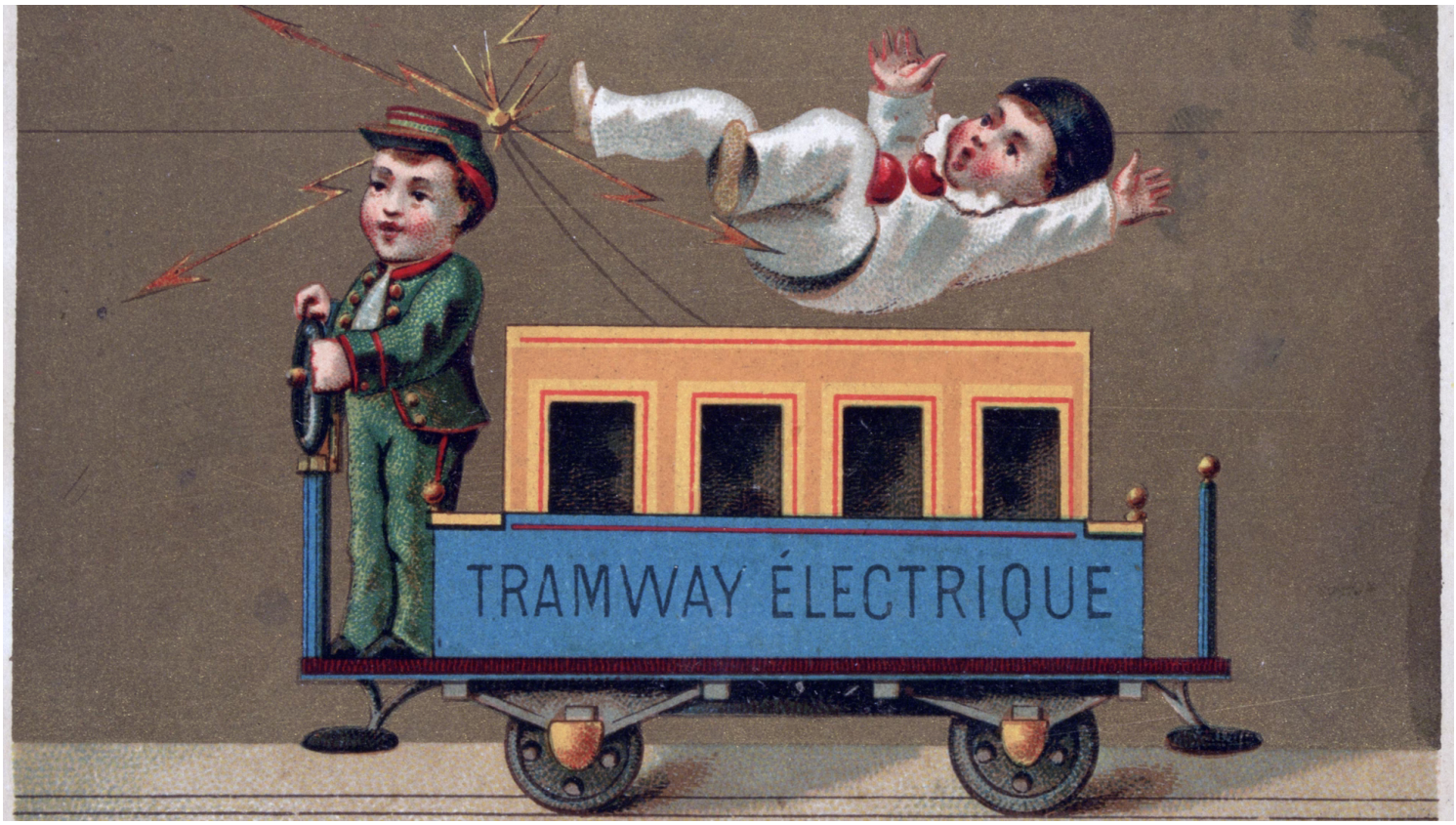
The Atlantic

TECHNOLOGY

Enough With the Trolley Problem

A 50-year-old philosophical thought experiment has been central to the debate about autonomous vehicles. It's time to give it up.

By Ian Bogost



Hulton Archive / Getty

MARCH 30, 2018

SHARE

You know the drill by now: A runaway trolley is careening down a track. There are five workers ahead, sure to be killed if the trolley reaches them. You can throw a lever to switch the trolley to a neighboring track, but there's a worker on that one as well who would likewise be doomed. Do you hit the switch and kill one person, or do nothing and kill five?

That's the most famous version of the trolley problem, a philosophical thought experiment popularized in the 1970s. There are other variants; the next most famous asks if you'd push a fat man off a bridge to stop the trolley rather than killing even one of the supposedly slim workers. In addition to its primary role as a philosophical exercise, the trolley problem has been used a tool in psychology—and more recently, it has become the standard for asking moral questions about self-driving cars.

