‘Would You Kill the Fat Man?’ and ‘The Trolley Problem’

By SARAH BAKEWELL • NOV. 22, 2013

You are walking near a trolley-car track when you notice five people tied to it in a row. The next instant, you see a trolley hurtling toward them, out of control. A signal lever is within your reach; if you pull it, you can divert the runaway trolley down a side track, saving the five — but killing another person, who is tied to that spur. What do you do? Most people say they would pull the lever: Better that one person should die instead of five.

Now, a different scenario. You are on a footbridge overlooking the track, where five people are tied down and the trolley is rushing toward them. There is no spur this time, but near you on the bridge is a chubby man. If you heave him over the side, he will fall on the track and his bulk will stop the trolley. He will die in the process. What do you do? (We presume your own body is too svelte to stop the trolley, should you be considering noble self-sacrifice.)

In numerical terms, the two situations are identical. A strict utilitarian, concerned only with the greatest happiness of the greatest number, would see no difference: In each case, one person dies to save five. Yet people seem to feel differently about the “Fat Man” case. The thought of seizing a random bystander, ignoring his screams, wrestling him to the railing and tumbling him over is too much. Surveys suggest that up to 90 percent of us would throw the lever in “Spur,” while a similar percentage think the Fat Man should not be thrown off the bridge. Yet, if asked, people find it hard to give logical reasons for this choice. Assaulting the Fat Man just feels wrong; our instincts cry out against it.

Nothing intrigues philosophers more than a phenomenon that seems simultaneously self-evident and inexplicable. Thus, ever since the moral philosopher Philippa Foot set out Spur as a thought experiment in 1967, a

whole enterprise of “trolleyology” has unfolded, with trolleyologists generating ever more fiendish variants. (Fat Man was developed by the philosopher Judith Jarvis Thomson, in 1985.)

Judy Garland performed "The Trolley Song" in the film "Meet Me in St. Louis," 1944. MGM/Photofest

Some find it frivolous: One philosopher is quoted as snapping, “I just don’t do

trolleys.” But it really matters what we do in such situations, sometimes on a vast scale. In 1944, new German V-1 rockets started pounding the southern suburbs of London, though they were clearly aimed at more central areas. The British not only let the Germans think the rockets were on target, but used double agents to feed them information suggesting they should adjust their aim even farther south. The government deliberately placed southern suburbanites in danger, but one scientific adviser, whose own family lived in South London, estimated that some 10,000 lives were saved as a result. A still more momentous decision occurred the following year when America dropped atom bombs on Hiroshima and Nagasaki on the argument that a quick end to the war would save lives — and by macabre coincidence, the Nagasaki bomb was nicknamed Fat Man. Similar calculations are being made right now. One has only to think of collateral damage in military strikes, or of the justifications offered for torturing terrorism suspects. At their best, such reasonings are valid; at their worst, they are a writhing nest of weasels.

No wonder trolleyology now forms part of the philosophy course undertaken by cadets at West Point. No wonder too that after several decades maturing in university philosophy departments, trolleyology has burst into the public eye with two books coming out at once. Both are jaunty, lucid and concise. Both explore an array of philosophical sources, from Aristotle and Aquinas to Bentham, Kant and Nietzsche. In “The Trolley Problem,” Thomas Cathcart, a co-author of “Plato and a Platypus Walk Into a Bar . . .” and other works, imagines a real-life trolley case on trial in the “Court of Public Opinion,” a clever but slightly cumbersome device. In “Would You Kill the Fat Man?” David Edmonds, also a seasoned philosophy writer, tells the story more plainly, yet with wit and panache.

Both books deal with difficult questions of reason and instinct, as well as with moral philosophy’s scope and methods. Trolleyology has attracted the interest of practitioners of “x-phi,” or experimental philosophy, who distinguish themselves from the “armchair” philosophers of old (that much-maligned piece of furniture), by borrowing empirical tools from sociology and

psychology to find out what makes people tick. They stop short of laying out sections of trolley track under bridges, but they do use such resources as Harvard University's online Moral Sense Test, where some 200,000 volunteers have tried out their moral intuitions on a range of situations. One experiment at Michigan State University even used a virtual-reality simulation.

