Res Publica (2022) 28:193–216
https://doi.org/10.1007/s11158-021-09519-y

The Ethics of Automated Vehicles: Why Self-driving Cars Should not Swerve in Dilemma Cases

Rob Lawlor[1]

Accepted: 28 May 2021 / Published online: 6 July 2021

Abstract

In this paper, I will argue that automated vehicles should *not* swerve to avoid a person or vehicle in its path, *unless they can do so without imposing risks onto others*. I will argue that this is the conclusion that we should reach *even if* we start by assuming that we *should* divert the trolley in the standard trolley case (in which the trolley will hit and kill five people on the track, unless it is diverted onto a different track, where it will hit and kill just one person). In defence of this claim, I appeal to the distribution of moral and legal responsibilities, highlighting the importance of safe spaces, and arguing in favour of constraints on what can be done to minimise casualties. My arguments draw on the methodology associated with the trolley problem. As such, this paper also defends this methodology, highlighting a number of ways in which authors misunderstand and misrepresent the trolley problem. For example, the 'trolley problem' is *not* the 'name given by philosophers to classic examples of unavoidable crash scenarios, historically involving runaway trolleys', as Millar suggests, and trolley cases should *not* be compared with 'model building in the (social) sciences', as Gogoll and Müller suggest. Trolley cases have more in common with *lab* experiments than model building, and the problem referred to in the trolley problem is *not* the problem of deciding what to do in any one case. Rather, it refers to the problem of explaining what appear to be conflicting intuitions when we consider two cases together. The problem, for example, could be: how do we justify the claim that automated vehicles should *not* swerve even if we accept the claim that we *should* divert the trolley in an apparently similar trolley case?

Keywords Trolley problem · Automated vehicles · Self-driving cars · Autonomous vehicles · Thought experiments · Gogoll and Müller · Nyholm and Smids

✉ Rob Lawlor
r.s.lawlor@leeds.ac.uk

[1] Inter-Disciplinary Ethics Applied, University of Leeds, 17 Blenheim Terrace, Leeds LS2 9JT, UK

Introduction

Most authors who appeal to the trolley problem when discussing automated vehicles appeal to something like Judith Jarvis Thomson's *Bystander at the Switch* case (Thomson 1985, p. 1397), which I will call *Switch:*[1]

> ***Switch***
>
> A trolley is heading towards five individuals on the track. The driver tried the brakes, but the brakes failed, and the driver fainted. You have seen this, and also seen that the trolley can be diverted onto another track, where there is only one individual. The switch is within reach. What do you do?

The expectation, typically, seems to be that finding the right answer to this question will take us a long way—if not all the way—to deciding what automated vehicles should do in similar cases.

In contrast to this approach, I will argue that automated vehicles should be programmed such that they do *not* swerve off course to avoid the person/vehicle in its path[2] in dilemma cases where swerving would put other humans at risk.[3] And I will argue that we should accept this conclusion even if we think that we *should* divert the trolley in *Switch*. But this is not because I consider the trolley problem to be irrelevant or unimportant. On the contrary, many of my arguments *rely* on the methodology associated with the trolley problem.

Before presenting the argument, I should acknowledge my assumptions and provide some clarifications. For the purpose of this paper, I will be assuming that the vehicles are *fully* automated. Therefore, I will not consider the further issue of whether, when or how the vehicle should pass control back to the human driver.[4] Furthermore, on this assumption, this means there will be no human driver in the automated vehicles. Anyone who is in the vehicle will be considered a passenger, without any responsibilities for the car's behaviour.[5]

[1] Thomson's case was, itself, a modification of a case originally presented by Philippa Foot (Foot 1967). See (Woollard and Howard-Snyder 2016) for a discussion of the history, and the differences between Foot's original case and Thomson's.

[2] Note that I said 'in its path', and not 'in front of it'. The phrase 'in its path' is importantly different from 'in front of it': essentially the idea is that, if the road is curved, rather than straight, the vehicle would continue following the curve of the road, staying in its lane.

[3] Obviously, on my view, the vehicle should swerve if it can do so without hitting anyone else. Other nuances will be considered throughout the paper.

[4] Besides, empirical evidence about the time that a human driver would need in order to resume control suggests that passing control to a human driver is unlikely to be a viable option in these dilemma cases (Merat et al. 2014; Allan 2018). It is also worth noting that, in Germany's guidelines for the motor industry, the guidance is that if the human driver fails to act, 'the vehicle simply tries to stop' (Tuffley 2017). This is very much in keeping with the approach I will be defending in this paper.

[5] I do acknowledge that there are further complications if we consider cases in which there is always a driver who retains responsibility, in the way that is comparable to a pilot in a plane with autopilot. And I do not dismiss these complications. However, there is a limit to what can be addressed in a single paper. In this paper, therefore, I will not consider these complications.

Also, for clarity, I should state that, in my examples, the cars are driving on UK roads—and therefore are driving on the left-hand side of the road. And, in all of my examples, I am assuming that automated vehicles share the roads with conventional cars, as well as bicycles, motorbikes and pedestrians.

Also, I embrace the (usually derogatory) term trolleyology. I do this, primarily, because the feature that is typically the focus of derision is in fact a crucial part of the methodology that I will defend. Recognising, and highlighting, this feature is an important part of highlighting the ways in which many critics misunderstand and misrepresent the methodology. The second reason is because using the term 'trolleyology' allows me to avoid the more cumbersome phrase, 'the trolley problem and the methodology associated with the trolley problem'.

Against Swerving

For the purposes of this paper, I will not defend any particular position in relation to *Switch*. Rather, I will assume (for the sake of argument) that the reader believes it would be permissible to divert the trolley. My claim will be that, *even if* you believe it would be permissible to divert the trolley in the *Switch* case, there is good reason to resist the (apparently) similar conclusion in the context of automated vehicles on our roads: automated vehicles should not swerve in these dilemma cases in which human casualties are unavoidable.

(If you do not think it is permissible to divert the trolley in the case above, this will not undermine my argument. On the contrary, this makes my position easier to defend, as you will be less resistant to my position to start with.)

So why shouldn't automated vehicles swerve? To answer this question, I will consider a number of dilemma cases involving automated vehicles, and I will highlight ways in which these cases differ from the Switch case. And, ultimately, I will appeal to these differences to justify the claim that we should not programme cars to swerve in dilemma cases, even if we assume that we *should* divert the trolley in *Switch*. Consider the following case:

Follow the Leader

In this case, there are seven adults waiting to cross the road. The first person looks, and sees a car coming. Nevertheless, she believes she has time to get across the road, and she starts to run across the road. Seeing their friend go to cross the road, five of the others run into the road as well, following behind the first person. One individual however decides that it doesn't look safe, and this individual waits on the pavement.

Now there is one extra crucial detail. The first person did judge the situation correctly: by running across the road, she was able to get to the middle of the road. As such, while she would not be on the pavement, safely off the road, she would be out of the on-coming car's lane, and therefore out of harm's way.

(She had looked both ways and had seen that there were no cars coming the other way.)

However, because the others left the pavement a little later, and because they didn't run as fast, they did not manage to get out of the car's way in time. They will still be in that car's lane.