Should an autonomous car endanger a driver over a pedestrian? What about an elderly person over a child? If the car can access information about nearby drivers it might collide with, should it use that data to make a decision? The trolley problem has become so popular in autonomous-vehicle circles, in fact, that MIT engineers have built a crowdsourced version of it, called Moral Machine, which purports to catalog human opinion on how future robotic apparatuses should respond in various conditions.

But there's a problem with the trolley problem. It does a remarkably bad job addressing the moral conditions of robot cars, ships, or workers, the domains to which it is most popularly applied today. Deploying it for those ends, especially as a source of answers or guidance for engineering or policy, leads to incomplete and dangerous conclusions about the ethics of machines.

The philosopher Judith Jarvis Thomson coined "trolley problem" in 1976, but another philosopher, Philippa Foot, first posed the scenario in a 1967 paper about the difference between what people *intend* and what they can *foresee*. Foot considers abortion as an example. A surgeon who performs a hysterectomy on a pregnant woman *intends* to extract the uterus, but *foresees* the baby's resulting death.

Meanwhile, a doctor who terminates a fetus to save a mother's life directly intends the infant's end. Similar cases suggest different moral conclusions.

Foot poses many related scenarios, one of which is the now-famous tram operator. Another imagines a mob that threatens revenge if a judge doesn't execute an innocent person. The second seems like the trolley problem at first—a choice between more or fewer deaths. But most people who would urge the trolley operator to imperil just one worker would also be appalled at framing the innocent man, Foot writes.

She concludes that there is a difference between what one *does* and what one *allows*. In particular, writes Foot, "the distinction between avoiding injury and bringing aid is very important indeed."

Foot's short paper offers the contemporary reader a more nuanced approach to thinking about the moral scenarios involving autonomous vehicles than does the singular tram example, which became known as the trolley problem. In part, that's because Foot follows a tradition known as virtue ethics, after Aristotle. For virtue ethicists, the quality of an individual's moral character and life is most important.

But the exercise of virtue does not drive most autonomous-car debates. Instead, a concern for the eventual outcome of a whole robocar society is of greatest concern. In moral philosophy, this approach, distinct from virtue ethics, is called *consequentialism*. Consequentialists—including utilitarians, the most famous kind—are concerned with the outcomes and consequences of actions first and foremost.

The utilitarian mindset is very deeply ingrained into the rhetoric of self-driving cars, even before its advocates start making decisions about whom to run down in the event of a calamity. One common rationale for autonomous vehicles is the massive increase in safety they could provide. More than 37,000 people were killed in car crashes in America in 2016. Since more than 94 percent of crashes are caused by driver error, replacing fallible humans with reliable machines seems like an obvious net benefit for society.

The problem is, focusing on outcomes risks blinding people to the virtues and vices of the robocar rollout. Just as Foot's trolley-and-workers scenario is morally different from her judge-and-rioters example, so it is that autonomous outcomes with the same human costs might entail quite different moral, legal, and civic consequences.

Recently, an autonomous Uber in Tempe, Arizona, struck and killed 49-year-old Elaine Hertzberg, a pedestrian walking a bicycle across a road. After I wrote about the possible legal implications of the collision, some readers responded with utilitarian sneers. After all, 5,376 pedestrians were killed by cars in the United States in 2015, and news outlets don't tend to cover each of those as if they are special cases. Soon enough, autonomous cars could reduce or eliminate pedestrian deaths. If you put this idea in trolley-problem terms, the tracks would represent time rather than space. One death is still a tragedy, but if it means making progress toward the prevention of thousands, then perhaps it is justified.

The problem is, that position assumes that Hertzberg's death is identical to any of the unfortunate thousands killed by conventional vehicles. Statistically that might be true, but morally, it isn't necessarily so.

In the future, if they operate effectively, autonomous cars (not to mention front-collision warning systems in traditional cars) are likely to prevent accidents like the Tempe collision with far greater success. Sensors and computers, which can respond to their surroundings better than people, are supposed to perform more effectively than human response and reason can. As details of the Uber collision have trickled in, some experts have concluded that the collision should have been avoided. Furthermore, Uber's cars appear to have fallen short of a company goal of 13 miles of autonomous behavior per human intervention as of March, when the crash occurred. Meanwhile, Google's sister company Waymo claims that its cars can go an average of 5,600 miles without needing a human to take the reins.

On Arizona's roads today, then, the difference between a Waymo autonomous vehicle and an Uber one might be more important than the difference between a human-

operated and a computer-operated vehicle. But in order to lure self-driving car research, testing, and employment to the state, Arizona Governor Doug Ducey allowed all such vehicles to share the roads without significant regulatory oversight.