The results of such studies have been fascinating, showing, for example, that women are less likely than men to sacrifice the Fat Man, or even to flip the lever in Spur. Other investigations reveal that people are more likely to approve the killing of the Fat Man if they have just seen a comedy clip as opposed to "a tedious documentary about a Spanish village." The contingent nature of our ethical responses in general emerges from other research. We are more generous toward a stranger if we have just found a dime; a judge's decision to grant parole depends on how long it has been since he or she had lunch. Are these the "deep-rooted moral instincts" on which we are willing to found decisions that may affect tens or hundreds of thousands of fellow humans?

Apparently our instincts only feel deep; in fact, they are fickle and easily manipulated. This malleability can be good or bad. Cathcart reminds us that many white people 150 years ago would have considered it instinctively obvious that black people were different and slavery was justifiable. Even a decade ago, many found it self-evident that gay people should not marry or have family lives. (A few still feel this, but now they have to argue their case rather than taking it for granted.) In both cases, rational moral arguments altered assumptions previously taken to be "just the way things are."

This is one reason moral philosophers need not worry about being out of a job yet. A cool utilitarian calculus has its place, and so do our subrational instinctive juices. If either were missing, we would make some truly terrible choices. Yet there is also still room for that quaint seated figure, thinking through the principles and working out a kind of pragmatic yet justifiable

wisdom. An armchair is also a useful place for reading books like these. With all this help, then perhaps when the trolley comes rattling around the corner, and with a half-second to decide, you might just do the right thing. Whatever that may be.

WOULD YOU KILL THE FAT MAN?

The Trolley Problem and What Your Answer Tells Us About Right and Wrong

By David Edmonds

Illustrated. 220 pp. Princeton University Press. $19.95.

THE TROLLEY PROBLEM; OR, WOULD YOU THROW THE FAT GUY OFF THE BRIDGE?

A Philosophical Conundrum

By Thomas Cathcart

Illustrated. 132 pp. Workman Publishing. $14.95.

Self-driving cars programmed to decide who dies in a crash

See how self-driving cars prepare for the real world inside a private testing facility owned by Google's autonomous car company, Waymo. USA TODAY

WASHINGTON — Consider this hypothetical:

It's a bright, sunny day and you're alone in your spanking new self-driving vehicle, sprinting along the two-lane Tunnel of Trees on M-119 high above Lake Michigan north of Harbor Springs. You're sitting back, enjoying the view. You're looking out through the trees, trying to get a glimpse of the crystal blue water below you, moving along at the 45-mile-an-hour speed limit.

As you approach a rise in the road, heading south, a school bus appears, driving north, one driven by a human, and it veers sharply toward you. There is no time to stop safely, and no time for you to take control of the car.

Does the car:

A. Swerve sharply into the trees, possibly killing you but possibly saving the bus and its occupants?

B. Perform a sharp evasive maneuver around the bus and into the oncoming lane, possibly saving you, but sending the bus and its driver swerving into the trees, killing her and some of the children on board?

C. Hit the bus, possibly killing you as well as the driver and kids on the bus?

In everyday driving, such no-win choices are may be exceedingly rare but, when they happen, what should a self-driving car — programmed in advance — do? Or in any situation — even a less dire one — where a moral snap judgment must be made?

It's not just a theoretical question anymore, with predictions that in a few years, tens of thousands of semi-autonomous vehicles may be on the roads. About $80 billion has been invested in the field. Tech companies are working feverishly on them, with Google-affiliated Waymo among those testing cars in Michigan, and mobility companies like Uber and Tesla racing to beat them. Automakers are placing a big bet on them. A testing facility to hurry along research is being built at Willow Run in Ypsilanti.

There's every reason for excitement: Self-driving vehicles will ease commutes, returning lost time to workers; enhance mobility for seniors and those with physical challenges, and sharply reduce the more than 35,000 deaths on U.S. highways each year.