The car is approaching and will not have time to stop. The car (I stipulate) has three choices:

1. Try to stop, but don't swerve—hitting the five who followed their friend
2. Swerve right onto other side of the road (while also trying to slow down)—hitting the one who crossed first.
3. Swerve left onto the pavement (while also trying to slow down)—hitting the one who stayed on the pavement.

What should we say about this case? The first thing to note is that option 3 looks particularly unattractive. The debate that most people will have here will be about the choice between 1 and 2. Before we get to the debate between options 1 and 2 though, we should reflect on this observation. This observation is important because our strong intuition about option 3 here may highlight a moral consideration that is also relevant in other cases. If we were committed to the simple principle that our aim should be to minimise casualties, option 3 should be no worse than option 2. But that seems implausible. So, we have to ask: what explains the difference?

Compare *Follow the Leader* with the following case:

Three Way Switch

This case is like the original *Switch* case, except there are three tracks, and seven people. On its current course, the trolley will hit and kill five people. If you divert the trolley right, the trolley will hit and kill one person. If you divert the trolley left, the trolley will hit and kill one other person.

Presumably, if you think it is permissible to divert the trolley in *Switch*, you will think it will be permissible to divert the trolley in *Three Way Switch*. The only difference is that you now have to consider the question, how should you decide which way to divert it?[6] But, essentially, if you think diverting the trolley in *Switch* is permissible, there is no obvious reason why you should not divert the trolley left in *Three Way Switch*.

So how do we explain the fact that we think the car should not swerve left, onto the pavement, in *Follow the Leader*? Ultimately, I suggest that the answer is in the question: that person is on the pavement! Before discussing this in more detail, however, it is worth noting that there could be more than one explanation—it could be

[6] Perhaps you should toss a coin. But I will not consider that complication, though I do discuss the value of tossing a coin in dilemma cases in (Lawlor 2006) and (Lang and Lawlor 2016).

overdetermined. Here, though, I will focus on one particular explanation, which has two virtues: first, its simplicity, and second its uncontroversial nature, at least in relation to swerving onto the pavement in this case. This example highlights the fact that we may not want to apply the conclusion we reach in the trolley case to automated vehicles. But, as I will argue in later, the fact that there are differences, and that we do not apply the answer we get in the trolley case to cases involving automated vehicles, does not undermine the trolley problem or the associated methodology.

So why shouldn't the automated vehicle swerve left, onto the pavement? As noted above, I suggest that we appeal to the simple fact that the person is on the pavement. That is, *independent* of any difficult questions about the individuals' responsibility for putting themselves in harm's way, it is a morally relevant consideration that this person is on the pavement while the other six individuals are on the road. It is part of our Highway Code that cars should not drive on the pavement. The pavement is a protected area. Of course, it is also true that cars should not drive on the wrong side of the road—but that norm is weaker than the norm that cars should not drive on the pavement (especially at speed). Cars can, in many circumstances, move onto the other side of the road. In many cases, this is permissible in overtaking, or manoeuvring around obstacles, such as road works or a parked lorry. In contrast, there is a much stronger norm against driving up onto the pavement—especially when travelling at speed. Essentially, I argue that certain areas can and should be thought of as safe spaces. The pavement, pedestrian crossings and traffic islands seem to be the most obvious candidates. My suggestion is that this basic principle should have *very* significant weight: certain designated safe spaces should be protected, such that they remain safe, even in these dilemma cases, even if a car could save more lives by swerving into a single pedestrian on the pavement. And this principle explains why the intuitions we have about what ought to be done in *Follow the Leader* do not match the intuitions we have in relation to *Three Way Switch*.

Safe Spaces and Responsibility

Above, I introduced the *Follow the Leader* case and appealed to the idea of safe spaces to explain the conclusion that the car should not swerve left onto the pavement. I noted at the time, however, that an appeal to safe spaces was not the only explanation one could appeal to.

Another plausible explanation is that the six have acted recklessly and, to some extent, were *responsible for putting themselves in danger*, while the last person made a decision to stay safe and to not run across the road. This is a consideration that did not feature in *Three Way Switch*. Therefore, we could appeal to this as a consideration that explains why the car should not swerve left, even though we do think it would be permissible to divert the *trolley* to the left. We might think that this is a significant consideration that ought to be given some weight at least. In isolation, this seems right to me. However, there are also complications which suggest that we might not want to rely on this principle when considering how we should programme automated vehicles.

In particular, there are practical concerns. Who should be considered responsible for their actions? For example, would we claim that the individuals should be considered responsible for putting themselves in harm's way, even if it was six young children running across the road, rather than adults? And if we wanted to draw this distinction, we would have to consider the fact that an automated vehicle is not likely to be able to accurately judge the ages of the individuals in front of it. Also, whether the individuals are adults or children, it is unlikely that the vehicles will be able to judge whether the individuals are responsible for putting themselves in danger or not. Consider the following case:

Freak Wind[7]

A car is driving in a built-up area, and there are a number of pedestrians on the pavement. It is an exceptionally windy day, and an unusually high wind blows two pedestrians into the road, into the path of the automated vehicle. A third pedestrian also struggles to stay on her feet, but she manages to stay on the pavement. There are also cars driving in the opposite direction in the other lane of the road. The car has the following three options:

1. Brake, but hit the two pedestrians who have been blown onto the road.
2. Swerve left, hitting the pedestrian on the pavement.
3. Serve right into oncoming traffic.

I do not want to say too much about this case, but I include it in order to present a case in which responsibility does not look like it could be the deciding factor.

Even in this case, even though we cannot suggest that the two people in the road are responsible for putting themselves at risk, I will argue that there should be a strong presumption against driving onto the pavement. In any case in which driving onto the pavement would impose a risk of harm onto someone who is on the pavement, there should be a very weighty consideration against doing so, and this is the case even if those in the road are *not* responsible for being there.

This is partly for reasons I discussed in the previous section, in relation to safe spaces, but we also need to consider the practical limits. Presumably, it will be very difficult to program a computer to be able to distinguish between a person stepping into the road voluntarily, a person being blown into the road by a strong wind or a person being pushed into the road by a psychopath. These practical considerations give us another reason to avoid approaches that would rely heavily on a system's ability to judge who is, and who is not, responsible for being in the road and in danger of being hit. (Also see the *Pillion Passenger* case, which I discuss below.)

These complications, I believe, count in favour of the simpler solution presented above.

[7] If you think these examples sound too far-fetched to be taken seriously, see (Sky News 2015), (Watts 2015) and (Shaw 2015).

Back to options 1 and 2

However, we have not yet considered the main issue, which is the choice between options 1 and 2:

1. Try to stop, but don't swerve—hitting the five who followed their friend
2. Swerve right onto other side of the road (while also trying to slow down)—hitting the one who crossed first.

For the sake of simplicity, I will assume that this is a case involving adults, as originally stipulated.

For many, I suspect it will be tempting to defend option 2 on the grounds that this would minimise casualties.[8] However, I will argue that we should reject this line of reasoning. To support my argument, I could appeal to the two principles considered above: one concerning responsibility and the other concerning safe areas. Remember, it could be overdetermined, so both accounts could be right. Again, however, I will ultimately defend the appeal to safe spaces. Nevertheless, it will be illuminating to discuss the ethical issues relating to responsibility.

