Are We Programming Killer Cars? The Ethics of Autonomous Vehicles

Self-driving cars have become technologically feasible. The question now is: are they ethically feasible?

By Ariella Brown

June 27, 2019

Tesla Autopilot**Tesla/YouTube**

Over the past several years, more and more autonomous features have been embedded in cars. And just a couple of months back, Tesla released the following video in which it boasted about having achieved “Full Self-Driving.”

A Techopedia article reported that even earlier Tesla cars contained "the necessary hardware for autonomous driving," though activating the ability depended on a software update. The article also envisioned the difference between the way autonomous cars built today will differ from the ones in the future.

Currently, Tesla cars are equipped with the necessary hardware for autonomous driving, but software updates are required to fully enable the feature. While it will allow fully autonomous driving, it will also still allow the human driver to take control when the situation calls for intervention.

The next generation of autonomous vehicles, however, would not need steering wheels, pedals or transmissions. The advantage of such cars is the possibility of reducing accidents and providing necessary transportation for people who are incapable of driving like the elderly or those with vision or physical disabilities.

But there is also a potential downside: the necessity for the human agency that sets up the car's programming to foresee all possible scenarios and to direct the car to make the kind of judgements people have to when the scenario calls for action that will inevitably cause some form of harm.

While Tesla may be the most famous name on the AI front for vehicles, it certainly is not the only player in this growing market. Some far more venerable names in the industry have also gotten into the act.

Bernard Marr recently wrote about Toyota’s billion dollar investment in self-driving cars and AI. The company has set goals for itself that it wants to reach by 2020:

> “Through Toyota’s investments in tech start-ups such as Perceptive Automata it hopes to create the technology to allow autonomous vehicles more human-like intuition when they are on the road more similar to how human drivers interact with pedestrians.”

Self-driving safety track record

Of course, we're not there yet. But the question is if that is the end goal and if it is something that we should pursue without full considerations of the ramifications of a fully independent car.

Every Single Self-Driving Car Accident And Death lists nine accidents involving autonomous vehicles, only four of which caused fatalities. Despite the claims of the title, though, the list is incomplete, as there have been fatalities from such accidents after the article was published.

The last fatality it reported was the one involving a Tesla Model X on March 23, 2018. The driver of the car died when it hit a highway barrier. Tesla blamed it on the barrier's interference with the autonomous driving system of the vehicle:

"The reason this crash was so severe is because the crash attenuator, a highway safety barrier which is designed to reduce the impact into a concrete lane divider, had been crushed in a prior accident without being replaced," Tesla said in its statement.

The company added: "We have never seen this level of damage to a Model X in any other crash."

Unfortunately, though, that was not the end of fatal crashes for Tesla's self-driving cars. A number of them occurred this year.

Among the incidents was one on March 1, 2019. It's been confirmed by the US National Transportation Safety Board (NTSB) that the semi-autonomous Autopilot software was engaged on a Tesla Model 3 when it slammed into a tractor-trailer attempting to cross a Florida highway and the car driver was killed.

Though they are still relatively rare, compared to the car accidents caused by human drivers, the fact that there are any accidents and fatalities caused by self-driving cars have made people concerned about their safety and programming. In fact, this year Quartz cast some doubt about Tesla's safety claims.

Like that Tesla accident, most autonomous car accidents result in the death of the person sitting in the driver's seat. However, there have been cases of people outside the car struck and killed by autonomous cars.

The most infamous incident of that sort may be the one involving Uber in the March 2018 death of Elaine Herzberg. The 49-year-old woman was walking and pushing her bicycle across the Mille Avenue in Tempe, Arizona when the Uber car struck her.

You can see the video of the incident released by the police here:

As a result of that, Uber adopted a policy of making sure to include human drivers in its cars. The story was reported here: Uber Puts Self-Driving Cars Back to Work but With Human Drivers.

This is a way for Uber to circumvent the problem that we will have to confront, if and when fully autonomous cars become the norm: how to program them to incorporate the instinct to preserve human life.

Programming AI with concern for ethics

As we saw in another article, Our Brave New World: Why the Advance of AI Raises Ethical Concerns, with the great power of AI comes great responsibility, to ascertain that technology does not end up making situations worse in the name of progress. The study of ethics for AI has captured the attention of people who think about what needs to be done ahead of implementing automated solutions.

One of those people is Paul Thagard, Ph.D., a Canadian philosopher, and cognitive scientist brought up some of the issues we have to now confront with respect to programming ethics into AI in How to Build Ethical Artificial Intelligence.