But there are also a host of nagging questions to be sorted out as well, from what happens to cab drivers to whether such vehicles will create sprawl.

And there is an existential question:

Who dies when the car is forced into a no-win situation?

“There will be crashes,” said Van Lindberg, an attorney in the Dykema law firm's San Antonio office who specializes in autonomous vehicle issues. “Unusual things will happen. Trees will fall. Animals, kids will dart out.” Even as self-driving cars save thousands of lives, he said, “anyone who gets the short end of that stick is going to be pretty unhappy about it.”

Few people seem to be in a hurry to take on these questions, at least publicly.

It’s unaddressed, for example, in legislation moving through Congress that could result in tens of thousands of autonomous vehicles being put on the roads. In new guidance for automakers by the U.S. Department of Transportation, it is consigned to a footnote that says only that ethical considerations are "important" and links to a brief acknowledgement that "no consensus around acceptable ethical decision-making" has been reached.

Whether the technology in self-driving cars is superhuman or not, there is evidence that people are worried about the choices self-driving cars will be programmed to take.

Last year, for instance, a Daimler executive set off a wave of criticism when he was quoted as saying its autonomous vehicles would prioritize the lives of its passengers over anyone outside the car. The company later insisted he'd been misquoted, since it would be illegal "to make a decision in favor of one person and against another."

Last month, Sebastian Thrun, who founded Google's self-driving car initiative, told Bloomberg that the cars will be designed to avoid accidents, but that "If it happens where there is a situation where a car couldn't escape, it'll go for the smaller thing."

But what if the smaller thing is a child?

How that question gets answered may be important to the development and acceptance of self-driving cars.

Azim Shariff, an assistant professor of psychology and social behavior at the University of California, Irvine, co-authored a study last year that found that while respondents generally agreed that a car should, in the case of an inevitable crash, kill the fewest number of people possible regardless of whether they were passengers or people outside of the car, they were less likely to buy any car "in which they and their family member would be sacrificed for the greater good."

Self-driving cars could save tens of thousands of lives each year, Shariff said. But individual fears could slow down acceptance, leaving traditional cars and their human drivers on the road longer to battle it out with autonomous or semi-autonomous cars. Already, the American Automobile Association says three-quarters of U.S. drivers are suspicious of self-driving vehicles.

“These ethical problems are not just theoretical,” said Patrick Lin, director of the Ethics and Emerging Sciences Group at California Polytechnic State University, who has worked with Ford, Tesla and other autonomous vehicle makers on just such issues.

While he can’t talk about specific discussions, Lin says some automakers “simply deny that ethics is a real problem, without realizing that they’re making ethical judgment calls all the time” in their development, determining what objects the car will "see," how it will predict what those objects will do next and what the car's reaction should be.

Does the computer always follow the law? Does it slow down whenever it "sees" a child? Is it programmed to generate a random "human" response? Do you make millions of computer simulations, simply telling the car to avoid killing anyone, ever, and program that in? Is that even an option?

“You can see what a thorny mess it becomes pretty quickly,” said Lindberg. “Who bears that responsibility? … There are half a dozen ways you could answer that question leading to different outcomes.”

The trolley problem

Automakers and suppliers largely downplay the risks of what in philosophical circles is known as “the trolley problem” — named for a no-win hypothetical situation in which, in the original format, a person witnessing a runaway trolley could allow it to hit several people or, by pulling a lever, divert it, killing someone else.

In the circumstance of the self-driving car, it’s often boiled down to a hypothetical vehicle hurtling toward a crowded crosswalk with malfunctioning brakes: A certain number of occupants will die if the car swerves; a number of pedestrians will die if it continues. The car must be programmed to do one or the other.

Philosophical considerations, aside, automakers argue it's all but bunk — it's so contrived.

"I don't remember when I took my driver's license test that this was one of the questions," said Manuela Papadopol, director of business development and communications for Elektrobit, a leading automotive software maker and a subsidiary of German auto supplier Continental AG.

