31 May 2016 | 15:00 GMT

Can You Program Ethics Into a Self-Driving Car?

When self-driving cars kill, it's the code (and the coders) that will be put on trial

By **Noah J. Goodall**

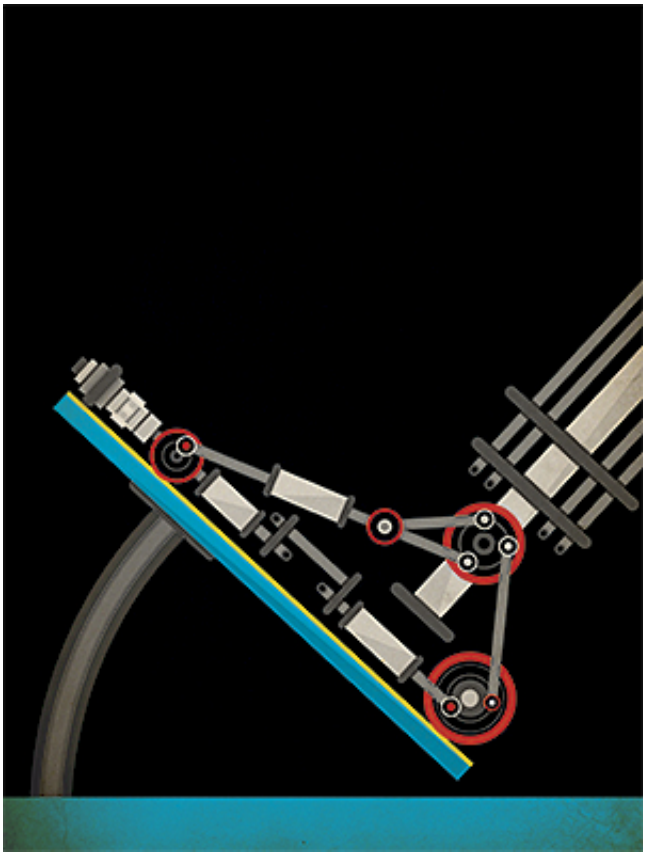


Illustration: Carl De Torres

It's 2034. A drunken man walking along a sidewalk at night trips and falls directly in front of a driverless car, which strikes him square on, killing him instantly. Had a human been at the wheel, the death would have been considered an accident because the pedestrian was clearly at fault and no reasonable person could have swerved in time. But the "reasonable person" legal standard for driver negligence disappeared back in the 2020s, when the proliferation of driverless cars reduced crash rates by 90 percent. Now the standard is that of the reasonable robot.

The victim's family sues the vehicle manufacturer on that ground, claiming that, although the car didn't have time to brake, it could have swerved around the pedestrian, crossing the double yellow line and colliding with the empty driverless vehicle in the next lane. A reconstruction of the crash using data from the vehicle's own sensors confirms this. The plaintiff's attorney, deposing the car's lead software designer, asks: "Why didn't the car swerve?"

Special Report
TRUSTING ROBOTS

Today no court ever asks why a driver does anything in particular in the critical moments before a crash. The question is moot as to liability—the driver panicked, he wasn't thinking, he acted on instinct. But when robots are doing the driving, "Why?" becomes a valid question. Human ethical standards, imperfectly codified in law, make all kinds of assumptions that engineers have not yet dared to make. The most important such assumption is that a person of good judgment will know when to disregard the letter of the law in order to honor the spirit of the law. What engineers must now do is teach the elements of good judgment to cars and other self-guided machines—that is, to robots.

The computerization of driving can be traced back at least to the 1970s, with the introduction of electronic antilock brakes. Now ever more advanced features, like automated steering, acceleration, and emergency braking, are coming every year. The testing of fully automated vehicles, provided that a test driver remains in the vehicle, is allowed in parts of the United Kingdom, the Netherlands, Germany, and Japan, as well as in the United States, where it is explicitly legal in four states and the District of Columbia and at least not prohibited in almost every other. Google, Nissan, and Ford, among others, have said they expect true driverless operation within 5 to 10 years.

Automated vehicles get information on their environments from a range of sensors, such as video cameras, ultrasonic sensors, radar, and lidar (laser-based ranging). Automated vehicles licensed for testing in California are required to provide the Department of Motor Vehicles with all of their sensor data for 30 seconds prior to any collision, of which there have been a score or so, including one with a Google car at fault. Engineers are thus gaining the ability to reconstruct the events around crashes with remarkable precision, using records of what a vehicle was able to sense, the alternatives it considered, and the logic behind its decisions. It will thus be possible to ask a computer to recapitulate its reasoning, much as we might ask human beings to annotate their every decision in a video game or a driving simulator.