If we focus on responsibility, we should not forget that the first individual *did* make an accurate judgement and did get out of the way of oncoming traffic and, while she may not have reached the pavement, she did get out of the way of the oncoming traffic, safely to the other side of the road (and remember there were no cars on that side of the road). Therefore, we should not treat the six individuals as individuals who made the *same* decision—each running into the road and putting themselves in danger. Rather, the first individual made a *better* judgement than the other five. She saw that there were no cars coming from one direction, and—regarding the car coming from the other direction—she judged, correctly, that she had time to make it safely to the middle of the road and could then make her way safely to the other side of the road. It may have been a risky decision, and we might criticise her for crossing the road in such a risky manner, but ultimately her judgement was *better* than the others' and she *was* able to get out of harm's way before the car reached her. The other five, in contrast, misjudged the situation, and they put themselves directly in harm's way. Furthermore, their actions put others at risk (at least, this is true *if* cars are likely to swerve).

Given the facts of the case (as I have *stipulated* them), why should the car swerve to avoid those who put themselves in harm's way, swerving instead into the one person who had judged the risk accurately and had got to the other side of the road? If we program automated vehicles to swerve in cases like this, then we effectively give people the power to put others at risk. The first adult running across the road made a decision and she would have been fine if no one followed her. But if the others

[8] It has certainly been the case that many students have opted for this option when I discussed this case in teaching.

follow, and if cars are programed to swerve to minimise casualties, the five following effectively get the first person killed.[9]

Considering this case in isolation, and ignoring the practicalities about what automated vehicles can be expected to consider, this explanation appealing to responsibility seems plausible. In relation to more general principles that can be applied to automated vehicles, however, I will resist this explanation, partly because of the practicalities noted above, and partly for reasons I will discuss in relation to the next case, *Pillion Passenger*.

Again, I suggest that we should appeal to the importance of safe areas. An obvious challenge here, though, is that the road does not look as credible a safe space as the pavement. My response is to argue that we must also consider context, and moral and legal responsibilities.

Consider this in terms of your own behaviour crossing the road. If the other side of the road is clear, one does not necessarily wonder, 'can I get to the pavement before the oncoming car reaches me'. One may simply consider whether it is possible to get past this one lane, to get out of the car's way (assuming the other side of the road is clear). My suggestion, therefore, is that a person should be considered to be in a safe space just as long as they are not somewhere they should not be.[10] The person on the pavement is clearly not somewhere they should not be. Although it is more controversial, I have argued that the first runner is also *not* somewhere she should not be. In contrast, the five are very clearly somewhere they should not be.

Safe Spaces and the Distribution of Responsibilities

Again, the benefit of this approach is that we can avoid the complications discussed in section 'Safe Spaces and Responsibility'. In addition, I will now present another case, which cannot be dealt with by asking who is responsible for putting themselves in danger.

Pillion Passenger

In this case, a car is approaching a crossroads with traffic lights. The car is travelling at 30 mph (the speed limit on the road). Next to the car, there is a cyclist in the cycle lane, also travelling (downhill) at 30 mph. Before they reach the point at which they would need to slow down for the junction, the traffic light changes from red to green. Therefore, both the cyclist and the car continue, confident that they will pass through the junction before the lights

[9] And note, this is a consideration that was not relevant in the *Switch* case. As such, it is a consideration we can appeal to, to explain why automated vehicles should not swerve, even if the trolley should be diverted in *Switch.*

[10] This may mean that, in some cases, we will reach different conclusions in different countries, as a result of different laws having different implications for where people should be. To the extent that there can be reasonable disagreements and reasonable differences between legal systems, I do not consider this to be an implausible implication of my view.

change back to red. And this, I stipulate, is a reasonable and accurate assessment. As such, this should be unproblematic.

However, just as they come to the junction, a motorcyclist with a passenger jumps a red light. Due to the details—the speed, direction and position of the vehicles—the car has three choices:

1. Try to stop, but don't swerve—hitting the back of the motorbike
2. Swerve left—avoiding the motorbike, but hitting the cyclist in the cycling lane.
3. Swerve right—still hitting the motorbike, as well as driving into the path of oncoming traffic.

I assume that we can rule out option 3. But what should we say about the other two options? How does this case compare with *Follow the Leader* or *Switch*? And what lessons can we draw from this case?

One thing to emphasise is that there is a reason I gave the motorcyclist a passenger. And it was not just so that there would be a dilemma between hitting two people or hitting one. The other relevant point is that the motorcyclist is responsible for putting herself in harm's way by jumping a red light—but the passenger is not. So how should we deal with this case?

Earlier, I discussed the issue of responsibility, asking whether we should give less weight to the lives of those who were responsible for putting themselves in danger. In this section, however, I consider a different sense of responsibility. Here, I talk about responsibility in the sense of particular individuals having particular responsibilities. For example, as a parent, it is *my responsibility* to feed *my* children—it is not your responsibility (at least in normal circumstances).

In this section of the paper, I will argue that *this* sense of responsibility is an important element in thinking about the ethics of automated vehicles in dilemma cases. We need to think in terms of *whose* responsibility it is to avoid particular harms. In this case, we need to consider the motorcyclist's responsibilities and the driver's responsibilities—or, in the case of automated vehicles, the car's responsibilities.[11] In particular: it is the motorcyclist's responsibility to avoid putting her passenger in danger. Of course, other drivers should also consider the safety of the motorcyclist and her passenger—and should be vigilant for motorbikes. But it is not the driver's responsibility—or the *car's* responsibility—to save the pillion passenger at the expense of the cyclist when the motorcyclist drives through a red light.[12] In contrast, it *is* the car's responsibility to ensure that it does not put the cyclist at risk by swerving into the cycle lane.

[11] I am not implying the car has true agency. This should simply be read as shorthand. Also, I am suggesting that the car should be programed such that it follows the law in the same way that a human driver should. As such, even if only as a shorthand, it does make sense to talk of the responsibilities that the car has in different situations (e.g. it has a responsibility to stick to the speed limit). The programers should then program the cars with these responsibilities in mind.

[12] Obviously, I agree that it should stop, if possible. And, likewise, it should swerve if it can do so without putting anyone at risk.

Some might challenge this, asking if I would be committed to the idea that the car should not brake hard to avoid the motorbike because doing so would impose additional risks on those behind the car. My response to this would be to appeal to responsibility again, and also to the law, appealing to the principle that it is every driver's responsibility to leave ample space between themselves and the car in front, precisely because the car in the front may need to brake suddenly in an emergency.

It is also worth noting that I am not just appealing to the law as a brute fact, saying that our choices are limited by what would be legally permissible. And I am not defending the crude view that we can ignore ethics, and just follow the law.[13] Rather, I am appealing to *particular* legal responsibilities, suggesting that these are *good* laws, and that we have good reason to support laws which *distribute responsibilities* in this way.

Is the car itself a safe space?

