The Tesla Problem: Trolleys, Proximity, and Duty

Charles Harry Smith
The London School of Economics

Introduction

I want to draw attention to problems that close physical proximity between the driver and bystander can reveal in a new take on Judith Jarvis Thomson's well-known *Trolley Problem*. I apply Thomson's reasoning to hypothetical situations involving semi-autonomous, self-driving vehicles in order to highlight how complicated and counter-intuitive the distinctions between the two roles can become when physical distance is taken into account. This '*Tesla Problem*' has clear implications for ethics and the laws governing autonomous vehicles, and they redound to how we should understand Thomson's original problem. Beginning with a brief explication of Thomson's original variations of the issue in '*The Trolley Problem'* (Thomson, 1985), I suggest that her reasoning does not give us the tools necessary to unravel the *Tesla Problem*. Moreover, I argue that the implications of the *Tesla Problem* impose more stringent conditions on moral duty in situations similar to Thomson's original *Bystander Problem*.

Trolley Problem Variations

Thomson's paper builds on Phillipa Foot's earlier articulation of a *Trolley Problem* which asks whether it is morally permissible to alter the course of an out of control tram that will kill five people caught on the tracks if not diverted[74]. Unfortunately, if the diversion is taken, thereby saving the lives of the five, it will result in the death of a lone individual on the other track. In Foot's original formulation, you are the *driver* of the trolley. Your choice is between killing one or killing five, depending on which track you choose to take. If you are the driver, killing one is taken to be better than killing five, all things being equal. Thomson's alteration makes you, the moral agent, merely a *bystander* to the affair, observing the trolley from afar and affecting its route by changing the tracks with a switch. Hence, she calls it the *Bystander Problem.* Now, the choice seems different: it becomes a matter of choosing between *acting* to kill one or *letting* five die through inaction.

In contrast with the driver's case, Thomson proposes that, at first, it

74 Thomson summarises their previous debate in her (1985) article which I engage with throughout this paper. It is worth also reading Foot's *The Problem of Abortion and the Doctrine of the Double Effect* (1967) and Thomson's *Killing, Letting Die, and the Trolley Problem* (1976) to understand the development of the literature.

seems that actively killing one is worse than letting five die because, in an analogous case, it would be an abuse of rights to allow a surgeon to harvest the organs of a healthy individual to save five patients. However, the *Bystander Problem* shows that this would mean it is morally impermissible for you, the bystander, to redirect the trolley and kill the one to save the five. Most people's moral intuition is the opposite: that it is not only permissible for the bystander to act, but that they *should* divert the trolley and save the five lives. Consequently, more work must be done to narrow down when it is or is not permissible to intervene.

Thomson proposes that to avoid purely utilitarian calculations, such as in the case of the organ-harvesting surgeon, we must appeal in a Kantian manner to the right not to be used as *merely* a 'means to an end'. And from Dworkin she adopts the premise that "rights trump utilities" (Dworkin 1977:223-239), arguing that we may not infringe on the 'stringent' rights of the one to save five lives thus preserving the difference between diverting the trolley to kill one and save five, which seems acceptable, and pushing an overweight bystander from a bridge in front of the trolley to stop it, which does not seem as acceptable though one also kills one to save five. Thomson calls this latter example the *Fat Man Problem*. This is where the big question arises: what makes the intuitive difference between these two cases?

As Thomson argues, the circumstances *here and now* are directly relevant to both what is permissible and to assessments of which acts are worse. Other things being equal, it is only permissible to kill one person to save the five if what threatens the five *here and now* threatens the one instead. The fat man on the bridge was in no danger of being hit by the trolley until you pushed him, directly infringing upon his right to life and assaulting him. You would be infringing his rights *merely* to use him as a 'means to an end', viz. to stop the trolley. In contrast, switching the track does not directly breach anyone's stringent rights, only the trivial property rights of the railway company. Pushing someone, on the other hand, would infringe upon their rights to bodily autonomy even if there was no trolley whatsoever. Furthermore, no matter whether you pulled the switch or not, some number of people on the track will be killed by the trolley. If, however, you pushed someone over the railings, you would be directly intervening, introducing them to a threat they were not exposed to before.