None of these conditions are addressed by pondering a trolley-problem scenario. To ask if the Uber should have struck Hertzberg or swerved off the shoulder (putting the operator at risk to avoid the pedestrian collision) presumes that the Uber vehicle can see the pedestrian in the first place and respond accordingly. It assumes that that ability is reliable and guaranteed—the equivalent of a mechanical act like throwing a lever to switch a trolley's tracks. This context, missing from the trolly-problem scenario, turned out to be the most important aspect of the outcome in Tempe, both in terms of consequences and morality.

Foot already anticipates the missing context of her cases, even before the tram example became the trolley problem. "In real life," she writes, "it would hardly ever be certain that the man on the narrow track would be killed. Perhaps he might find a foothold on the side ... and cling on as the vehicle hurtled by." One solution to this infinity of possibilities is just to run an infinity of trolley problems, gleaning patterns from the public response to them. That's the Moral Machine's approach, one that matches the way machine-learning systems work best: with a large data set. But another approach would involve considering specific problems in the most appropriate moral context.

As it happens, Foot offers a different example that shares more in common with what actually transpired in Tempe than the trolley does. Imagine five patients in a hospital. Their lives could be saved by being administered a certain gas, but the use of it releases lethal fumes into the room of another patient, who cannot be moved. In this case, the calculus of effect is identical to the classic trolley problem, and yet, to many the conclusion is not nearly so obvious. That's just because of a difference between intended and foreseeable effect, but also because the moral desire to avoid causing injury operates differently.

In the trolley problem, the driver is faced with a conflict between two similar harms, neither of which he or she chooses. But in the hospital-gas example, the doctor is

faced with a conflict between delivering aid and causing harm. In truth, Uber's situation is even more knotted, because none of the parties involved seemed to possess sufficient knowledge of the vehicle's current (not future) capacity for doing harm—not the company that makes the car, the driver who operates it, or the government that regulates it. That makes the moral context for the Uber crash less about the future of vehicular casualty, and more about the present state of governmental regulation, corporate disclosure, and transportation policy. But those topics are far less appealing to think about than a runaway trolley is.

If it's a precedent in moral philosophy that technologists, citizens, and policy makers really want, they might do better to look at Uber's catastrophe as an example of *moral luck*, an idea proposed by the philosopher Thomas Nagel. Here's a classic example: An intoxicated man gets in his car to drive home at night. Though drunk, he reaches his destination without incident. Now consider a man in the same circumstances, but while driving he strikes and kills a child who crosses the road unexpectedly. It seems natural to hold the latter man more blameworthy than the former, but both took the same voluntary actions. The only difference was the outcome.

Seen in this light, the Uber fatality does not represent the value-neutral, or even the righteous, sacrifice of a single pedestrian in the interest of securing the likely safety of pedestrians in a hypothetical future of effective, universally deployed robocars. Instead, it highlights the fact that positive outcomes—safer cars, safer pedestrians, and so on—might just as easily be functions of robocars' moral luck in not having committed a blameworthy act. Until now, of course.

Moral luck opens other avenues of deliberation for robocars, too. In the case of self-driving cars, voluntary action is harder to pin down. Did the Uber driver know and understand all the consequences of their actions? Is it reasonable to assume that a human driver can intervene in the operation of a machine he or she is watching, and not actively operating? Is Uber blameworthy even though the State of Arizona expressly invited experimental, autonomous-car testing on real roads traversed by its

citizenry? All of these questions are being asked now, in Arizona and elsewhere. But that's cold comfort for Elaine Hertzberg.

RECOMMENDED READING

Can You Sue a Robocar?

IAN BOGOST

Would You Pull the Trolley Switch? Does it Matter?

LAUREN CASSANI DAVIS

Is One of the Most Popular Psychology Experiments Worthless?

OLGA KHAZAN

The point of all this isn't to lay blame or praise on particular actors in the recent Uber pedestrian collision. Nor is it to celebrate or lament the future of autonomous vehicles. Rather, it is to show that much greater moral sophistication is required to address and respond to autonomous vehicles, now and in the future.

Ethics isn't a matter of applying a simple calculus to any situation—nor of applying an aggregate set of human opinions about a model case to apparent instances of that model. Indeed, to take those positions is to assume the utilitarian conclusion from the start. When engineers, critics, journalists, or ordinary people adopt the trolley problem as a satisfactory (or even just a convenient) way to think about autonomous-vehicle scenarios, they are refusing to consider the more complex moral situations in which these apparatuses operate.

For philosophers, thought experiments offer a way to consider unknown outcomes or to reconsider accepted ides. But they are just tools for thought, not recipes for ready-made action. In particular, the seductive popularity of the trolley problem has allowed people to misconstrue autonomous cars as a technology that is already present, reliable, and homogeneous—such that abstract questions about their hypothetical moral behavior can be posed, and even answered. But that scenario is years away, if it ever comes to pass. In the meantime, citizens, governments, automakers, and technology companies must ask harder, more complex questions about the moral consequences of robocars today, and tomorrow. It's time to put the brakes on the trolley before it runs everyone down.