Commentary on "Three Scenarios with Self-Driving Vehicles"

Commentary On
Three Scenarios with Self-Driving Vehicles
Autonomous vehicles combine the ethical concerns found in robotics, autonomous systems, and connected devices — and synthesize novel concerns out of these. As with systems that demonstrate increasing autonomy, they also raise questions about moral responsibility. To be *morally responsible* for an outcome is to be the proper object of praise or blame as a result of that outcome. Because the status of autonomous vehicles as *agents* is unclear, and because they are involved in complex socio-technical systems involving many agents and institutions, answering questions about the moral responsibility for outcomes they are involved in is especially vexing.

We can use several lenses to help us determine who is morally responsible for an outcome:

(1) Who caused this outcome? Whose *causal contributions* to the outcome were most significant?

We almost always think that the person who is responsible for an outcome is the person who caused it most directly.

In the case of autonomous vehicles, this is difficult to determine, since autonomous systems are said to “launder” the agency of the people who operate them.[1] If an autonomous vehicle gets in a crash, should we blame the driver, even if they were not operating the car? Or did the car “make” the decision that led to the crash? This is further exacerbated by the fact that autonomous vehicle design is the result of a complex of legal and economic incentives that ultimately *help explain* why the cars are built and programmed the way they are, and thus why they cause the outcomes they do.

(2) Who is it appropriate to blame or praise for this outcome? Who would it be appropriate to punish for this outcome?

Moral responsibility is closely bound up with what philosophers call “reactive attitudes”: moralized attitudes we express in response to what someone has done. [2] Blame and praise are the most obvious of these, but indignation, resentment, and gratitude are other examples. When trying to locate responsibility, we should think about who it would be fair to express these attitudes towards.

Autonomous systems present especially difficult cases for identifying responsible parties because, according to a significant thread in the literature, there is a “responsibility gap” that is created between human agency and the decisions of autonomous machines. If an autonomous machine — like a car — were to make a “mistake,” there would be *no one we could fairly punish*. That argument is a clear example of this way of approaching the question of responsibility: find the people it would be fair to blame or praise, and you have found the responsible party.

Consider scenario 1A: the driver of Car A *should have been paying attention* because their car has merely level 3 autonomy. It is reasonable to think that they have a greater share of the responsibility than the driver of Car B. The driver of Car B could have reasonably believed that their car would be able to handle such a situation, and this attenuates their blameworthiness.

In scenario 1B, when both cars have V2V communication, then our sense of responsibility shifts from the drivers of the cars to the designers of the V2V system. As long as this failure took place in a relatively normal situation, this is the kind of situation that the V2V system should have been able to handle. Therefore, in turn, the drivers would have been reasonable to delegate the decision making to their cars. (Still, the driver of Car B can claim even greater justification for offloading this decision making since, again, their car has level 4 autonomy.)

(3) What is it reasonable to expect or demand of a person, given the role that they occupy?

In a perfect world, the several components of the socio-technical system that designs, manufacturers, regulates, and operates autonomous vehicles would be performing ably and diligently. Each has a complementary role to play:

Institutions can shape the design decisions of autonomous vehicles in a way that individual consumers never could. They can also shape the environment and infrastructure in which they operate. Did the car crash because the lane markings were eroded or unclear, for example?

Designers and manufacturers have a responsibility to test their designs to establish their reliability throughout the spectrum of scenarios that drivers will tend to face. (At least, as far as is practicable, since the permutations of those situations are in fact infinite.) They then have a responsibility to then communicate transparently the capabilities and shortcomings of their vehicles to consumers, and perhaps include designs that nudge — or force — drivers to behave responsibly.

Drivers, finally, need to operate cars responsibly only within their limits.

In these cases, we can ask: Who has failed in their duties to contribute to this harmonious interdependent system? Is there a shortcoming that regulators are uniquely placed to anticipate, but they failed to do so? Or a scenario that manufacturers should have tested for? Should the driver have kept their hands on the wheel, but was texting instead (and is there car level 3, 4, or 5?)? Deciding who failed in their specific obligations, and *how far their behavior departed* from what society can reasonably demand of them, will help us apportion responsibility.

Consider the specific scenarios:

Scenario 2[RJ1] : All three ways of thinking about how to distribute responsibility seem to point to the driver of the standard car that rear-ends the autonomous car: they directly caused the crash; they are more to blame than the autonomous vehicle, and we should expect more of them as a human driver. Note that the autonomous vehicle did something that was unexpected, i.e. it stopped while the light in front of it was green. However, just because it behaved unexpectedly does not mean it behaved recklessly. In fact, the autonomous vehicle behaved as it should have, since the alternative would have likely been to injure the pedestrian crossing in the crosswalk. Thus, it is difficult to blame the autonomous vehicle or its designer[RJ2] .

The pedestrian is also clearly responsible — perhaps as much or more than the driver of the standard car. It is the pedestrian’s recklessness that initiates this chain reaction that results in the crash. They seem, in fact, to make the greatest causal contribution to the situation.

Scenario 3[RJ3] : In this scenario, we again have a system that would normally prevent the crash, which has failed because of a rare situation. Both of the drivers involved behave irresponsibly. Should the designers of the system have designed it

better, or given it a failure mode to cope with situations like this? Which of the drivers is more at fault?

It is hard to say which of the drivers is more at fault. Both are equally reckless in being distracted and both make equal causal contributions to the crash.

The more interesting locus of responsibility may be the automated intersection. Should the designers have tested its capabilities in inclement weather? (Is this an area with a rainy season, or a desert that's experiencing a once-in-a-lifetime downpour? That is to say: *what should they have expected*?) A more graceful failure mode would probably have been to turn the intersection into a four-way stop. This requires all of the drivers who approach the intersection to be much more cautious, increasing safety at the (plainly acceptable) cost of efficiency.

Ideally these approaches would all align. but they don't always. This is why philosophers, lawyers, and others continue to tussle over which method of determining responsibility is most appropriate. However, we can certainly separate the viable from the non-viable answers; and these lenses can help focus our intuitions and show us the path forward in apportioning blame. By clearing the way for productive conversations, moral philosophy has thus shown itself useful.

General readings:

- Jenkins, Ryan. "Autonomous Vehicles Ethics and Law." New America Foundation. September, 2016.

Apportioning responsibility for automated systems:

- Matthias, Andreas. "The responsibility gap: Ascribing responsibility for the actions of learning automata." *Ethics and information technology* 6.3 (2004): 175-183.
- Mittelstadt, Brent Daniel, et al. "The ethics of algorithms: Mapping the debate." *Big Data & Society* 3.2 (2016): 2053951716679679.

Trolley problems and the distribution of harm:

- Himmelreich, Johannes. "Never mind the trolley: The ethics of autonomous vehicles in mundane situations." *Ethical Theory and Moral Practice* 21.3 (2018): 669-684.

- Nyholm, Sven, and Jilles Smids. "The ethics of accident-algorithms for self-driving cars: An applied trolley problem?." Ethical theory and moral practice 19.5 (2016): 1275-1289.