Consequently, Thomson arrives at her conclusion: it is permissible to act to minimise the number of deaths caused by a threat *if and only if* the action does not directly infringe on anyone's 'stringent' rights (Thomson 1985:1409). So, it is morally permissible for the bystander to pull the switch and kill a person to minimise the deaths *caused by the trolley,* but not to minimise the total deaths *à la* utilitarianism such as in the *Fat Man Problem*. She calls this the *distributive*

exemption to the 'rights trump utilities' principle. It "permits arranging that something that will do harm anyway shall be better distributed than it otherwise will be [in order to] do harm to fewer rather than more" (Thomson 1985:1408). The situation is not altered in a significant manner with respect to rights, as the circumstances *here and now* – the rights infringed – have not changed, only the number of deaths. And, if we can minimise the deaths, then this is the preferable course of action.

Trolleys and Cars

The *Trolley Problem* is a situation which few people will ever find themselves in given the fact that very few people drive trams. That said, the intuitions and implications of the *Trolley Problem* can be readily transposed onto analogous situations. One natural extension is to apply it to driving a car. Of course, if you had time to think the problem through whilst driving, it would most likely be avoidable: you would have time to brake, swerve or otherwise dodge the accident if you could deliberate. We most often do not have the luxury of time when faced with these sorts of moral dilemmas though. The fact that 94% of car crashes can be attributed to human error[75] makes a strong case for the adoption of autonomous cars as once most vehicles on the road are autonomous, incidents of human error should be reduced and there should be fewer crashes and accidents. Autonomous cars will reduce crashes because they can process information faster than we can. They would have more time to deliberate as they run on software that chooses between different courses of action thousands of times a second. It is certainly conceivable that at least some of these choices could involve variations of the *Trolley Problem*, in that whatever the outcome, some unavoidable harm will be caused. Thus, the moral implications of the *Trolley Problem, Bystander Problem* and *Fat Man Problem* apply to the context of autonomous vehicles as the cars' programming will need to be able to reflect those implications. Autonomous cars must be able to choose between harms in morally permissible ways.

Programming a finalised, ordered list of every conceivable moral outcome, ranked relatively to one another, would be impractical. It is far more likely that moral principles would be encoded and used to process the sensory data that the car acts on much like a form of deontology or rule utilitarianism. The final program, whatever moral rules it follows, will be propagated to every vehicle in the fleet, so we need to know what we consider to be desirable moral outcomes so that the vehicles can implement them[76]. As of yet, there is no

75 Singh, S. (2015). *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey.* (No. DOT HS 812 115).

76 The cynic might quip that autonomous cars are unlikely to be bought, at great expense, if they are programmed to kill their occupants. Let's assume that government legislation is passed to regulate the software so all autonomous vehicles

agreed 'solution' to the various *Trolley Problems* and so, until the answers (if they exist) are settled on, we cannot hope to program vehicles in a way that is consistent with an agreed upon vehicular morality. In the meantime, before fully autonomous cars become the norm, I suggest that there is another moral quandary to be explored that involves the intermediate step of so-called 'semi-autonomy.' I shall name this the '*Tesla Problem*.'

The Tesla Problem

It is a quirk of current autonomous car legislation that any steering decision is *overdetermined*. The most advanced road-legal autonomous vehicles today are made by Tesla. The vehicles are semi-autonomous in 'autopilot' mode, that is, they are capable of fully driverless navigation but are restricted from it by law. Hence the driver must keep their hands on the wheel at all times whilst the autopilot is engaged, both for safety, and to assume liability should something go wrong. At any moment the driver can retake control of the vehicle and override the autopilot. This gives rise to an interesting problem, hitherto unexplored: in an accident, the person in the driver's seat of a Tesla can, in an instant, be *either* the driver or a bystander, depending on if they act or not. The extreme proximity significantly blurs the line between the two roles. This, as I later show, has interesting repercussions for the *Bystander Problem* where a lack of proximity plays a significant role.