Manufacturers and software developers will have to defend a car's actions in ways unimaginable to today's human drivers.

Regulators and litigators will thus be able to hold automated vehicles to superhuman safety standards and to subject them to intense scrutiny following the inevitable, if rare, crashes. Manufacturers and software developers will have to defend a car's actions in ways unimaginable to today's human drivers.

All driving involves risk, and deciding how to distribute that risk among drivers, pedestrians, cyclists, and even property has an ethical component. For both engineers and the general public, it's important that a car's decision-making system weigh the ethical implications of its actions.

Photos: Santa Clara Valley Transportation Authority/AP Photo

I, Fender Bender: Early this year a Google car had a scrape

with a bus—the first accident in which a self-driving car is thought to have been at least partly at fault.

A common response to morally ambiguous situations is to follow the law while minimizing damage as much as possible. This strategy is appealing because it not only allows a developer to justify the car's actions without a lot of effort ("We were in total compliance with the law"), it also passes the responsibility of defining ethical behavior to lawmakers. Unfortunately, it also assumes that the law covers far more than it does.

For instance, in most states, the law relies on a driver's common sense, and it has very little to say about behavior immediately before a crash. In the example above, a vehicle programmed to follow the letter of the law refuses to swerve across the double yellow line even though it risks running over a drunken pedestrian—and even though the other side of the road has only a driverless car that is known to be empty. The law rarely makes exceptions for an emergency as specific as a man falling into the road, and when it does make exceptions, as the state of Virginia's does, the language seems to imply that a movement is legal so long as the car does not crash (the exact language is "provided such movement can be made safely"). In this case, it would be up to the car's developer to decide when it is safe to cross the double yellow line.

Rarely will a self-driving car be absolutely certain that the road is clear and that crossing the double yellow line is safe. Instead, it will estimate the level of confidence at, say, 98 percent or 99.99 percent. Engineers will have to decide ahead of time just how high the confidence level must be to cross a double yellow and how the threshold might vary depending on what the car is trying to avoid, whether it's a plastic bag or a fallen pedestrian.

Even now, self-driving cars are using what might be called judgment to break the law. Google has acknowledged allowing its vehicles to exceed the speed limit to keep up with traffic when going slower would actually be dangerous. Most people would probably favor speeding in other situations as well, such as in an emergency trip to the hospital. Chris Gerdes and Sarah Thornton of Stanford University have argued against encoding laws into software as hard constraints, because drivers seem to consider most laws as costs that are at least somewhat malleable when they can make gains in speed. You don't want to be stuck behind a cyclist for miles because your car refuses to briefly edge over the double yellow line.

Even while staying within the law, an automated vehicle can make many subtle safety decisions. For example, the law is largely silent on how a vehicle should position itself within a lane. Most travel lanes are nearly twice as wide as the typical vehicle, and drivers can use the extra room to maneuver around debris, or position themselves away from erratic vehicles.

Even now, self-driving cars are using what might be called judgment to break the law.

In a 2014 patent, Google takes this concept further, describing how an automated vehicle might position itself in a lane to minimize its risk exposure. The company cites the example of an automated car driving on a three-lane road with a large truck on its right and a small car on its left. To optimize its own safety, the automated car would position itself slightly off-center in the lane, closer to the small car and away from the large truck.

This seems sensible, and it's probably something that most people do, either consciously or unconsciously. Still, it raises ethical concerns. By moving toward the smaller vehicle, the automated car has decreased the overall risk but is now unfairly distributing it. Should the small car have to take on more risk simply because it's small? If this problem involved a single driver's habits, it wouldn't matter much. But if such risk redistribution were formalized and applied to all driverless cars, it could have substantial consequences.

In each of these examples, a car is making a decision about several values—the value of the object it might hit as well as the value of its occupant. Unlike people, who make these decisions instinctively, an automated vehicle would do so as the result of a carefully planned strategy of risk management, which defines a risk as the magnitude of misfortune associated with the feared event multiplied by its likelihood.

Google also patented an application of this type of risk management in 2014. In this patent, the company describes a vehicle that may want to change lanes to get a better view of a traffic light. Or the vehicle could choose to remain in its current lane, where it would avoid taking on the small risk of crashing—say, because of a reading from a faulty sensor—at the cost of that traffic-light information. Each potential outcome is assigned a likelihood as well as a positive or negative magnitude (either a benefit or a cost). Each event's magnitude is multiplied by its likelihood, and the resulting values can be summed. If the benefits outweigh the costs by a reasonable margin, the vehicle will execute the action being considered.

