Cars That Think | Transportation | Self-Driving

23 Jun 2016 | 18:00 GMT

People Want Driverless Cars with Utilitarian Ethics, Unless They're a Passenger

We want autonomous cars to be as safe for everyone as possible, as long as they're safest for us first

By **Evan Ackerman**

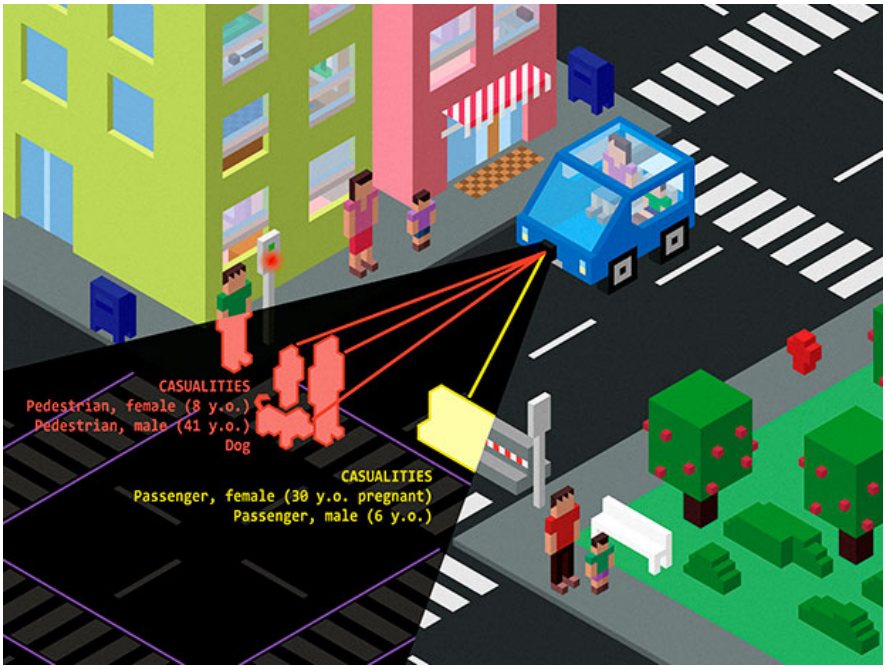

Illustration: Iyad Rahwan

At some point in the nearer-than-might-be-comfortable future, an autonomous vehicle (AV) will find itself in a situation where something has gone wrong, and it has two options: either it can make a maneuver that will keep its passenger safe while putting a pedestrian at risk, or it can make a different maneuver that will keep the pedestrian safe while putting its passenger at risk. What an AV does in situations like these will depend on how it's been programmed: in other words, what ethical choice its software tells it to make.

If there were clear ethical rules that society could agree on about how AVs should behave when confronted with such decisions, we could just program those in and be done with it. However, there are a near infinite number of possible ethical problems, and within each one, the most ethical course of action can vary from person to person. Furthermore, it's not just the passengers who have a say in how AVs behave, but also the manufacturers, and more likely than not, government regulators.

To try to understand how users feel about the potential for AVs to make ethical decisions, Jean-Francois Bonnefon from CNRS in France, Azim Shariff from University of Oregon, and Iyad Rahwan from the MIT Media Lab conducted a series of online surveys, full of questions about AVs in ethical quandaries, as well as how ethical decisions made by AVs might influence the user's perception of them. And like many ethical problems, the results reveal inherent contradictions that may make the adoption of AVs more difficult.

Their paper, recently published in *Science*, shares the results of a series of Amazon Mechanical Turk surveys based on a set of trolley problem variations. The trolley problem, if you're not familiar with it, is essentially this:

> A runaway trolley is headed towards a group of five people standing on the tracks. You are standing next to a lever, and if you pull this lever, the trolley will be switched onto a different track, with a single person standing on it. Do you pull the lever?

The appeal of this problem for anyone studying ethics or morality is that it's very easy to make slight changes to the scenario which result in a much different ethical question. One common variation is replacing the lever with a "fat man" that you have to push in front of the trolley to save the five people. Or, five people becomes ten people, but the single person on the other track is a family member. Or, the five people are all elderly, and the single person is a child.