Suppose that you are cruising along a narrow mountain road, hemmed in by traffic barriers preventing you from careening off down the precipice, autopilot activated and driving for you. Ahead the road dips away, downhill. Although you're travelling at considerable speed, you are within the speed limit. As you crest the dip you're met with an unexpected sight: some hikers are dangerously crossing the road mere metres in front of you. Aghast, you realise that you cannot brake in time to avoid hitting them. Either you can continue in your lane, in which case you run five ramblers down, or you can swerve into the opposite lane (there is no oncoming traffic) and run down one. In short, we are faced with something close to a *Trolley Problem.*

In this situation, who are you: bystander, or driver? For Thomson, this matters, as some people intuitively "feel a difference between these two cases" because the driver both has a responsibility for anyone who is harmed through their actions and, in this case, they are *actively* killing five if they do not turn (Thomson 1985:1397). If, on the other hand, you are merely a bystander, you are simply "a private person who just happens to be there," who would personally do the five no harm if you were not to interfere with the vehicle (). Inaction would be tantamount to 'letting five die,' not equivalent to killing five. Consequently, there would be no reason you *ought* to act, unlike if you were the

can only follow moral rules.

driver. Your role in the Tesla is unclear because of the autopilot and your relationship to it, and because of your physical proximity to it all as a bystander – if indeed that is what you are. After all, in Thomson's original problem, the bystander is not in the cabin along with the trolley captain.

To begin, we shall assume that the autopilot is the driver. It is, of course, doing all the actual driving. You may be in the driver's seat, but you are not steering, accelerating or braking as the vehicle moves. Even if your hands are on the wheel, you could be asleep, or blind, and it would make no difference to the actions of the autopilot or the vehicle.

The question whether computers can have moral responsibilities is well beyond the scope of this paper. We can, however, assume there is a sense in which the autopilot makes rational, logical deductions about the environment around it in order to decide on courses of action. It is *controlling* the vehicle, for all intents and purposes. We can assume that it is certainly logically and physically possible for a sufficiently advanced computer to take on the role of driver. Then arguably that would make you a mere bystander to the whole affair.

For now, grant that you are indeed the bystander and the computer is the driver. Upon seeing an impending accident, Thomson says that "it would be permissible for you to *take* charge, *take* responsibility" and intervene despite the moral risk that this involves (Thomson 1985:1398). You do not have a moral *duty* to intervene, but it is certainly *permissible*, according to Thomson, to employ the distributive exemption and make what would kill five instead kill one person. You would be infringing upon no more rights in doing so than would have been infringed upon if you had not acted and, hence, this is morally acceptable.

However, this is where the *Tesla Problem* becomes interesting. If you do decide to act, then you take control of the vehicle. Unlike with trolleys, by choosing to act you are no longer a bystander, it seems that *you become the driver*. The person in the driver's seat, steering the car, is surely and unequivocally the driver. Equally, if you did not act, you would remain passive – by definition, a non-acting bystander. If you act at all, that is, take the wheel, no matter what role you had to begin with, you assume the role of the driver. In other words, if you act you cannot choose to remain a mere bystander. That is the first problem that your proximity as a bystander creates. The roles become even more puzzling the further we push them.

At this point, we can address an initial response. It is not clear that just because the autopilot is *driving* the vehicle, it is the *driver*. For instance, upon activating a more familiar feature like cruise control, we do not assume that the driver of the car has stopped being the human being in the driver's seat. Cruise control is merely an aid, automating some elements of driving to help the driver. Autopilot goes further in controlling almost all elements of driving but does this

mean it is more than simply an aid? The autopilot, some might respond, does not decide where to drive its passengers to. Choosing a destination remains the job of a human being and this makes the crucial difference. Yes, once a destination is set and has been programmed into the trip computer autopilot completely decides everything about the route you take but it does not perform the task of choosing destinations itself.