If anything, self-driving cars could almost eliminate such an occurrence. They will sense such a problem long before it would become apparent to a human driver and slow down or stop. Redundancies — for brakes, for sensors — will detect danger and react more appropriately.

"The cars will be smart — I don't think there's a problem there. There are just solutions," Papadopol said.

Alan Hall, Ford's spokesman for autonomous vehicles, described the self-driving car's capabilities — being able to detect objects with 360-degree sensory data in daylight or at night — as "superhuman."

"The car sees you and is preparing different scenarios for how to respond," he said.

Lin said that, in general, many self-driving automakers believe the simple act of braking, of slowing to a stop, solves the trolley problem. But it doesn't, such as in a theoretical case where you're being tailgated by a speeding fuel tanker.

Should government decide?

Some experts and analysts believe solving the trolley problem could be a simple matter of regulators or legislators deciding in advance what actions a self-driving car should take in a no-win situation. But others doubt that any set of rules can capture and adequately react to every such scenario.

The question doesn't need to be as dramatic as asking who dies in a crash either. It could be as simple as deciding what to do about jaywalkers or where a car places itself in a lane next to a large vehicle to make its passengers feel secure or whether to run over a squirrel that darts into a road.

Chris Gerdes, who as director of the Center for Automotive Research at Stanford University has been working with Ford, Daimler and others on the issue, said the question is ultimately not about deciding who dies. It's about how to keep no-win situations from happening in the first place and, when they do occur, setting up a system for deciding who is responsible.

A driverless shuttle made its debut in Las Vegas Wednesday with a bump. Police say a semi-truck had a minor collision with the shuttle, less than two hours after the shuttle began carrying passengers. No injuries were reported. (Nov. 8) AP

For instance, he noted California law requires vehicles to yield the crosswalk to pedestrians but also says pedestrians have a duty not to suddenly enter a crosswalk against the light. Michigan and many other states have similar statutes.

Presumably, then, there could be a circumstance in which the responsibility for someone darting into the path of an autonomous vehicle at the last minute rests with that person — just as it does under California law.

But that "forks off into some really interesting questions," Gerdes said, such as whether the vehicle could potentially be programmed to react differently, say, for a child. "Shouldn't we treat everyone the same way?" he asked. "Ultimately, it's a societal decision," meaning it may have to be settled by legislators, courts and regulators.

That could result in a patchwork of conflicting rules and regulations across the U.S.

“States would continue to have that ability to regulate how they operate on the road,” said U.S. Sen. Gary Peters, D-Mich., one of the authors of federal legislation under consideration that would allow for tens of thousands of autonomous vehicles to be tested on U.S. highways in theyears to come. He says that while design and safety standards will rest with federal regulators, states will continue to impose traffic rules.

Peters acknowledged that it would be “an impossible standard” to eliminate all crashes. But he argued that people need to remember that autonomous vehicles will save tens of thousands of lives a year. In 2015, the consulting firm McKinsey & Co. said research indicated self-driving cars could reduce traffic fatalities by 90% once fully deployed. More than 37,000 people died in U.S. roads in 2016 -- the vast majority because of human error.

But researchers, automakers, academics and others understand something else about self-driving cars and the risks they may still pose, namely, that for all their promise to reduce accidents, they can't eliminate them.

“It comes back to whether you want to find ways to program in specifics or program in desired outcomes,” said Gerdes. “At the end of the day, you’re still required to come up with what you want the desired outcomes to be and the desired outcome cannot be to avoid any accidents all the time.

“It becomes a little uncomfortable sometimes to look at that."

The hard questions

While some people in the industry, like Tesla’s Elon Musk, believe fully autonomous vehicles could be on U.S. roads within a few years, others say it could be a decade or more — and even longer before the full promise of self-driving cars and trucks is realized.

The trolley problem is just one that has to be cracked before then.

There are others, like those faced by Daryn Nakhuda, CEO of Mighty AI,

which is in the business of breaking down into data for self-driving cars all the objects they are going to need to "see" in order to predict and react. A bird flying at the window. A thrown ball. A mail truck parked so there is not enough space in the car's lane to pass without crossing the center line.