Another issue that some have discussed in relation to the ethics of automated vehicles is the question of whether it is legitimate to put the safety of the person in the car ahead of the safety of other road users—with the added dimension that there is obviously an incentive for car manufacturers to put the owner's safety ahead of other people's, unless legislation takes this choice away.[14]

Mountain Side

A car, with one passenger, is driving on a narrow, winding mountain road. To the left of the car, there is a pedestrian area, with six pedestrians. To the right, there is a small verge (just wide enough for pedestrians to stand) and then the cliff edge. There is no barrier. Suddenly, five of the six pedestrians run across the road from the pavement towards the verge. The car has three options.

1. Try to stop, but don't swerve—killing the five
2. Swerve left (while trying to stop)—killing a pedestrian on the pavement
3. Swerve right (while trying to stop)—driving off the side of the cliff, killing the person in the car.

Considered in isolation (i.e. not in the context of *this* paper), the natural question would be, do we prioritise pedestrians, or do we prioritise those in the car? Or, as Millar puts it, 'Should your robot driver kill you to save a child's life?' (Millar 2014b). But given the context of this paper, and the view that I have defended thus far, it is natural to ask a more specific question: is the car itself a safe space, for those

[13] In places, this does seem to be the view defended by Casey (Casey 2017). At other times, however, he does seem to recognise the distinction between good laws and bad laws. (I discuss Casey in more detail in (Lawlor 2021.)

[14] For example, see (Gogoll and Müller 2016).

inside it? I will argue that the answer to this question is yes, and therefore that the 'robot driver' should not kill you (the passenger of the car) to save a pedestrian.[15]

This is a self-driving car, so presumably it is driving legally: it is not speeding or driving recklessly. It is not doing anything it should not. And you are simply a passenger. As such, *you* are not doing anything you should not—and you are not somewhere you should not be. As such, appealing to the arguments I have presented so far, I suggest that your life should not be sacrificed because someone else—whether through their own risky behaviour or by accident—has left a safe space. Therefore, I suggest that there is a strong case for saying that the car itself should be considered a safe space.

Safe Spaces as Constraints

At this point, I should clarify my position regarding safe spaces. They are not safe spaces in the sense that we must do *everything* we can to protect them and keep them safe. That is, I am not arguing that we should make safe spaces as safe as possible. Rather, they are safe spaces in the sense that their designation as safe spaces imposes limits on what it is permissible for vehicles to do in order to minimise casualties. That is, I am appealing to the idea of safe spaces as constraints.

If you are in a safe space, the law protects you in the sense that it makes it clear that it would be unlawful for a vehicle to drive into a safe space (unless this can be done without imposing a risk on anyone else). This is clear and significant protection (warranting the label 'safe space') even if it cannot *guarantee* that you will not be hit by malfunctioning cars or drunk drivers. This is a form of protection that is consistent with other protections we find in the law.

To see why this is significant, consider the following case:

Blocking

> In this case, an automated vehicle notices that another car has veered out of its lane and is now heading towards the pavement and is likely to hit and kill a couple of pedestrians. The car also calculates that if it accelerates and changes course slightly, it will *intercept* and collide with the stray vehicle, thus blocking the car's trajectory towards the pedestrians, saving their lives, but putting the passenger of the automated car in danger—possibly risking serious injury, and even a possibility of death.

On my interpretation of safe spaces, the automated vehicles should *not* intercept and block the stray vehicle. As I argued above, the car itself should be considered a safe space, and—regarding the pavement—the aim is not to make the safe space as safe as possible.

[15] I should also note that I agree with (Gogoll and Müller 2016) that there is also a problem with leaving this particular issue to personal choice, and ultimately to the market.

Even with these cases in mind, however, the law does offer some protection for those on the pavement, in the form of a deterrent, making it clear that whoever is responsible will be held legally responsible and will be prosecuted if appropriate (whether this is a negligent company or a negligent or reckless human driver).

Safe Spaces and Probabilities

A common complaint against the trolley problem is that, in the standard trolley cases, it is assumed that we know the outcomes that will follow from each decision. There is no uncertainty. Discussing a case in which a car could swerve to avoid an oncoming truck, Nyholm and Smids suggest that:

> To many people, imposing a 1% chance of death on an innocent pedestrian in order to save five car-passengers might appear to be the morally right choice. (Nyholm and Smids 2016, p. 1286)

They then assert:

> The trolley cases do not require any such judgments. (Nyholm and Smids 2016, p. 1286)

But *which* trolley cases do not require any such judgements? Trolleyology is all about creating *new* cases—creating cases to explore the particular variable that we are interested in. Consider the following case, which is in fact based largely on a case introduced by Nyholm and Smids themselves.

> ***Truck***
>
> A heavy truck suddenly appears in the path of a self-driving car carrying five passengers, and the only way for the self-driving car to avoid the truck is to swerve, with the possibility of hitting a pedestrian on the sidewalk.[16] For various reasons, there will be significant uncertainty. Ultimately though, I will assume that the automated vehicle will be programed to make certain assumptions or judgements, and ultimately (whether accurately or not) it will reach some sort of conclusion about probabilities. So let us suppose that the system runs the calculations, resulting in the following judgements.
>
> 1. Do not swerve: 90% chance of a collision. If there is a collision, each of the individuals in the car has a 40% chance of being killed.

[16] Astute readers may have noticed that I removed the reference to the pedestrian being elderly. Following the standard methods of trolleyology, I want to remove variables where possible, to isolate the one ethical issue we want to explore. Of course, if we want to discuss ageism we could construct two cases that are identical except that one case involves a 20-year-old who will be killed and another case with an 80-year-old who will be killed, and we can consider whether that should change our judgements in the two cases. My view—though I will not defend it here—is that we should *not* program automated vehicles to discriminate in this way, based on age (or race, or status or likelihood of curing cancer…).

2. Swerve: 10% chance of hitting the pedestrian. If the pedestrian is hit, there is a 10% chance of the pedestrian being killed. (So, ultimately, the probability of the pedestrian being killed is judged to be 1%.)

(And the person in the truck is unlikely to be seriously injured either way.) Finally, we should remember that we do not know how accurate these probabilities are. These are not likely to be the actual probabilities. These are just the results of the system's calculations, based on limited information and certain assumptions. Nevertheless, these probabilities are the best we have.

If we just ask people for their *initial* intuitions, without significant reflection or argument, and without appealing to the importance of safe spaces and individual responsibilities, I suspect that the majority of people would say that we should program the vehicle to swerve.

However, there is an important point to make here. Although trolley cases have been adopted by psychologists, who do surveys asking respondents what they would do, the methodology I am defending here is the methodology of moral philosophers, not psychologists.[17] And the thing to emphasise here is that, for moral philosophers, any particular case (or pair of cases) is the starting point not the end point. What follows from this starting point is argument: argument about what it would be permissible to do, and argument about what explains the permissibility of one choice and the impermissibility of the other. Ultimately, the aim is to construct a plausible set of moral principles which can explain what we should do in each case, and why, without contradicting ourselves. Therefore, we should explore the issue and argue for different solutions.