For comparison, let us examine the trolley driver as there are considerable similarities. He does not steer the trolley either – his trolley is on tracks. All he can control is the trolley's speed, rendering him little more than a human cruise control. Nevertheless, what sets him apart from being merely a cruise control, to my mind, is his role in choosing destinations for the trolley. In this regard he has more in common with the person in the Tesla's driver's seat than with its autopilot. We could push this further though and strip the driver of more responsibility. For instance, if the trolley also had an advanced automatic cruise control we might still plausibly think of the captain as the 'driver' despite having no real control over the trolley apart from setting station stops. Speed and direction would be beyond his control whilst in motion. The Tesla **'**driver**'** is in a similar position: they decide a route, program it into the navigation system and away the car drives. Save for this strategic planning they do nothing. Importantly, this tells us that we could also construe the autopilot as *not* being the driver – granting that our imagined trolley captain with an advanced cruise control is still worthy of the title 'driver' (although I admit that they quite plausibly are not).

Even granting the driver is the human being when autopilot is engaged, *prima facie* this changes nothing. It would still be morally permissible for them to act under the distributive exemption by taking the wheel and killing a person. A stronger claim than this, if we recall the trolley driver, is that the driver of the car *ought* to steer and kill the lone person; they have a moral duty to minimise the harm caused. Equally, if they shirked their moral duty to intervene and did not act, allowing the autopilot to steer, navigate and accelerate, they would be a non-acting bystander[77]. In either case the stringent rights of the pedestrian have not been infringed upon.

But now the moral quandary name in the title of this paper fully emerges. Again, imagine that you are driving along the mountain road, but this time you crest the rise to see that ahead it narrows to a single lane. Five hikers are dangerously crossing where the road is at its thinnest. No one is coming in the opposite direction. You also notice that there is a rather overweight straggler on the pavement, much closer to you and your car than the group is. Although

77 Of course, in this case the car, programmed in line with our moral consensus, would also employ the distributive exemption to swerve and kill the one. There would be no difference in the results, but different actions would have been taken.

the bystander is walking safely off the road, you realise that if you were to swerve and hit them, their mass would be sufficient to stop your car in time before hitting the five hikers further on. This is analogous to Thomson's footbridge example; we have already covered how the moral reaction people have to this situation differs from their reaction to the *Trolley Problem*: "Here, the consensus is that it is not okay to save five lives at the expense of one" by pushing the overweight straggler from the bridge, or, analogously, by running down the straggler on the footpath (Greene, 2007:42). The autopilot, programmed to follow this moral consensus, would not swerve either.

Abusing a person's rights either by pushing them from a bridge when they were not in danger or by mounting the curb and killing them when they were walking safely, is impermissible according to Thomson. Minimising death under the distributive exemption is only permissible if a person's stringent rights are not infringed which they would be if the person was not under threat from the car unless you mounted the pavement to hit them. Realising this we can assume that you would *not* act; stronger than this, you *should* not act. If you do not act though, you are certainly not the driver. The autopilot is. Thus, you must watch in horror – a mere bystander – as your car smashes through the five hikers crossing the road, killing them all. As the driver you would have had to be steering and in control of the car, which you were not.

Perhaps it could be argued that, on the contrary, you were in fact the driver as you had ultimate responsibility for the destination in a manner similar to the trolley driver with an advanced automatic cruise control. But this cannot be right; in moments your car will stop because of the accident, aborting your destination. Or, if this reasoning is rejected, then one might instead argue that your tacit acceptance of the autopilot's actions does not amount to a forfeiture of your role as driver. I believe this line of thinking cannot be accepted, however. If 'driving' requires only continued tacit acceptance of the actions and decisions made by the autopilot, then it becomes a completely passive process. This would make the person sat in the driver's seat utterly extraneous – and morally insignificant – to the decisions being made. If you always defered to the autopilot there's no one in the driver's seat at all and car proceeds on its own. Deference is not driving which requires the distinctly active role a driver guiding a vehicle. So, this line of thinking fails too as it would imply that you are *never* the driver when autopilot is engaged for you quite literally could be absent without making a moral difference. You are entirely extraneous, and so not the driver in any traditional sense of the word. So I will continue under the assumption that you are a bystander in this situation despite the fact you occupy the driver's seat.