The trouble is that the risk of crashing is incredibly small—the typical driver in the United States crashes once every 257,000 kilometers (160,000 miles) or about every 12 years. Therefore, even with the avalanche of driving data that will come once automated driving takes off, it will be some time before we have plausible crash probabilities for each of the many possible scenarios.

And assigning the magnitude of damage is even harder. Property damage costs are simple enough to estimate—the insurance industry has a lot of experience with it—but injuries and deaths are another story. There's a long history of assigning value to a life, and it is normally expressed as the amount of money one could justify spending to prevent a statistical fatality. A safety improvement that has a 1 percent chance of saving a life for 100 people represents one statistical fatality. The United States Department of Transportation recommends spending US $9.1 million to prevent a fatality, a number inferred from market data, including the premiums that people demand for taking hazardous jobs and what people are willing to pay to buy safety equipment, such as smoke alarms. Not only safety must be weighed in the balance but also the cost of lost mobility or time, which the USDOT puts at $26.44 per hour for personal travel.

It all seems very tidy. But viewing risk in terms of lost lives and wasted commuting time fails to capture much of the moral considerations surrounding how we expose people to risk. For example, an automated vehicle that treated every human life alike would have to give more room on the road to a motorcyclist riding without a helmet than to another one wearing full protective gear because the unprotected one would be less likely to survive a crash. This seems unfair—why should the safety-conscious rider be punished for his virtues?

Another difference between robot ethics and the human kind is that theirs can be warped, even by programmers who had only the best of intentions. Imagine that the algorithm operating a driverless car adjusted the buffering space it assigns to pedestrians in different districts, which it might identify by analyzing settlements from civil proceedings involving crashes. Although this is a perfectly reasonable, well-intentioned, and efficient way of controlling a vehicle's behavior, it can also lead to bad outcomes if, for example, the actual reasons injured pedestrians settled for less were because they lived in low-income neighborhoods. The algorithm would then inadvertently penalize the poor by providing them smaller buffers and slightly increasing their risk of being hit when out for a walk.

It is tempting to dismiss such concerns as idle academic maunderings, but there is no way around them, because computer programs take things quite literally. The time to figure out the consequences of an action is before they happen—in design, rather than the patching phase.

And this is partly why so many researchers use hypothetical situations in which the vehicle must decide between two or more bad outcomes. The most famous of these is the "trolley problem," [pdf] in which a trolley is threatening to collide with unsuspecting children and the only way to stop it is to throw a fat man over the side of a bridge and onto the track's switch. (The man's weight matters: Otherwise, a self-sacrificing onlooker could jump off the bridge himself.) Do you sacrifice one life to save many by such a positive action? If your answer is "no," consider this: You'd no doubt be willing to sacrifice one life to save many by refusing to act—so how can you justify the apparent contradiction?

The ethics of road-vehicle automation is a solvable problem; other fields have handled comparable risks and benefits in a safe and reasonable way.

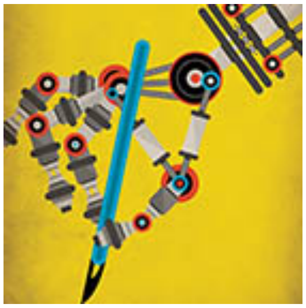
There is a substantial literature on such thought experiments, and indeed, they allow you to stress-test simple and straightforward ethics systems and to find areas where a bit more nuance would be helpful. Suppose an automated vehicle were programmed to avoid pedestrians at all costs. If a pedestrian were to suddenly appear in a two-lane tunnel, and the vehicle couldn't stop in time, the vehicle would be forced to swerve, even into the path of an oncoming bus loaded with passengers. The plausibility of that specific scenario is less important than the flaw it exposes in the vehicle's logic—that valuing pedestrian safety as categorically more important than that of any other road users can actually be much more dangerous in certain situations.

The ethics of road-vehicle automation is a solvable problem. We know this because other fields have handled comparable risks and benefits in a safe and reasonable way. Donated organs are distributed to the sick based on metrics based on quality-adjusted life years and disability-adjusted life years, among other variables. And the military draft has added exemptions for certain useful professions, such as farmer and teacher.

Automated vehicles face a greater challenge. They must decide quickly, with incomplete information, in situations that programmers often will not have considered, using ethics that must be encoded all too literally in software. Fortunately, the public doesn't expect superhuman wisdom but rather a rational justification for a vehicle's actions that considers the ethical implications. A solution doesn't need to be perfect, but it should be thoughtful and defensible.

About the Author

Noah J. Goodall is a research scientist at the Virginia Transportation Research Council, in Charlottesville, Va.



Would You Trust a Robot Surgeon to Operate on You?

Precise and dexterous surgical robots may take over the operating room