I will now argue against the option of swerving, *even in this case*. On the face of it, we have a clash of moral considerations. The probability of harm is clearly a morally relevant consideration. As the probability of hitting/killing the pedestrian goes down, the more tempting it becomes to suggest that the car should swerve. The likelihood of harm is clearly a relevant consideration. At the same time, however, I have argued in favour of constraints, which would seem to rule out—or, at least, provide a strong case against—the option of swerving. The question then is, how weighty are the considerations relating to the probabilities involved, and how do they compare to the principles I have appealed to so far, regarding responsibilities and safe spaces? Are the constraints I have argued for absolute, or could they be outweighed in some cases?

I suggest that we should give priority to the protection of those in safe spaces, and that we should *not* consider it permissible to impose risks on some to save others, *unless* the risk is negligible. That is, the car should not swerve if this would impose a non-negligible risk on the pedestrian on the pavement. Of course, this raises a new

[17] For an interesting contrast between the approach of philosophers and the approach of psychologists, see (Kamm 2008, pp. 422–449). And for arguments highlighting the value of the *empirical* study of trolley dilemmas, using surveys to explore people's views, see (Wolkenstein 2018). I have no objection to the empirical study of attitudes to dilemma cases—but this empirical research is not the focus of this paper.

difficulty. What counts as non-negligible? That is not a question I will answer here. I will say, however, that 1% is far from non-negligible. Even a 'mere' 1% chance of being killed is a very significant probability of death. I suggest that a negligible risk of death is *very significantly* less than 1%.

Now, let us compare and contrast two cases in order to consider the difference that a single change makes. I will start with another case from Thomson.

> ***Transplant***
>
> A surgeon has 5 patients who will die imminently if they do not receive organ transplants. Coincidentally, 'a young man who has just come into [the] clinic for his yearly check-up has exactly the right blood-type' (Thomson 1985, p. 1396). The surgeon realises that if he took all the organs from this young man he could provide organs to her other patients—saving the lives of five patients, at the cost of a single life.

> ***Transplant 1%***
>
> This case is similar to the previous case. In this case, however, the surgeon does not need to take the young man's organs in order to save the lives of the five. Rather, he would only need to perform a procedure which, although not exactly safe, would be *unlikely* to kill the patient. More specifically, the doctor estimates, there is a 10% chance of injuring the young man in a particular way. And, if he is injured in this way, there is a 10% chance that he will die. (So, ultimately, the probability of the young man being killed is judged to be 1%.)[18]

I take it as a given that it would be impermissible to harvest the patient's organs in order to save five lives in the first case. And I have already acknowledged that the probability of harm is a morally relevant consideration—and a significant one. But how significant is it? In this case, I suggest that the reduced probability of death is not sufficient to outweigh the other considerations that count against imposing this risk on the young man, without his consent, to save the five. I take this to be relatively uncontroversial—and if we appeal to the law and codes of ethics in medical practice, the answer is very clear: it would be impermissible.

While I acknowledge that there are also differences between *Transplant 1%* and *Truck 1%* which could warrant further discussion, here I will just emphasise two points:

(1) My argument here is based primarily on the comparison between *Transplant* and *Transplant 1%*. My point is that the moral considerations that count against sacrificing one to save five in a case of certainty remain significant even when the harm is far from certain. This principle can be applied, straightforwardly, to cases involving automated vehicles.

[18] Note the similarity of the probabilities of the outcomes for the pedestrian in *Truck*.

(2) Although I concede that there are differences between the *Transplant* cases and the *Truck* cases, the similarities are significant enough to shift the burden to my opponent, such that it is up to them to argue that the differences are significant enough, such that we can justify imposing risks onto the one to save the five in *Truck 1%*, even if we can't justify this in *Transplant 1%*.

Objections

In this section of the paper, I will respond to two possible objections to this paper—one relating to the methodology of the paper, and one relating to the conclusion I defend.

Objections to the use of the Trolley Problem

In my attempt to defend my position, I have appealed to the trolley problem and made use of the *methodology* associated with the trolley problem. However, many have argued that the trolley problem has little or no value when considering the ethics of automated vehicles. In this paper, I will focus primarily on (Gogoll and Müller 2016) and (Nyholm and Smids 2016), but a number of other papers, with similar objections, will also be discussed in passing.

I will respond to these objections, highlighting ways in which many authors have misunderstood the trolley problem, particularly in the literature on automated vehicles.

To a large extent, much of this work has been done for me in a couple of excellent papers (Keeling 2020) (Hübner and White 2018). However, I will argue that there are some significant flaws which were not exposed explicitly enough in these papers. I suggest that these arguments are important if we are to fully appreciate the weakness of the arguments against the trolley problem, and if we are to fully understand the value of trolleyology.

The Trolley Problem is not the Name of a Dilemma Case

First, it is important to note that the 'trolley problem' is *not* the 'name given by philosophers to classic examples of unavoidable crash scenarios, historically involving runaway trolleys and innocent bystanders' (Millar 2014a) or the name given to a 'type of lesser-of-evils dilemma, where injury is both inevitable and variable' (Casey 2017, p. 1353) (Also see (Lin 2015, p. 78) and (Etzioni and Etzioni 2017, p. 415)).

The term was coined by Thomson, and Thomson is quite explicit in stating what the trolley problem refers to. In the following passage, Thomson is contrasting the trolley case with *Transplant*:

> Why is it that the bystander may turn his trolley, though the surgeon may not remove the young man's lungs, kidneys, and heart? Since I find it particularly

> puzzling that the bystander may turn his trolley, I am inclined to call this The Trolley Problem. Those who find it particularly puzzling that the surgeon may not operate are cordially invited to call it The Transplant Problem instead. (Thomson 1985, p. 1401)

The key question is *not,* 'what would you do?' but rather, 'what explains the difference?'[19]

This highlights another misunderstanding.

Differences are Essential

One of the most common ways in which authors challenge the trolley problem is by appealing to what Keeling calls 'The Moral Difference Argument', according to which:

> Trolley cases and real-world collisions are different in at least some morally significant respects; and these differences render trolley cases of little or no relevance to the moral design problem. (Keeling 2020)

Keeling highlights a number of significant problems with this objection, and his arguments are compelling. However, despite the force of his arguments, Keeling understates the extent to which Nyholm and Smids (and others) have missed the point. Keeling argues that differences between trolley cases and automated vehicles do not render the trolley problem irrelevant. However, he does not emphasise the more important point. It is *not just* that the trolley problem can be relevant, despite differences. Rather, the differences are, in fact, an *essential* element of trolleyology. The methodology necessarily involves *comparisons* between cases, and these comparisons are possible *only* if there are differences between one case and another. If you miss this, you completely miss the point of the trolley problem. But this is exactly what Nyholm and Smids do (along with Gogoll and Müller and many other authors writing about the trolley problem in the context of automated vehicles).

For example, Nyholm and Smids argue that it is problematic that trolley cases are so different from cases involving automated vehicles, and they highlight (among other differences) differences relating to moral and legal responsibilities.[20] (Nyholm and Smids 2016, p. 1287) To see why we should not be persuaded by this objection, consider my arguments in this paper. My arguments explicitly focus on moral and legal responsibilities, and the arguments are based on the very differences that Nyholm and Smids worry about. The differences that Nyholm and Smids worry about do not cause problems for me. They are an essential part of my argument. I

[19] Similarly, note how this question is emphasised by (Woollard and Howard-Snyder 2016), in their summary of the history of the trolley problem.

