The Moral Science Behind Self-Driving Cars

July 07, 2016

In 2016, Joshua Brown of Canton, Ohio became the first known fatality of a self-driving car. Brown was killed in May when his Tesla Model S crashed into a tractor trailer while engaged in a self-driving Autopilot mode. Although Tesla's Autopilot mode allows its vehicles to automatically apply brakes, steer, and change lanes, the company has insisted that this technology is not meant to replace a human driver—at least not quite yet.

"When used in conjunction with driver oversight, the data is unequivocal that Autopilot reduces driver workload and results in a statistically significant improvement in safety," Tesla said in a statement.

And they have a good point — according to Tesla, their cars have logged 130 million miles on the road using Autopilot with just a single fatality. By comparison, human drivers averaged 1.08 crash deaths per 100 million miles driven in 2014, according to the most recent data from the US Department of Transportation.

Of course, no matter how safe the technology behind autonomous vehicles becomes, people will still be killed in car crashes. One of the stickiest problems facing car manufacturers will be programming the algorithms that govern the safety functions of autonomous vehicles: Should your self-driving car purposely sacrifice a passenger if it means saving two pedestrians? Or should these vehicles be programmed to protect their passengers at any cost?

An interdisciplinary team of researchers has started conducting experiments to learn more about how people might react to the moral quandaries posed by self-driving cars. Psychological scientists Jean-François Bonnefon (University of Toulouse Capitole) and Azim Shariff (University of Oregon) teamed up with MIT computer scientist Iyad Rahwan on a series of experiments examining the moral dilemmas posed by autonomous vehicles.

“Autonomous vehicles (AVs) should reduce traffic accidents, but they will sometimes have to choose between two evils, such as running over pedestrians or sacrificing themselves and their passenger to save the pedestrians,” the researchers write in *Science*.

Across six online surveys including nearly 2,000 participants, the researchers found that people had somewhat contradictory feelings about how autonomous vehicles (AVs) should handle moral dilemmas.

“Although people tend to agree that everyone would be better off if AVs were utilitarian (in the sense of minimizing the number of casualties on the road), these same people have a personal incentive to ride in AVs that will protect them at all costs,” Bonnefon and colleagues explain.

That is, people thought that self-driving cars should be programmed to hurt the fewest possible people: An AV with a single passenger should swerve off the road to avoid hitting a group of 10 pedestrians. But, at the same time, people weren’t very keen to actually use a car that might be programmed to kill them.

Participants were given a scale from 0 (protect the passenger at all costs) to 100 (minimize the total number of deaths) and asked to rate a series of scenarios. The majority of participants preferred utilitarian programming; in one study, 76% of participants thought the more moral choice was for a self-driving car to sacrifice its single passenger’s life rather than kill 10 pedestrians.

However, when people were asked whether they would purchase or ride in a vehicle with this kind of utilitarian programming themselves, approval of the algorithm dropped significantly.

“In other words, even though participants still agreed that utilitarian AVs were the most moral, they preferred the self-protective model for themselves,” the researchers point out. “This is the classic signature of a social dilemma, in which everyone has a temptation to free-ride instead of adopting the behavior that would lead to the best global outcome.”

Paradoxically, the researchers argue, people’s hesitance to use AVs with utilitarian programming may actually increase unnecessary road deaths by postponing the widespread adoption of safer technology.

“For the time being, there seems to be no easy way to design algorithms that would reconcile moral values and personal self-interest—let alone account for different cultures with various moral attitudes

regarding life-life trade-offs—but public opinion and social pressure may very well shift as this conversation progresses," the researchers conclude.

Reference

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573-1576. doi: 10.1126/science.aaf2654