Automakers will have to decide what the car "sees" and what it doesn't. Seeing everything around it — and processing it — could be a waste of limited processing power. Which means another set of ethical and moral questions.

Then there is the question of how self-driving cars could be taught to learn and respond to the tasks they are given — the stuff of science fiction that seems about to come true.

While self-driving cars can be programmed — told what to do when that school bus comes hurtling toward them —- there are other options. Through millions of computer simulations and data from real self-driving cars being tested, the cars themselves can begin to learn the "best" way to respond to a given situation.

For example, Waymo — Google's self-driving car arm — in a recent government filing said through trial and error in simulations, it's teaching its cars how to navigate a tricky left turn against a flashing yellow arrow at a real intersection in Mesa, Ariz. The simulations — not the programmers — determine when it's best to inch into the intersection and when it's best to accelerate through it. And the cars learn how to mimic real driving.

More: Driverless cars can transform lives — if we change the rules and let them

More: Your new self-driving car will be pioneered by a farmer

More: Google and AutoNation partner on self-driving car program

Ultimately, through such testing, the cars themselves could potentially learn how best to get from Point A to Point B, just by having programmed them to

discern what "best" means — say the fastest, safest, most direct route. Through simulation and data shared with real world conditions, the cars would "learn" and execute the request.

Here's where the science fiction comes in, however.

Playing 'Go'

A computer programmed to “learn” how to play the ancient Chinese game of Go by just such a means is not only now beating grandmasters for the first time in history — and long after computers were beating grandmasters in chess — it is making moves that seem counterintuitive and inexplicable to expert human players.

What might that look like with cars?

At the American Center for Mobility in Ypsilanti, Mich., where a testing ground is being completed for self-driving cars, CEO John Maddox said vehicles will be able to put to the test what he calls “edge” cases that vehicles will have to deal with regularly —such as not confusing the darkness of a tunnel with a wall or accurately predicting whether a person is about to step off a curb or not.

The facility will also play a role, through that testing, of getting the public used to the idea of what self-driving cars can do, how they will operate, how they can be far safer than vehicles operated by humans, even if some questions remain about their functioning.

“Education is critical,” Maddox said. “We have to be able to demonstration and illustrate how AVs work and how they don’t work.”

As for the trolley problem, most automakers and experts expect some sort of standard to emerge — even if it's not entirely clear what it will be.

At SAE International — what was known as the Society of Automotive

Engineers, a global standard-making group — Chief Product Officer Frank Menchaca said reaching a perfect standard is a daunting, if not impossible, task, with so many fluid factors involved in any accident: Speed. Situation. Weather conditions. Mechanical performance.

Even with that standard, there may be no good answer to the question of who dies in a no-win situation, he said. Especially if it's to be judged by a human.

“As human beings, we have hundreds of thousands of years of moral, ethical, religious and social behaviors programmed inside of us,” he added. “It’s very hard to replicate that.”

What the Fatal Uber Crash Doesn’t Tell Us About Self-Driving Cars

This sad accident won’t set any useful precedents.

Jesse Kirkpatrick and Ryan Jenkins March 23, 201812:50 PM

Future Tense

Autonomous vehicles are supposed to be safer than human drivers.

titi-kako/iStock

On Sunday night, a self-driving Uber struck and killed a pedestrian in Arizona.

This appears to be the first fatality in which a self-driving car was involved.

The facts of the case remain unclear, and interpretation of the events will likely change as the local authorities and the National Transportation Safety Board investigate. And while a tremendous amount of ink had already been spilled about the looming ethical challenges of autonomous vehicles—both in academia and public fora—this case ultimately might not help us settle the burning issues that surround this new technology. Here's what we know.