[20] Bryan Casey also has a paper focusing on the importance of legal responsibility when criticising approaches that use abstract cases like trolley cases (Casey 2017). Casey complains that a focus on ethics essentially misses the point, and he argues that companies and their engineers will just focus on the law: 'Lawyers have got this one', he asserts in the abstract. Similarly, see (Marshall 2017). Even if we accept Casey's argument (which I do not), Casey's argument only shows that *engineers and businesses* need not think about trolley cases—but trolley cases, and the arguments in this paper, would remain relevant to those trying to work out what the law ought to be. (I discuss Casey in more detail in (Lawlor 2021.)

appeal to these differences to explain why, in *Follow the Leader,* it is not permissible for the automated vehicles to swerve left (onto the pavement) even if it is permissible for the trolley to be diverted left in *Three Way Switch*.

There are no Limits

Similarly, Nyholm and Smids also emphasise that, in trolley cases, there 'is only a very limited number of considerations that are allowed to be taken into account'. They continue, 'This is not the ethical-decision situation that is faced by the multiple stakeholders who together need to decide how to program self-driving cars to respond to different types of accident-scenarios' (Nyholm and Smids 2016, p. 1281). And they emphasise, 'they can bring any and all considerations they are able to think of as being morally relevant to bear on their decision about how to program the cars. They can do that and should do so' (Nyholm and Smids 2016, p. 1282).

The point they miss, however, is that trolleyology is all about creating new cases. It is *because* there are *numerous* variations of trolley cases, rather than just a single case, that some people refer to this growing literature as trolleyology (Bakewell 2013) (Brown 2010) (Levy 2014). And there are so many variations because, with each new variation, authors aim to highlight new considerations. This is the very feature of the methodology that is derided by those who use 'trolleyology' as a term of derision. Once we recognise this, we can see that the objection is misguided. Trolleyologists *can* discuss 'any and all considerations' that they consider to be morally relevant. One simply needs to construct the appropriate case or cases. The key point, though, is that trolleyologists seek precision and clarity by constructing cases to exclude other variables, as discussed in the next section.

Lab Experiments, not Model Building

Ultimately, the mistakes considered above are closely related to a more fundamental misunderstanding of the trolley problem, which suggests that the methodology involves something like the following three-stage approach.

First, we start by identifying a trolley case, typically *Switch*.

Second, we reach a judgement about whether it would be permissible or impermissible to divert the trolley in *Switch*.

Finally, we then apply the conclusion from stage two to cases involving automated vehicles.

According to this approach, if it is permissible, in *Switch*, to divert the trolley, to minimise the number of people killed, then we should conclude that it would be similarly permissible to program cars to swerve when this would reduce the number of people killed. This seems to be implicit in many papers which discuss the trolley problem in relation to automated vehicles, but it can be seen clearest when authors *explicitly* spell out their understanding of the methodology. For example, Gogoll and Müller suggest that trolley cases are like scientific models, intended to capture 'the correct set of variables' (Gogoll and Müller 2016, p. 690).

They claim that in applied ethics…

> We use thought experiments as proxies for moral problems in the real world. Thought experiments in applied ethics are useful only insofar as they manage to abstract away distracting details, while retaining the important moral properties and variables of the initial problem X. If we fail to include an important variable of the initial problem in our thought experiment, then the elicited intuitions and the corresponding underlying moral principles will not teach us anything about how to regulate problem X. Creating a moral thought experiment is then essentially similar to what is called model building in the (social) sciences. In creating a model, it is important that we are able to identify the relevant variables at work in a certain situation. The tricky part in modeling, of course, is identifying the correct set of variables. (Gogoll and Müller 2016, p. 690)[21]

With this 'model' analogy, they add a little detail to the three-stage process, such that we have an approach that I will call the 'modelling' approach.

> ***The Modelling Approach***
>
> Step one: we identify a trolley case (or other thought experiment) which captures 'the correct set of variables', which will then function as our 'model'.
> Step two: we work out what the right response would be in our model.
> Step three: based on the principle that 'in applied ethics, we … use thought experiments as proxies for moral problems in the real world' (Gogoll and Müller 2016, p. 690) we apply the conclusion we reached in step two to the real-world problem we are considering.

My claim is that this is a mistake. Most thought experiments—and trolley cases in particular—should *not* be understood as being analogous to modelling. For example, consider another (now infamous) trolley case from Thomson, which I call *Footbridge*.[22]

> ***Footbridge***
>
> This case is like *Switch*, in that you can do nothing, and five people will die, or you can intervene and only one will die. This time, though, the only way you can save the five is by pushing a fat man off a bridge into the path of the trolley. Thomson stipulates that this would kill the person who was pushed into the path of the trolley, but it would also be sufficient to stop the trolley. (It is also stipulated that, while you could jump in front of the trolley yourself, you would not be big enough to stop the trolley.)

[21] This quote does not explicitly mention trolley cases. It talks, more generally, about thought experiments. However, this argument is presented in the context of their criticism of trolley cases being unrealistic.

[22] Thomson calls this case *Fat Man* (Thomson 1985, p. 1409).

It would be *extremely* uncharitable to suggest that Thomson thought that her *blatantly* unrealistic example was intended to mirror the real world—or to act as a '[proxy] for moral problems in the real world' (Gogoll and Müller 2016, p. 690).

Thought experiments like Thomson's are *not* the philosopher's equivalent of models. Rather, they are the philosopher's equivalent of *lab* experiments. As such, I reject the claim that 'Thought experiments in applied ethics are useful only insofar as they manage to abstract away distracting details, while retaining the important moral properties and variables of the initial problem X' (Gogoll and Müller 2016, p. 690). Perhaps *some* thought experiments can be compared to models, acting as proxies for moral problems in the real world, but this is not the case for trolley cases like *Switch* or *Footbridge*. Like lab experiments, they are not intended to be accurate representations of the world. Like lab experiments, the aim is to *isolate* a single variable, and ideally to keep *everything* else constant. If one particular feature is the only difference between two cases, and if we have different intuitions about the two cases, then we can examine that one feature in isolation.[23] In both *Switch* and *Footbridge*, we sacrifice one life but save five. So why is it okay to turn the trolley in *Switch*, but it is not okay to push the man off the bridge in *Footbridge*? Can we identify a moral principle that would allow us to explain the different conclusions we reach in each case?

While this approach can result in examples that seem bizarre or comical, we should remember that the details are chosen for a reason. And we should remember that *lab* experiments are not intended to be realistic representations of the world. The real world is messy and quite different from the lab. And, of course, it is true that one should *not* appeal to the results from lab experiments without being careful to consider the complexities that come with the real world (see Goldacre 2009, pp. 93–97). Despite this, we learn a lot from lab experiments (because of the ability to remove variables), and they remain an *essential* part of science, despite the messiness of the world outside of the lab.