Now we arrive at the crucial point: in this situation *there is no difference between your actions as the driver and the actions of a bystander*. For

a bystander too, it would be impermissible to breach a person's rights – you cannot throw them off the bridge and you cannot swerve to kill the straggler. You have a moral duty to remain a non-acting bystander. Indeed, once we realise this, we see that the role of driver collapses into the role of 'bystander who does not act' whenever Thomson's rights-based justifications enter the moral decision-making process and ensure you cannot act. And conversely, if rights are not infringed upon by your action, then it is permissible to act and you *ought* to act, thereby collapsing the role of acting bystander into that of the acting driver if either of them moves to save the five. *Proximity* destroys the distinction between the driver and non-acting bystander if their acting would infringe someone's rights. Put differently, it is impossible to be either an acting bystander or a driver who does not act, as the acting bystander collapses into the role of the acting driver and the non-acting driver collapses into the role of a non-acting bystander. This is illustrated by the following table:

	Acting Infringes Rights?	Is it Permissible?	Role Change if You Act?	Ought to Act?
You are the Driver	No	Yes	Remains the Driver	Yes
	Yes	No	Becomes a non-acting Bystander	No
You are a Bystander	No	Yes	Becomes the Driver	Yes
	Yes	No	Remains a non-acting Bystander	No

Having ascertained this, we see that when it comes to semi-autonomous vehicles, one can only act in a morally permissible way as either a non-acting bystander or acting driver, depending on the circumstances. First, if you are the driver it is certainly morally *permissible* to make what threatens the five here and now (your car) threaten the other person instead, but this is only true if all other things are equal. This means that we *can* actively swerve and kill the straggler in the first example but we *cannot* swerve and kill the lone walker in the second. We would be infringing upon a person's stringent rights in the latter case, so it would be morally impermissible. Consequently, if the person is safely walking on the pavement, then we must remain a non-acting bystander and resign ourselves to letting the car kill the five.

Secondly, and closely linked to this, we must notice your moral duty (or lack of it) to act in the various situations. If you are the driver, or would become the driver by acting, and no rights are being infringed upon, then many people, including Foot, would say that you have a moral duty to act. Thomson does not go quite so far but nevertheless agrees that if a driver faces a choice between killing the five or one, then, other things being equal, "perhaps he ought

to [kill the one]" under the distributive exception (Thomson, 1985:1415).

In the two versions of the *Tesla Problem* you *ought* to act when it is a) permissible to act and b) when no rights would be infringed upon through your acting as either the bystander or driver. When you are the driver, the widely shared intuition is that there is a moral duty to act and save the five, at least if the distributive exemption is valid and applies in this scenario. Thus the same holds for the bystander as I have shown that through their acting, they would become a driver and so would inherit the driver's moral duty to act. Proximity thus imposes a duty on the bystander that Thomson does not recognise. If it is a moral duty for the driver to act, then it must be for the bystander too.

Turning Back to Trolleys

These considerations of the role of proximity reflect new light on the distinctions in Thomson's original paper. Consider, for the final time, the original *Bystander Problem*. Thomson sees a clear moral difference between the two cases: the driver actively kills five if he does nothing whilst the bystander merely lets five die. This is *despite the fact that both driver and bystander, in their respective situations, have equal control over the tram* in question. Indeed, Thomson goes as far as to say that "the bystander will do the five no harm at all if he does not throw the switch", all but absolving him of moral responsibility (Thomson, 1985:1397). According to Thomson, the bystander *may* intervene but he has *no* moral duty to intervene. After all, if they do nothing, the bystander "merely fails to save them – he lets them die" but is not morally responsible for *killing* five people (Thomson, 1985:1398).

But when proximity brings the bystander physically closer to the driver – when the roles overlap as in the *Tesla Problem* – it becomes clear that the driver/bystander *must* act if the distributive exemption holds. Thus, if Thomson stands by her exemption, then she faces a difficult challenge of delineation: when does proximity no longer impose moral responsibility on the bystander? In effect, this raises, I believe, a dilemma in the manner of a sorites paradox[78].