The pedestrian, Elaine Herzberg, was struck by Uber's Volvo XC90 SUV as she attempted to cross a street in Tempe, Arizona. While the Tempe police have made no conclusions about who is at fault, video footage of the crash shows Herzberg attempting to cross the street with a bike, outside of a protected crosswalk, about 10 p.m. The Uber vehicle had a "safety driver"—a human in the driver seat—who is supposed to take control in the event of an emergency. It appears the driver was looking down and only realized what was happening when she heard the sound of the impact. Tempe Police Chief Sylvia Moir stated, "It's very clear it would have been difficult to avoid this collision in any kind of mode." Nevertheless, Uber has suspended its testing of self-driving cars throughout North America—in Arizona, Pittsburgh, San Francisco, and Toronto.

So what does this mean for the testing and rollout of self-driving vehicles? Will states like Arizona, which serves as the anarchic frontier of self-driving regulations since the governor issued an executive order on the testing of self-driving cars, apply the brakes and tighten regulations? Is self-driving technology, even in beta mode, not yet ready for prime time? Many discussing the Uber crash have reiterated familiar worries about the safety of self-driving vehicles. That's an important topic, one that needs debate. But an even greater risk is that, when considering these questions, we might lose sight of one of the major projected benefits of self-driving cars: that they are expected to ultimately *save* lives of tens of thousands of drivers, passengers, and pedestrians.

Perhaps a better question than "Are self-driving cars safe?" is "Should we

blame an autonomous vehicle more than we would a human driver in a similar case?" Autonomous vehicles are supposed to be safer than human drivers—this has been sold to us as their principal benefit. Given that, isn't it only fair to hold AVs to higher standards than humans? If that were true, we might think this crash is actually *worse* than if it had been the result of a human driver.

For example, imagine you are walking along a sidewalk when you see a stranger clutch his chest and keel over. You stoop down to perform CPR. Now, in this situation, if you are a trained physician, a bystander would reasonably demand more of you, precisely because of your greater capacities. If you fail to revive this person, you have done something that is *less easily excused* than if someone else with no medical training had failed in the same task. What it is reasonable to blame someone for is proportionate to what we could have *expected* of her, and what we can expect of someone is proportionate to her capacities.

The manufacturers of autonomous vehicles may find themselves in an analogous situation: Because they have trumpeted the safety benefits of autonomous vehicles, they have dramatically raised the public's expectations. While autonomous vehicles can be expected to lower the number of traffic fatalities greatly, we might also view the harms and deaths they will inevitably cause as worse than those caused by human drivers. How much worse will be a matter of debate. But we are stuck in a Catch-22 for the time being, and the public seems to want it both ways: Autonomous vehicles will be superior to human drivers, despite the deaths they may cause, at the same time that each particular self-driving car crash is worse.

How good is "good enough" when bringing autonomous vehicles onto the road?

However, the Uber fatality probably cannot help us answer the most urgent

and compelling ethical questions that have haunted discussions of autonomous vehicles: the question of responsibility for harms or deaths. Similar worries confound the development of all systems that act so autonomously that they take a human out of the loop for decisions that could cause harm.

In legal circles, they say that "bad facts make bad law": A precedent built on facts that are not favorable to begin with is not going to be useful. This particular crash provides an imperfect test case. Judging from early reports, with the caveat that the investigation is still ongoing, it seems that the pedestrian was at fault and that it would have been nearly impossible for any driver—human or machine—to prevent this accident. (Some things are simply ruled out by the laws of physics.)

Finally, there was a human in the car monitoring the car's activity, so the blame would not fall to the machine itself or the creators of its algorithmic programming. Manufacturers, the public, and regulators still await what would be the true test case of the ethics of autonomous vehicles: an *empty* car in full autonomous mode striking a pedestrian who is in no way at fault, especially if a human driver could have been expected to avoid causing that harm. The slim solace that this tragic accident brings is that it presents the opportunity to have a full and open debate about the appropriate standards for blame in the event of such accidents. Doing so can help us shed some light on just how good is *good enough* when bringing autonomous vehicles onto the road. Unfortunately, this may turn out to be a case of what the Uber crash doesn't tell us.

Jesse Kirkpatrick is assistant director of the Institute for Philosophy and Public Policy at George Mason University.

Ryan Jenkins studies the ethics of new technologies with the potential to impact human life. He teaches philosophy at Cal Poly.