Designing ethical self-driving cars

January 25 2023, by Katharine Miller

Credit: Unsplash/CC0 Public Domain

The classic thought experiment known as the "trolley problem" asks: Should you pull a lever to divert a runaway trolley so that it kills one person rather than five? Alternatively: What if you'd have to push someone onto the tracks to stop the trolley? What is the moral choice in each of these instances?

For decades, philosophers have debated whether we should prefer the utilitarian solution (what's better for society; i.e., fewer deaths) or a solution that values individual rights (such as the right not to be intentionally put in harm's way).

In recent years, automated vehicle designers have also pondered how AVs facing unexpected driving situations might solve similar dilemmas. For example: What should the AV do if a bicycle suddenly enters its lane? Should it swerve into oncoming traffic or hit the bicycle?

According to Chris Gerdes, professor emeritus of mechanical engineering and co-director of the Center for Automotive Research at Stanford (CARS), the solution is right in front of us. It's built into the social contract we already have with other drivers, as set out in our traffic laws and their interpretation by courts. Along with collaborators at Ford Motor Co., Gerdes recently published a solution to the trolley problem in the AV context. Here, Gerdes describes that work and suggests that it will engender greater trust in AVs.

How could our traffic laws help guide ethical behavior by automated vehicles?

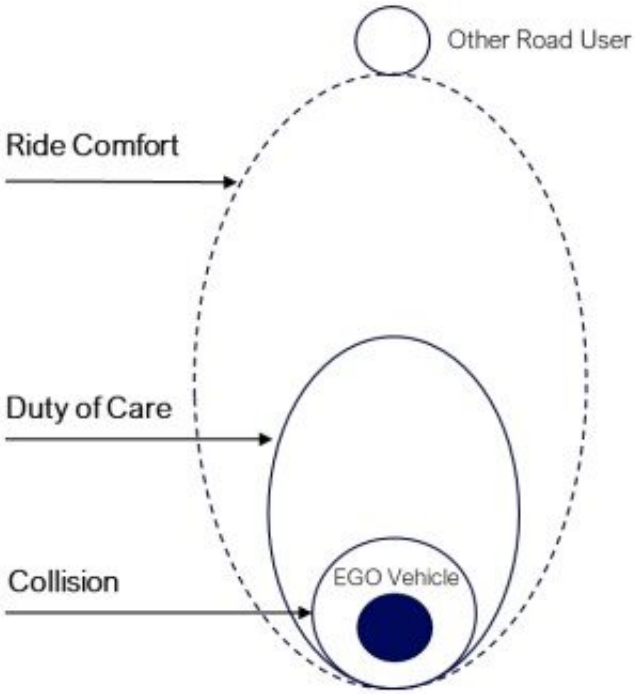
Ford has a corporate policy that says, Always follow the law. And this project grew out of a few simple questions: Does that policy apply to automated driving? And when, if ever, is it ethical for an AV to violate the traffic laws?

As we researched these questions, we realized that in addition to the traffic code, there are appellate decisions and jury instructions that help flesh out the social contract that has developed during the hundred-plus years we've been driving cars. And the core of that social contract revolves around exercising a duty of care to other road users by

following the traffic laws except when necessary to avoid a collision. Essentially: In the same situations where it seems reasonable to break the law ethically, it is also reasonable to violate the traffic code legally.

From a human-centered AI perspective, this is kind of a big point: We want AV systems ultimately accountable to humans. And the mechanism we have for holding them accountable to humans is to have them obey the traffic laws in general. Yet this foundational principle—that AVs should follow the law—is not fully accepted throughout the industry. Some people talk about naturalistic driving, meaning that if humans are speeding, then the automated vehicle should speed as well. But there's no legal basis for doing that either as an automated vehicle or as a company that says that they follow the law.

So really the only basis for an AV to break the law should be that it's necessary to avoid a collision, and it turns out that the law pretty much agrees with that. For example, if there's no oncoming traffic and an AV goes over the double yellow line to avoid a collision with a bicycle, it may have violated the traffic code, but it hasn't broken the law because it did what was necessary to avoid a collision while maintaining its duty of care to other road users.



Example visualization of the various envelopes and considerations for their relative properties. Credit: Exceptional Driving Principles for Autonomous Vehicles: https://repository.law.umich.edu/jlm/vol2022/iss1/2/

What are the ethical issues that AV designers must deal with?

The ethical dilemmas faced by AV programmers primarily deal with exceptional driving situations—instances where the car cannot at the same time fulfill its obligations to all road users and its passengers.