He raises the following 3 obstacles:

1. **Ethical theories are highly controversial. Some people prefer ethical principles established by religious texts such as the Bible or the Quran. Philosophers argue about whether ethics should be based on rights and duties, on the greatest good for the greatest number of people, or on acting virtuously.**
2. **Acting ethically requires satisfying moral values, but there is no agreement about which values are appropriate or even about what values are. Without an account of the appropriate values that people use when they act ethically, it is impossible to align the values of AI systems with those of humans.**
3. **To build an AI system that behaves ethically, ideas about values and right and wrong need to be made sufficiently precise that they can be implemented in algorithms, but precision and algorithms are sorely lacking in current ethical deliberations.**

Thagard does offer an approach to overcome those challenges, he says, and references his book, *Natural Philosophy: From Social Brains to Knowledge, Reality, Morality, and Beauty*. However, in the course of the article, he does not offer a solution that specifically addresses self-driving car programming.

Self-driving cars and the Trolley Problem

Ideally, drivers avoid hitting anything or anyone. But it is possible to find oneself in a situation in which it is impossible to avoid a collision, and the only choice is which person or people to hit.

This ethical dilemma is what is known as the Trolley Problem, which, like the trolley itself, goes back over a century. It's generally presented as follows:

You see a runaway trolley moving toward five tied-up (or otherwise incapacitated) people lying on the tracks. You are standing next to a lever that controls a switch. If you pull the lever, the trolley will be redirected onto a side track, and the five people on the main track will be saved. However, there is a single person lying on the side track.

You have two options:

1. **Do nothing and allow the trolley to kill the five people on the main track;**
2. **Pull the lever, diverting the trolley onto the side track where it will kill one person.**

Of course, there is no really good choice here. The question is which one is the lesser of two bad options. It was just this kind of a dilemma that the Green Goblin presented Spiderman in the 2002 movie, attempting to force him to choose between rescuing a cable-car full of children or the woman he loves:

Being a superhero, Spiderman was able to use his web-spinning abilities and strength to save both. But sometimes even superheroes have to make a tragic choice, as was the case in the 2008 film *The Dark Knight* in which Batman's choice was to leave the woman he loved in the building that exploded.

So even those who have superior abilities cannot always save everyone, and the same situation can apply to AI-enabled cars.

The question then is: Which code of ethics do we apply to program them to make such choices?

What should the self-driving car do?

MIT Technology Review drew attention to some researchers working on formulating the answers a few years ago in How to Help Self-Driving Cars Make Ethical Decisions. Among the researchers in the field is Chris Gerdes, a professor at Stanford University who has been looking into "the ethical dilemmas that may arise when vehicle self-driving is deployed in the real world."

He offered a simpler choice: that of dealing with a child running into the street, which forces the car to hit something but allows it to choose between the child and a van on the road. For a human that should be a no-brainer that protecting the child is more important than protecting the van or the autonomous car itself.

But what would the AI think? And what about the passengers in the vehicle who may end up sustaining some injuries from such a collision?

Gerdes observed, "These are very tough decisions that those that design control algorithms for automated vehicles face every day."

The article also quotes Adriano Alessandrini, a researcher working on automated vehicles at the University de Roma La Sapienza, in Italy who has served as head for the Italian portion of the European-based CityMobil2 project to test automated transit vehicle. See the video about it below:

She encapsulated the Trolley problem for drivers and self-driving cars in this summation:

> **“You might see something in your path, and you decide to change lanes, and as you do, something else is in that lane. So this is an ethical dilemma.”**

Another prominent expert in the field is Patrick Lin, a professor of philosophy at Cal Poly, that Geerdes has worked with. Lin's TED-Ed take on the ethical problems in programming self-driving cars to make life or death decisions, is presented as a thought experiment in this video:

If we were driving that boxed in car in manual mode, whichever way we'd react would be understood as just that, a reaction, not a deliberate decision," Lin says in the video. Accordingly, it would be understood to be "an instinctual panicked move with no forethought or malice."

The very real possibility of deaths occurring as a result not of a malfunction but as a result of the cars following their programming is what makes it so important to think ahead about how to handle what Lin describes as "a targeting algorithm of sorts."

He explains that such programs would be "systematically favoring or discriminating against a certain type of object to crash into."

As a result, those in "the target vehicles will suffer the negative consequences of this algorithm through no fault of their own."

He does not offer a solution to this problem but it's a warning that we do have to think about how we are going to handle it:

"Spotting these moral hairpin turns now will help us maneuver the unfamiliar road of technology ethics, and allow us to cruise confidently and conscientiously into our brave new future."

That will likely prove an even bigger challenge to navigate than the roads the autonomous vehicles have to drive on.