a while.) But in the near future, it doesn’t matter which emotions machines have. What is important is which emotions they can produce in humans, and how well they learn to master and manipulate these human responses. Instead of AI with empathy, we should be more concerned about humans having misplaced empathy with AI.

Colin Allen is a philosopher and cognitive scientist who has been teaching at Indiana University since 2004, but is moving to the University of Pittsburgh in fall 2017. Fritz Breithaupt is a humanities scholar and cognitive scientist at Indiana University. This essay is part of a Zócalo inquiry, Is Empathy the 20th Century's Most Powerful Invention?