Nyholm and Smids are not as explicit as Gogoll and Müller, but their arguments appear to rest on the same mistake. As noted above, they also highlight numerous ways in which trolley cases differ from cases involving automated vehicles. This, in itself, suggests something like the three-stage approach. At least, it is not clear why they would consider the differences to be problematic, unless they have something like this three-stage process in mind. The three-stage process also seems to be what Nyholm and Smids have in mind when, in the title of their paper, they ask whether

[23] Here, it is worth mentioning why I do not discuss Millar's 'Tunnel Case' (Millar 2014a), which has been discussed by others (Hogarth 2016), (Crisp 2015), (Stokes 2014). This case is primarily presented as a dilemma between saving the passenger of the car and a pedestrian in the road. In addition, however, it also stipulates that the pedestrian is a child, which raises additional issues. Additionally, the child trips, which introduces an element of luck, perhaps changing our view on the question of whether the child was responsible for putting himself in harm's way. This is not to say the case has no value. But, for the methodology I am defending here, the aim should be (as far as possible) to isolate each issue, to consider each in isolation. Similarly, see the truck case in (Nyholm and Smids 2016, p. 1285) and their inclusion of the pedestrian's age, along with the other variables, discussed in footnote 16. (Age is also discussed in (Etzioni and Etzioni 2017).)

accident-algorithms for self-driving cars are an example of 'an applied trolley problem?' (Nyholm and Smids 2016). Similarly, it seems to be what Etzioni and Etzioni have in mind when they complain that hard cases make bad laws (Etzioni and Etzioni 2017), and it seems to be what many critics of the trolley problem have in mind when they complain that trolley cases are unrealistic, absurd or simply *different* from the real-life ethical issues we are interested in. As such, I suggest that Gogoll and Müller are not alone in understanding the trolley problem in this way. However, it is noteworthy that it is an interpretation that is only found amongst critics of the trolley problem. This is not the interpretation that you find in the papers that actually use the trolley problem. If the methodology that the critics criticise is not the same as the methodology that trolleyologists actually use, it seems that the critics have missed their target. They are attacking an approach that no one actually defends.

So how should we understand the trolley problem. I will call this the 'contrast and explain approach.' This is the approach that has been used by Foot and Thomson, and many others, and it is the approach I used in this paper.

The Contrast and Explain Approach

First, identify (or construct) two cases about which people are likely to have different intuitions, reaching different conclusions, despite the fact that the cases are largely similar.
Second, ask the question, what explains the difference?
Third, provide an explanation.

In *contrast* to the modelling approach, this approach *relies* on the cases being different. The differences do not undermine the argument (suggesting that one case is an imperfect model for the other). They *inform* the argument.

As a final note, however, it is worth noting that this approach is not as linear or as simple as the three steps above suggest. The method involves a lot of back and forth, revising intuitions and explanations as new arguments highlight new implications and new problems. Similarly, the process is not as insulated as the three steps above suggest. As the history of the trolley problem highlights, debates do not typically stick to the original cases. New cases are introduced, highlighting new morally relevant considerations and new challenges. As such, the characterisation presented above is a gross simplification. Nevertheless, in the context of this paper, it is sufficient to highlight the extent to which it differs from the model-building interpretation, and to highlight the fact that differences are an essential part of the methodology, rather than flaws that undermine the approach.

The 'Irrational' Objection (and the Value of Predictability)

Regardless of the methodological issues discussed above, I suspect that some will challenge the conclusion that I have defended, arguing that it is irrational. The key point in favour of programing cars to swerve, in any case in which this will save more people than it will kill, is that this approach will reduce the number of deaths on the roads. Therefore, for each individual considering what rules or principles they

would prefer manufacturers to use, it would be rational to want cars to swerve and to do whatever is necessary to save as many lives as possible. Statistically, for each of us, if we think only of our own self-interest, and if we want to minimise our chances of being killed on the roads, we should *want* car manufacturers to program their cars to minimise the number of people killed, and to ignore other details.

I will argue that we should not accept this argument.

First, the empirical assumption can be challenged. Even if it is true, as a statistical *average*, that this policy would reduce people's chances of being killed in a road accident, it does not follow that the policy would actually be better for *every* individual. Remember, an implication of programming vehicles to swerve, in these dilemma cases, is that a pedestrian who is acting responsibly and keeping themselves safe, might be sacrificed in order to save the lives of a larger number who have put themselves at risk. Consider the *Mountain Side* case again.

In this case, and in others, there are two important implications of the fact that cautious pedestrians might be sacrificed to save reckless pedestrians who have put themselves at risk:

First, from an individual's point of view, programing a car to make decisions based on how many people will be killed (or hit), introduces unpredictability. As stated above, the reckless behaviour of others can put you at risk. In contrast, an approach that emphasises safe spaces introduces a predictability and clarity, which gives you more control over your own safety. If you stay on the pavement (where possible), and only cross the road at designated pedestrian crossings or only cross the road when it is very clearly safe to do so, then you are in a good position to keep yourself safe. Of course, I am not suggesting that you would have a 0% chance of being killed. There is always the possibility of a reckless driver, for example, losing control and mounting the pavement. Nevertheless, for individuals who are *more cautious than average*, it is plausible to suggest that *they* would be safer if cars were not programed to swerve than they would if cars were programed to swerve to save lives, without constriants. And, even if you are not typically cautious, it is in your power to be—you can make decisions that will keep you safe.

This also points to the second point. From society's perspective, the (apparently) consequentialist approach could potentially backfire. If my safety depends primarily on my own behaviour, this gives me more of an incentive to be cautious. Therefore, we cannot rule out the possibility that this incentive could be sufficient to have the effect that the roads would be safer if vehicles were *not* programed to minimise casualties. Therefore, in contrast to the rest of this paper, this particular argument could provide a *consequentialist* justification for the position I am defending.

Ultimately though, my main arguments will not rely on claims about what the actual consequences would be. Rather, I put more weight on the non-consequentialist arguments. And this should not be a surprise. Trolley cases have typically been used to challenge consequentialism, and to fine-tune non-consequentialist principles. In particular, arguments in the trolley literature emphasise that consequentialism is often more plausible when considered in the abstract, and less plausible when

one considers the *implications* of the theory.[24] Cases like *Footbridge* and *Transplant* highlight the counter-intuitive implications of consequentialism, *challenging* the principle that we should simply aim to save as many lives as possible, and they highlight the moral significance of other considerations (the idea of appealing to trolley cases, and the trolley problem, to challenge consequentialism and an—often unthinking—commitment to harm minimisation without constraints is also a significant part of (Keeling 2020) and (Hübner and White 2018)).

Perhaps most importantly, the argument that my position is irrational relies on a problematic understanding of what is rational. We should not accept it. It is *not* irrational to appeal to moral considerations. This is true even if they are contrary to one's self-interest. It is not irrational for me to recognise that it would be morally wrong to kill one patient, to take his organs, to save the lives of five other individuals. It is not irrational to think this, even if I am one of the patients whose life depends on an organ transplant. Even if it would be against my self-interest, it would *not* be irrational for me to vote against a proposal to make it legal for doctors to kill someone when doing so would save more lives overall, by providing organs for life-saving surgery.