Imagine that there was a passenger in the caboose of the runaway trolley talking with the trolley captain. Suddenly, the driver slumps at the controls and jams the accelerator. The passenger has only to reach out and turn the trolley to avoid the five in the now-familiar *Trolley Problem* situation. If the bystander in the autopiloted Tesla ought to take the wheel and drive, thus triggering a collapse of their role into the role of the driver, then it seems that this passenger must too, thus their role collapses into the driver's also. Repeat the example but move the passenger a foot further away from the trolley driver

78 The classic example of the *Sorites Paradox* involves a heap of sand, made up of millions of grains. If we remove one grain, the heap is surely still a heap. But if we repeat this we end up with only one grain of sand: when did the heap disappear?

each time. When is the passenger free of the moral duty to intervene like the bystander in the *Bystander Problem*? Are they actively killing five when seated right in front of the controls yet only letting five die when they are**,** say, two feet away?

Whilst I do not profess to know how Thomson would respond to this problem, my own opinion is that moral duty does *not* diminish with distance. Rather than the bystander 'letting five die' if they do not act, the *Tesla Problem* persuasively suggests that any bystander truly *ought* to act whenever they can – and if they do not, then they actively 'kill the five'. Otherwise the non-acting bystander in the driver's seat of the Tesla can shirk their moral responsibility even when they have their hands on the wheel and could have intervened with ease. If a driver did not act (regardless of the autopilot's actions), even if they would infringe upon no other rights by killing a person, we would certainly hold them responsible for not steering. They would have failed in their moral duty to minimise death by redistributing harm. If this is true, then drawing the line between the roles of bystander and driver is, I believe, almost impossible regardless of the distance between them and the driver so long as they can influence the direction of the tram/trolley/car. Proximity should mean nothing.

Conclusion

To conclude, unless you are *actively* steering the car, then you are a non-acting bystander in a semi-autonomous vehicle. However, because of your proximity to the driver the distinction between the roles collapses should you have to act. It follows from Thomson's paper that the person in the Tesla's driver's seat can assume the role of driver only if it does not infringe upon the pedestrian's rights. If stringent rights are infringed upon, driving is not permissible and so the driver must become merely a non-acting bystander. Further, if the driver ought to act, then the bystander must too if they can influence the vehicle's path. Thus, proximity strengthens the moral duty that the bystander is subject to. This demonstrates just how much of a concealed effect proximity has on the outcomes and the moral judgements for the moral actors of the original *Trolley Problem*, an effect which I believe has, so far, been overlooked. In effect, Thomson's original *Bystander Problem* incorrectly absolves the bystander of moral responsibility simply because of their distance from the incident. Ultimately, however, more work must be done to fully explore this '*Tesla Problem'* if moral duty is to be adequately understood in situations involving the autonomous or semi-autonomous vehicles of the future[79].

79 One area for further inquiry would be Thomson's more recent paper (2008) which incorporates self-sacrifice into the *Trolley Problem.* This, she believes, makes acting to save the five impermissible even when no stringent rights would be infringed upon. In an autonomous car, there is ample opportunity for moral problems to present

Bibliography

Dworkin, R. (1977). *Taking Rights Seriously*. Duckworth, 223-239.

Foot, P. (1967). *The Problem of Abortion and the Doctrine of Double Effect*. Oxford Review, 5, 5–15.

Greene, J. D. (2007). *The Secret Joke of Kant's Soul*. Moral Psychology: Historical and Contemporary Readings, 359-372.

Singh, S. (2015). *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey.* (No. DOT HS 812 115).

Tesla Motors (2016) *Autopilot: Full Self-Driving Hardware on All Cars*. Accessed 2016-10-21 at www.tesla.com/autopilot/.

Thomson, J. J. (1976). *Killing, Letting Die, and the Trolley Problem*. The Monist, 59(2), 204-217.

Thomson, J. J. (1985). *The Trolley Problem.* Yale Law Journal, 94(6), 1395-1415.

Thomson, J. J. (2008). *Turning the Trolley.* Philosophy & Public Affairs, 36(4), 359-374.

themselves involving potential self-sacrifice on the driver's part. How this would affect the findings of this paper would be an interesting avenue to pursue.