Self-driving cars: why we can't expect them to be 'moral'

24 January 2019, by John Mcdermid

Credit: Marat Marihal/Shutterstock

Ever since companies began developing self-driving cars, people have asked how designers will address the moral question of who a self-driving car should kill if a fatal crash is unavoidable. Recent research suggests this question may be even more difficult for car makers to answer than previously thought because the moral preferences people have vary so much between countries.

The researchers, based at Harvard University and MIT, developed an online game simulating situations where a fatal car accident was inevitable. They asked around 40m people from over 200 countries to choose between various accident outcomes, such as killing pedestrians rather than the car's passengers.

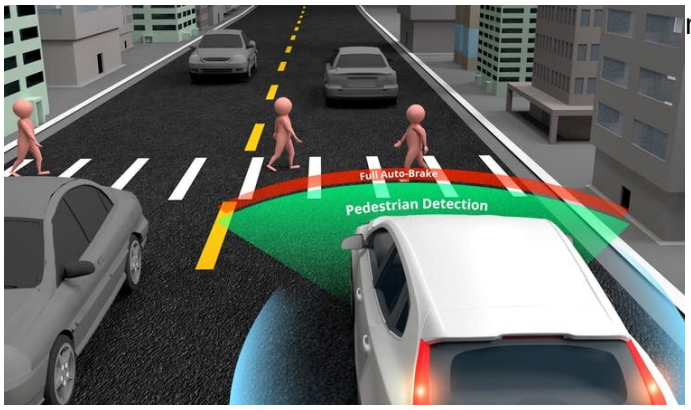
The results revealed three cultural clusters where there were significant differences in what ethical preferences people had. For example, in the Southern cluster (which included most of Latin America and some former French colonies), there was a strong preference for sparing women over men. The Eastern cluster (which included many Islamic countries as well as China, Japan and Korea) had a lower preference for sparing younger people over older people.

The researchers concluded by saying that this information should influence self-driving car developers. But is that really the case? While this paper highlights an interesting discovery about global variations in moral preferences, it also highlights a persistent misunderstanding about AI, and what it can actually do. Given the current AI technology used in self-driving cars, the idea that a vehicle could make a moral decision is actually impossible.

The 'moral machine' fantasy

Self-driving cars are trained to make decisions about where to steer or when to brake, using specific (or weak) AI (artificial intelligence that is focused on completing one narrow task). They're designed with a range of sensors, cameras and distance-measuring lasers (lidar), which give information to a central computer. The computer then uses the AI to analyse these inputs and make a decision.

Though the technology is currently relatively simple, cars will eventually outperform humans in these basic driving tasks. However, it's unrealistic to think that self-driving cars should also be capable of making an ethical decision that even the most moral of human beings wouldn't have time to make in an accident scenario. A car would need to be programmed with general AI if it were expected to do this.



Specific AI allows self-driving cars to make basic judgements about objects in its surroundings. Credit: Akarat Phasura/ Shutterstock

General AI is the equivalent of what makes us human. It's the ability to converse, enjoy music, find things funny or make moral judgements. Producing general AI is currently impossible because of the complexity of human thought and emotions. If we require moral autonomous vehicles we will not get there for several decades, if ever.

Another problem with the new research was that many of the situations that participants were asked to judge were unrealistic. In one scenario echoing the famous "trolley problem", participants were asked who the car should kill if its brakes failed: its three passengers (an adult man, an adult woman and a boy) or three elderly pedestrians (two men and one woman).

People can carefully consider these kinds of problems when answering a questionnaire. But in most real-life accidents, a driver wouldn't have time to make such judgements in the split second before it happens. This means the comparison is invalid. And given the current AI technology of self-driving cars, these vehicles won't be able to make these judgements either.

Current self-driving cars have sophisticated sensing abilities and can distinguish pedestrians from other objects, such as lamp posts or road signs. However, the research authors suggest that self-driving cars can, will and perhaps should be able to make more advanced distinctions. For example, they could recognise people deemed more desirable to society, such as doctors or athletes, and choose to save them in a crash scenario.

The reality is that designing cars to make such advanced judgements would involve producing general AI, which is currently impossible. There is also the question of whether this is even desirable. If it is ever possible to programme a car to decide which life should be saved, it is not something I believe we should allow. We should not allow the preferences identified in research, however large the sample size, to determine the value of a life.

In their most basic form, self-driving cars are being designed to avoid accidents if they can, and minimise speed at impact if they can't. Although, like humans, they aren't able to make a moral decision before an unavoidable accident. But self-driving cars will be safer than human drivers, as they're more attentive, can react quicker and will use braking system capabilities to the full in an accident scenario.

Currently, the biggest ethical challenge that self-driving car designers face is determining when there's enough evidence of safe behaviour from simulations and controlled on-road testing to introduce self-driving cars to the road. But this doesn't mean that they are "moral", or will be any time soon. To say that is to confuse the specific AI of doing driving tasks with general AI, which likely won't exist in our lifetime.

Ultimately, self-driving cars will be safer than humans. They will achieve this through design and avoiding accidents wherever possible, and reducing damage where not. However, the cars won't be able to make moral decisions that even we couldn't. This notion remains a far-fetched fantasy, and not one we should hold out hope for. Instead, let's concentrate on safety: something we will be able to have justified confidence in.

Provided by The Conversation