Conclusion

In this paper, I have highlighted a number of flaws in common objections to the trolley problem (and other thought experiments), and I have identified and argued against a three-stage process, which I call the 'modelling approach.' I have argued that this approach involves a significant misunderstanding of the trolley problem and fails to recognise the real value of the methodology. In contrast, I have defended an approach which I call the 'contrast and explain approach', and I have used this approach to argue that automated vehicles should *not* be programed to swerve in dilemma cases, where there is (for example) a choice between hitting one or hitting five. Furthermore, I argue that this is the conclusion that we should reach *even* if we start by assuming that we *should* divert the trolley in *Switch*.

Acknowledgements I am grateful to Kevin Macnish for comments on an earlier draft of this paper, and I would like to thank the anonymous referees for this journal, and the editor, for very helpful comments on my earlier drafts, and for a particularly well-managed review process.

[24] I consider this in more detail in (Lawlor 2009)—pages 135 to 138.

References

Allan, M. 2018. Drivers need up to three seconds to retake control of autonomous cars. *i*. https://inews.co.uk/essentials/lifestyle/cars/car-news/drivers-need-three-seconds-retake-autonomous-car-control/. Accessed 7 Aug 2018.

Bakewell, S. 2013. 'Would you kill the fat man?' and 'The trolley problem'. *The New York Times*. http://www.nytimes.com/2013/11/24/books/review/would-you-kill-the-fat-man-and-the-trolley-problem.html. Accessed 1 Apr 2017.

Brown, A. 2010. Trolleyology and morals. *The Guardian*. https://www.theguardian.com/commentisfree/andrewbrown/2010/oct/28/phillipa-foot-trolley-morality. Accessed 1 Apr 2017.

Casey, B. 2017. Amoral machines, or: How roboticists can learn to stop worrying and love the law. *Northwestern University Law Review* 111 (5): 1347–1366.

Crisp, R. 2015. The tunnel problem | practical ethics. *Practical Ethics, University of Oxford*. http://blog.practicalethics.ox.ac.uk/2015/07/the-tunnel-problem/. Accessed 11 Sept 2018.

Etzioni, A., and O. Etzioni. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics* 21 (4): 403–418. https://doi.org/10.1007/s10892-017-9252-2.

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review* 5: 5–15.

Gogoll, J., and Müller, J. F. 2016. Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics*, pp. 1–20. https://doi.org/10.1007/s11948-016-9806-x

Goldacre, B. 2009. *Bad science*. London: Fourth Estate.

Hogarth, R. 2016. Driverless cars will take us into a moral maze. *The Times*. https://www.thetimes.co.uk/article/driverless-cars-will-take-us-into-a-moral-maze-kg7z582vx. Accessed 11 Sept 2018.

Hübner, D., and L. White. 2018. Crash algorithms for autonomous cars: How the trolley problem can move us beyond harm minimisation. *Ethical Theory and Moral Practice* 21 (3): 685–698. https://doi.org/10.1007/s10677-018-9910-x.

Kamm, F.M. 2008. *Intricate ethics: Rights, responsibilities, and permissable harm*. New York, Oxford, USA: Oxford University Press.

Keeling, G. 2020. Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics* 26 (1): 293–307. https://doi.org/10.1007/s11948-019-00096-1.

Lang, G., and R. Lawlor. 2016. Numbers scepticism, equal chances and pluralism: Taurek revisited. *Politics, Philosophy & Economics* 15 (3): 298–315. https://doi.org/10.1177/1470594X15618967.

Lawlor, R. 2006. Taurek, numbers and probabilities. *Ethical Theory and Moral Practice* 9 (2): 149–166. https://doi.org/10.1007/s10677-005-9004-4.

Lawlor, R. 2009. *Shades of goodness*. Basingstoke: Palgrave Macmillan.

Lawlor, R. 2021. Automated vehicles: Amoral economics, law and professional ethics. https://www.academia.edu/45495954/Automated_Vehicles_Amoral_Economics_Law_and_Professional_Ethics. Accessed 13 Mar 2021.

Levy, F. (2014, May 6). Trolleyology in Oklahoma. *Huffington Post*. http://www.huffingtonpost.com/francis-levy/trolleyology-in-oklahoma_b_5267234.html. Accessed 1 Apr 2017.

Lin, P. 2015. Why ethics matters for autonomous cars. In *Autonomes fahren*, ed. M. Maurer, J.C. Gerdes, B. Lenz, and H. Winner, 69–85. Berlin, Heidelberg: Springer, Berlin Heidelberg. https://doi.org/10.1007/978-3-662-45854-9_4.

Marshall, A. 2017. Lawyers, not ethicists, will solve the robocar "Trolley Problem." *Wired*. https://www.wired.com/2017/05/autonomous-vehicles-trolley-problem/. Accessed 1 Aug 2018.

Merat, N., A. H. Jamson, F. C. H. Lai, M. Daly, and O. M. J. Carsten. 2014. Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation Research Part f: Traffic Psychology and Behaviour* 27: 274–282. https://doi.org/10.1016/j.trf.2014.09.005.

Millar, J. 2014a. An ethical dilemma: When robot cars must kill, who should pick the victim? | Robohub. http://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/. Accessed 1 Apr 2017.

Millar, J. 2014b. Should your robot driver kill you to save a child's life? *The Conversation*. http://theconversation.com/should-your-robot-driver-kill-you-to-save-a-childs-life-29926. Accessed 2 Aug 2018.

Nyholm, S., and J. Smids. 2016. The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice* 19 (5): 1275–1289. https://doi.org/10.1007/s10677-016-9745-2.

Shaw, M. 2015. OAP and pet dog died after being 'blown into road by gale-force winds'. *The Mirror*. http://www.mirror.co.uk/news/uk-news/pensioner-pet-dog-died-after-6243901. Accessed 5 Nov 2016.

Sky News. 2015. Man, 90, dies after being blown into bus. *Sky News*. http://news.sky.com/story/man-90-dies-after-being-blown-into-bus-10337067. Accessed 5 Sept 2017.

Stokes, A. 2014. The ethics of driverless cars. https://phys.org/news/2014-08-ethics-driverless-cars.html. Accessed 11 Sept 2018.

Thomson, J. J. 1985. The trolley problem. *The Yale Law Journal* 94 (6): 1395–1415. https://doi.org/10.2307/796133.

Tuffley, D. 2017. At last! The world's first ethical guidelines for driverless cars. *The Conversation*. http://theconversation.com/at-last-the-worlds-first-ethical-guidelines-for-driverless-cars-83227. Accessed 5 Sept 2017.

Watts, M. 2015. Shopkeeper describes how freak wind blew pensioner into path of bus. *Evening Standard*. http://www.standard.co.uk/news/crime/finchley-crash-horrifying-cctv-shows-moment-90yearold-man-was-blown-into-path-of-bus-by-gust-of-wind-a3131156.html. Accessed 2 Apr 2017.

Wolkenstein, A. 2018. What has the trolley dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars. *Ethics and Information Technology* 20 (3): 163–173. https://doi.org/10.1007/s10676-018-9456-6.

Woollard, F., and Howard-Snyder, F. 2016. Doing vs. allowing harm. In: *The Stanford Encyclopedia of Philosophy* (Winter 2016.) ed. E. N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/doing-allowing/. Accessed 1 Apr 2017.