Here are two examples from the paper, to give you a sense of some of the variations that the researchers were testing:

> *A man is the sole passenger in an autonomous self-driving vehicle traveling at the speed limit down a main road. Suddenly, 10 people appear ahead, in the direct path of the car. The car could be programmed to: swerve off to the side of road, where it will impact a barrier, killing the passenger but leaving the 10 pedestrians unharmed, or stay on its current path, where it will kill the 10 pedestrians, but the passenger will be unharmed.*

> *You and your child are in the car traveling at the speed limit down a main road on a bridge. Suddenly, 20 pedestrians appear ahead, in the direct path of the car. The car could be programmed to: swerve off to the side of road, where it will impact a barrier and plunge into the river, killing you and your child but leaving the pedestrians unharmed; or stay on your current path, where it will kill the 20 pedestrians, but you and your child will be unharmed.*

In total, the researchers conducted six online surveys of nearly 2000 people. Here's a summary of what they found, taken from the paper:

- 76% of participants thought that it would be more moral for AVs to sacrifice one passenger rather than kill 10 pedestrians. They overwhelmingly expressed a moral preference for utilitarian AVs programmed to minimize the number of casualties [a utilitarian approach].
- Participants' approval of passenger sacrifice was even robust to treatments in which they had to imagine themselves and another person, particularly a family member, in the AV.
- Respondents indicated a significantly lower likelihood of buying the AV when they imagined the situation in which they and their family member would be sacrificed for the greater good. It appears that people praise utilitarian, self-sacrificing AVs and welcome them on the road, without actually wanting to buy one for themselves.
- People were reluctant to accept governmental regulation of utilitarian AVs. Participants were much less likely to consider purchasing an AV with such

regulation than without.

All of which leads to this conclusion:

> *"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. Accordingly, if both self-protective and utilitarian AVs were allowed on the market, few people would be willing to ride in utilitarian AVs, even though they would prefer others to do so."*

When confronted with situations where enforcing individual behavior leads to a better global outcome, it's often necessary for regulators to get involved. The researchers offer vaccines as an example: nobody really wants to get stuck with a needle, but if everybody does it, we're all better off. In the context of driverless cars, this means that given the option, most people would choose to ride in or buy an AV that prioritizes their own personal safety above the safety of others, and consequently, car companies will be incentivized to sell cars programmed this way, which is why regulation might be necessary to achieve utilitarian ethics.

Unfortunately, the study also shows that having the government regulate AVs to enforce utilitarian ethical decisions would therefore result in fewer people wanting to buy them, slowing the pace of adoption and leading to more traffic accidents anyway. It may be necessary to enforce the utilitarian ethics that most people want in general, but it's not going to be popular for AV buyers. "Car-makers and regulators alike should be considering solutions to these obstacles," the researchers helpfully suggest.

While there's certainly potential for all kinds of complex and thorny ethical situations for AVs (only a handful of which were explored in this study), I have to wonder just how meaningful these kinds of questions are in the context of a future full of autonomous cars. It's not necessarily clear how often situations like these will come up in practice, especially considering how much safer overall autonomous cars seem likely to be.

To get a better understanding of the context in which AVs will be required to make ethical decisions, we asked the researchers for some clarifications.

IEEE Spectrum: About how often can such situations be reasonably expected to come up?

Jean-François Bonnefon: This is a very hard question: how frequent are these situations? And, does it matter for the decision to program moral algorithms?

At the moment, cars don't have a black box, so after an accident, it's impossible to know if the accident could have been avoided by, for example, the sacrifice of the driver. Even people who have been in accidents are very unlikely to have realized that they were in such a situation, or to remember the decision process during that situation. So, I think it's not possible to get any data about the frequency of situations right now.

Even if the probability turned out to be very small, or if that situation would never occur, we should still program the car before it hits the market. In a way, the frequency of the situation is irrelevant to the programming of the car, but it will be very relevant to the choices that people will make when buying the car.

Azim Shariff: I would add [another] variable, which is the psychological impact of the possibility, which I think is likely to be outsized. I think people already have this inbuilt fear of autonomous technologies, such that just knowing that there is a possibility that the car could make decisions which would imperil the passengers is going to be very psychologically salient for people. It's probably going to cause people to have a lot of pause about going the autonomous route altogether, and the negative consequences for that are quite profound. In order to avoid what's going to be a potentially very rare occurrence, people would be electing not to take the autonomous cars, which will minimize a much larger proportion of accidents.

IEEE Spectrum: And what kind of expectations do you think that passengers will have for autonomous cars in these ethical situations, keeping in mind that the expectations that we have for humans are minimal?

Jean-François Bonnefon: I think that what's going to happen at the start is that people will simply anthropomorphize the cars, and will essentially expect from the cars the same kind of moral center that they would expect from other people. Maybe that will change, maybe progressively we'll reach an understanding of machine ethics that will be different from humans, but at the moment, people are projecting on the cars the moral centers they expect from human drivers.

