Why Self-Driving Cars Must Be Programmed to Kill

Self-driving cars are already cruising the streets. But before they can become widespread, carmakers must solve an impossible ethical dilemma of algorithmic morality.

October 22, 2015

When it comes to automotive technology, self-driving cars are all the rage. Standard features on many ordinary cars include intelligent cruise control, parallel parking programs, and even automatic overtaking—features that allow you to sit back, albeit a little uneasily, and let a computer do the driving.

So it'll come as no surprise that many car manufacturers are beginning to think about cars that take the driving out of your hands altogether (see "Drivers Push Tesla's Autopilot Beyond Its Abilities"). These cars will be safer, cleaner, and more fuel-efficient than their manual counterparts. And yet they can never be perfectly safe.

And that raises some difficult issues. How should the car be programmed to act in the event of an unavoidable accident? Should it minimize the loss of life, even if it means sacrificing the occupants, or should it protect the occupants at all costs? Should it choose between these extremes at random? (See also "How to Help Self-Driving Cars Make Ethical Decisions.")
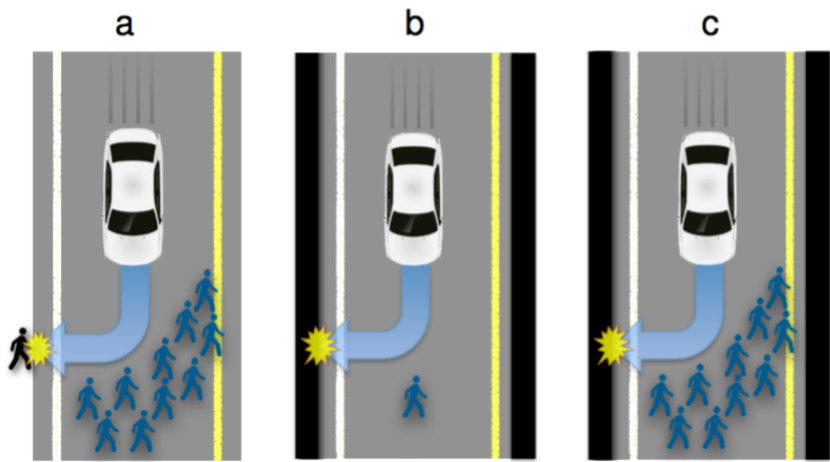
The answers to these ethical questions are important because they could have a big impact on the way self-driving cars are accepted in society. Who would buy a car programmed to sacrifice the owner?

So can science help? Today, we get an answer of sorts thanks to the work of Jean-Francois Bonnefon at the Toulouse School of Economics in France and a couple of pals. These guys say that even though there is no right or wrong answer to these questions, public opinion will play a strong role in how, or even whether, self-driving cars become widely accepted.

So they set out to discover the public's opinion using the new science of experimental ethics. This involves

posing ethical dilemmas to a large number of people to see how they respond. And the results make for interesting, if somewhat predictable, reading. "Our results provide but a first foray into the thorny issues raised by moral algorithms for autonomous vehicles," they say.

Here is the nature of the dilemma. Imagine that in the not-too-distant future, you own a self-driving car. One day, while you are driving along, an unfortunate set of events causes the car to head toward a crowd of 10 people crossing the road. It cannot stop in time but it can avoid killing 10 people by steering into a wall. However, this collision would kill you, the owner and occupant. What should it do?

One way to approach this kind of problem is to act in a way that minimizes the loss of life. By this way of thinking, killing one person is better than killing 10.

But that approach may have other consequences. If fewer people buy self-driving cars because they are programmed to sacrifice their owners,

then more people are likely to die because ordinary cars are involved in so many more accidents. The result is a Catch-22 situation.

Bonnefon and co are seeking to find a way through this ethical dilemma by gauging public opinion. Their idea is that the public is much more likely to go along with a scenario that aligns with their own views.

So these guys posed these kinds of ethical dilemmas to several hundred workers on Amazon's Mechanical Turk to find out what they thought. The participants were given scenarios in which one or more pedestrians could be saved if a car were to swerve into a barrier, killing its occupant or a pedestrian.

At the same time, the researchers varied some of the details such as the actual number of pedestrians that could be saved, whether the driver or an on-board computer made the decision to swerve and whether the participants were asked to imagine themselves as the occupant or an anonymous person.

The results are interesting, if predictable. In general, people are comfortable with the idea that self-driving vehicles should be programmed to minimize the death toll.

This utilitarian approach is certainly laudable but the participants were willing to go only so far. "[Participants] were not as confident that autonomous vehicles would be programmed that way in reality—and for a good reason: they actually wished others to cruise in utilitarian autonomous vehicles, more than they wanted to buy utilitarian autonomous vehicles themselves," conclude Bonnefon and co.

And therein lies the paradox. People are in favor of cars that sacrifice the occupant to save other lives—as long they don't have to drive one themselves.

Bonnefon and co are quick to point out that their work represents the first few steps into what is likely to be a fiendishly complex moral maze. Other issues that will need to be factored into future thinking are the nature of uncertainty and the assignment of blame.

Bonnefon and co say these issues raise many important questions: "Is it acceptable for an autonomous vehicle to avoid a motorcycle by swerving into a wall, considering that the probability of survival is greater for the passenger of the car, than for the rider of the motorcycle? Should different decisions be made when children are on board, since they both have a longer time ahead of them than adults, and had less agency in being in the car in the first place? If a manufacturer offers different versions of its moral algorithm, and a buyer knowingly chose one of them, is the buyer to blame for the harmful consequences of the algorithm's decisions?"

These problems cannot be ignored, say the team: "As we are about to endow millions of vehicles with autonomy, taking algorithmic morality seriously has never been more urgent."

A roundup of the week's most interesting papers from the arXiv:

Quantum Centipedes: Collective Dynamics of Interacting Quantum Walkers

Birdsong Dialect Patterns Explained Using Magnetic Domains

Quantifying the Global Atmospheric Power Budget

Revealing the Hidden Patterns of News Photos: Analysis of Millions of News Photos Using GDELT and Deep Learning-based Vision APIs

Computing

The Best of the Physics arXiv (week ending March 26, 2016)

A selection of the most interesting and intriguing arXiv papers from the past week.