Until now, there's been a lot of discussion centered around the utilitarian approach, suggesting that automated vehicle manufacturers must decide who lives and who dies in these dilemma situations—the bicycle rider who crossed in front of the AV or the people in oncoming traffic, for example. But to me, the premise of the car deciding whose life is more valuable is deeply flawed. And in general, AV manufacturers have

rejected the utilitarian solution. They would say they're not really programming trolley problems; they are programming AVs to be safe. So, for example, they've developed approaches such as RSS [responsibility-sensitive safety], which is an attempt to create a set of rules that maintain a certain distance around the AV such that if everyone followed those rules, we would have no collisions.

The problem is this: Even though the RSS does not explicitly handle dilemma situations involving an unavoidable collision, the AV would nevertheless behave in some way—whether that behavior is consciously designed or simply emerges from the rules that were programmed into it. And while I think it's fair on the part of the industry to say we're not really programming for trolley car problems, it's also fair to ask: What would the car do in these situations?

So how should we program AVs to handle the unavoidable collisions?

If AVs can be programmed to uphold the legal duty of care they owe to all road users, then collisions will only occur when somebody else violates their duty of care to the AV—or there's some sort of mechanical failure, or a tree falls on the road, or a sinkhole opens. But let's say that another road user violates their duty of care to the AV by blowing through a red light or turning in front of the AV. Then the principles we've articulated say that the AV nevertheless owes that person a duty of care and should do whatever it can—up to the physical limits of the vehicle—to avoid a collision, without dragging anybody else into it.

In that sense, we have a solution to the AV's trolley problem. We don't consider the likelihood of one person being injured versus various other people being injured. Instead, we say we're not allowed to choose actions that violate the duty of care we owe to other people. We therefore

attempt to resolve this conflict with the person who created it—the person who violated the duty of care they owe to us—without bringing other people into it.

And I would argue that this solution fulfills our social contract. Drivers have an expectation that if they are following the rules of the road and living up to all their duties of care to others, they should be able to travel safely on the road. Why would it be OK to avoid a bicycle by swerving an automated vehicle out of its lane and into another car that was obeying the law? Why make a decision that harms someone who is not part of the dilemma at hand? Should we presume that the harm might be less than the harm to the bicyclist? I think it's hard to justify that not only morally, but in practice.

There are so many unknowable factors in any motor vehicle collision. You don't know what the actions of the different road users will be, and you don't know what the outcome will be of a particular impact. Designing a system that claims to be able to do that utilitarian calculation instantaneously is not only ethically dubious, but practically impossible. And if a manufacturer did design an AV that would take one life to save five, they'd probably face significant liability for that because there's nothing in our social contract that justifies this kind of utilitarian thinking.

Will your solution to the trolley problem help members of the public believe AVs are safe?

If you read some of the research out there, you might think that AVs are using crowdsourced ethics and being trained to make decisions based upon a person's worth to society. I can imagine people being quite concerned about that. People have also expressed some concern about cars that might sacrifice their passengers if they determined that it would

save a larger number of lives. That seems unpalatable as well.

By contrast, we think our approach frames things nicely. If these cars are designed to ensure that the duty to other road users is always upheld, members of the public would come to understand that if they are following the rules, they have nothing to fear from automated vehicles. In addition, even if people violate their duty of care to the AV, it will be programmed to use its full capabilities to avoid a collision. I think that should be reassuring to people because it makes clear that AVs won't weigh their lives as part of some programmed utilitarian calculation.

How might your solution to the trolley car problem impact AV development going forward?

Our discussions with philosophers, lawyers, and engineers have now gotten to a point where I think we can draw a clear connection between what the law requires, how our social contract fulfills our ethical responsibilities, and actual engineering requirements that we can write.

So, we can now hand this off to the person who programs the AV to implement our social contract in computer code. And it turns out that when you break down the fundamental aspects of a car's duty of care, it comes down to a few simple rules such as maintaining a safe following distance and driving at a reasonable and prudent speed. In that sense, it starts to look a little bit like RSS because we can basically set various margins of safety around the vehicle.

Currently, we're using this work within Ford to develop some requirements for automated vehicles. And we've been publishing it openly to share with the rest of the industry in hopes that, if others find it compelling, it might be incorporated into best practices.

More information: Exceptional Driving Principles for Autonomous Vehicles: repository.law.umich.edu/jlm/vol2022/iss1/2/

Provided by Stanford University