Iyad Rahwan: I would add that people may even start to expect more from the cars. We have mental models of what we expect from other people, and these mental models include things like, if a dog jumps in front of a car all of a sudden and the person didn't have any time to respond, then we sort of excuse them, because we know that humans are not able to make rational analytical calculations about what's right and wrong in a scenario like this. They will probably freeze, they will probably be scared, and that's okay. But I think once cars are driven by artificial intelligence that may be able to process this kind of information much faster, we may very well expect more from the cars, in terms of moral standards.

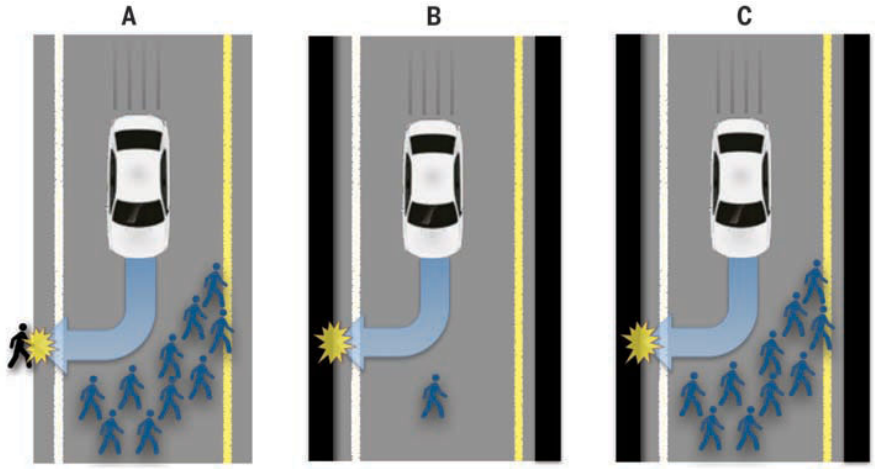

Image: MIT/Science

The researchers' conclusion: "If both self-protective and utilitarian AVs were allowed on the market, few people would be willing to ride in utilitarian AVs, even though they would prefer others to do so."

So now that we know all of this, what should we do? The researchers suggest that "as we are about to endow millions of vehicles with autonomy, a serious consideration of algorithmic morality has never been more urgent... These types of decisions need be made well before AVs become a global commodity." While I can understand the basis for this assertion, for better or worse, I doubt that in practice, in the near term, algorithmic morality is going to make much of a difference to the near future of

autonomous car development or availability.

Whenever we interview people at self-driving car startups, we always ask them about this issue of autonomous ethics, and nuTonomy's Karl Iagnemma sums up where most people doing research in the field seem to be right now:

> *"As of today, we don't have any procedure for what we would commonly think of as ethical decision making. I'm not aware of any other group that does either. I think the topic is a really important one. It's a question that's very important to pose, but it's going to take a while for us to converge to a technical solution for it. We'd love to be able to address that question today, but we just don't have the technology.*
>
> *The other part of it, not that this is a bad thing, is that we're putting more of a burden on the autonomous car than we do on the human driver. Human drivers, when faced with emergency situations where they might have to make a difficult ethical decision, aren't always able to make a reasonable ethical decision in that short amount of time. What level of performance are we going to hold autonomous cars to? The answer is, quite probably, a higher level of performance than we would hold a human driver to, or most people won't accept the technology. That may be unfair, but it doesn't necessarily mean that it's wrong."*

In our special report on trusting robots, Noah J. Goodall extends Iagnemma's thought a little bit farther:

> *"[Automated vehicles] 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."*

Whether or not an autonomous car can be programmed with an ethical system that all drivers (and pedestrians) can agree on, what seems more important is that autonomous cars continue to do what they're being programmed to do already: make decisions that maximize safety in a way that's understandable. This is the approach that most companies seem to be taking right now, whether by technological necessity or for lack of an encodable generalized ethical framework. Based on the fact that more than 90 percent of car crashes are caused by human error, we're still better off with safety-focused autonomous vehicles, whatever their ethics may be.

The Cars That Think Newsletter

Biweekly newsletter about the sensors, software, and systems that are making cars smarter.

About the Cars That Think blog

IEEE Spectrum's blog about the sensors, software, and systems that are making cars smarter, more entertaining, and ultimately, autonomous.

Philip E. Ross, Senior Editor

Willie D. Jones, Assistant Editor

Evan Ackerman, Senior Writer

Lucas Laursen, Contributor

Follow @/CarsThatThink

Subscribe to RSS Feed