What ethical models for autonomous vehicles don't address—and how they could be better

6 July 2020, by Matt Shipman

Photo credit: Denys Nevozhai.

There's a fairly large flaw in the way that programmers are currently addressing ethical concerns related to artificial intelligence (AI) and autonomous vehicles (AVs). Namely, existing approaches don't account for the fact that people might try to use the AVs to do something bad.

For example, let's say that there is an autonomous vehicle with no passengers and it is about to crash into a car containing five people. It can avoid the collision by swerving out of the road, but it would then hit a pedestrian.

Most discussions of ethics in this scenario focus on whether the autonomous vehicle's AI should be selfish (protecting the vehicle and its cargo) or utilitarian (choosing the action that harms the fewest people). But that either/or approach to ethics can raise problems of its own.

"Current approaches to ethics and autonomous vehicles are a dangerous oversimplification—moral judgment is more complex than that," says Veljko Dubljevi?, an assistant professor in the Science, Technology & Society (STS) program at North Carolina State University and author of a paper outlining this problem and a possible path forward. "For example, what if the five people in the car are terrorists? And what if they are deliberately taking advantage of the AI's programming to kill the nearby pedestrian or hurt other people? Then you might want the autonomous vehicle to hit the car with five passengers.

"In other words, the simplistic approach currently being used to address ethical considerations in AI and autonomous vehicles doesn't account for malicious intent. And it should."

As an alternative, Dubljevi? proposes using the so-called Agent-Deed-Consequence (ADC) model as a framework that AIs could use to make moral judgements. The ADC model judges the morality of a decision based on three variables.

First, is the agent's intent good or bad? Second, is the deed or action itself good or bad? Lastly, is the outcome or consequence good or bad? This approach allows for considerable nuance.

For example, most people would agree that running a red light is bad. But what if you run a red light in order to get out of the way of a speeding ambulance? And what if running the red light means that you avoided a collision with that ambulance?

"The ADC model would allow us to get closer to the flexibility and stability that we see in human moral judgment, but that does not yet exist in AI," says Dubljevi?. "Here's what I mean by stable and flexible. Human moral judgment is stable because most people would agree that lying is morally bad. But it's flexible because most people would also agree that people who lied to Nazis in order to

protect Jews were doing something morally good.

"But while the ADC model gives us a path forward, more research is needed," Dubljevi? says. "I have led experimental work on how both philosophers and lay people approach moral judgment, and the results were valuable. However, that work gave people information in writing. More studies of human moral judgment are needed that rely on more immediate means of communication, such as virtual reality, if we want to confirm our earlier findings and implement them in AVs. Also, vigorous testing with driving simulation studies should be done before any putatively 'ethical' AVs start sharing the road with humans on a regular basis. Vehicle terror attacks have, unfortunately, become more common, and we need to be sure that AV technology will not be misused for nefarious purposes."

The paper, "Toward Implementing the ADC Model of Moral Judgment in Autonomous Vehicles," is published in the journal *Science and Engineering Ethics*.

More information: Veljko Dubljevi?, Toward Implementing the ADC Model of Moral Judgment in Autonomous Vehicles, *Science and Engineering Ethics* (2020). DOI: 10.1007/s11948-020-00242-0