**The New Trolley Problem:  Driverless Cars and Deontological Distinctions**

**Abstract**

Discussion of the ethics of driverless cars has often focused on supposed real-life versions of the famous Trolley Problem. In these cases, a driverless car is in a position where crashing is unavoidable and all possible crashes risk harm: for example, it can either continue on its current path and crash into five pedestrians or swerve and crash into one pedestrian. There are significant disanalogies between the human versions of the Trolley Problem and situations faced by driverless cars which affect the application and moral significance of key deontological distinctions, such as the distinction between doing and allowing harm. The application and moral significance of the doing/allowing distinction in the context of the behaviour of driverless cars are not just currently unclear, but as yet to be determined. Nonetheless, Trolley Problems may still be useful in thinking about the ethics of driverless cars. Considering Trolley Problems can help us to negotiate how deontological distinctions should apply in this new context.

**Introduction**

Discussion of the ethics of driverless cars has often focused on supposed real-life versions of the famous Trolley Problem.[[1]](#footnote-2) In these cases, a driverless car is in a position where crashing is unavoidable and all possible crashes risk harm: for example, it can either continue on its current path and crash into five pedestrians or swerve and crash into one pedestrian.

A growing body of work criticises this approach to the ethics of driverless cars. Concerns include that driverless cars will never encounter trolley problem cases in the real world; that driverless cars will not be programmed with top-down ethical principles; and that there are other, more pressing ethical issues surrounding driverless cars from which trolley problems distract us.[[2]](#footnote-3)

Geoff Keeling provides convincing responses to two of these objections.[[3]](#footnote-4) As Keeling argues, consideration of trolley problems may be relevant to the ethics of driverless cars even if driverless cars will never encounter trolley problems and even if they will not be programmed top down from ethical principles. Trolley problems could still be used to help us understand which features might be morally relevant to the behaviour of driverless cars[[4]](#footnote-5) and to inform what Keeling describes as “a process of value-sensitive design: that is, a process in which engineers, moral philosophers and other stakeholders work together to determine the ethical implications of technical decisions made in the design-process”[[5]](#footnote-6). In response to the final objection, we can note that while there are clearly other interesting ethical issues surrounding driverless cars, the current discussion of trolley problems need not be a distraction from this. There is room for discussion of trolley problems and other ethical issues.

I will identify another significant disanalogy between the human versions of the Trolley Problem and situations faced by driverless cars. In the human case, the outcome depends upon the direct action (or inaction) of a human agent who is also the morally responsible agent.  In the driverless car case, we do not have a morally responsible agent who responds directly, but an autonomous machine. I will argue that this disanalogy has important implications for the use of Trolley Problems in the ethics of driverless cars.

As Keeling argues, if discussion of Trolley Problems is going to be helpful in the ethics of driverless cars, this is likely to be by helping us to identify which features of situations involving driverless cars matter morally and how.[[6]](#footnote-7) An important group of potentially morally relevant features are *deontological distinctions* like the distinction between doing and allowing harm. I think of a deontological distinction as a distinction between ways in which harm comes about, or the relationships between agents, victims and harms, which appears to matter morally even though it does not affect the severity or type of harm suffered. In this paper, I’ll focus on the distinction between doing and allowing harm.

I will argue that the disanalogies between the human versions of the Trolley Problem and situations faced by driverless cars affect the application and normative implications of the distinction between doing and allowing harm. In standard trolley cases, we use our understanding of the behaviour of the human agent and its relationship to the outcome to classify the case as doing or allowing harm and to inform our understanding of the normative significance of that behaviour. Without a human agent, we can expect either the classification of behaviour as doing or allowing to be different or for the normative implications of that classification to be different.

Indeed, I will argue that the application and moral significance of the doing/allowing distinction in the context of the behaviour of an autonomous machine are not just currently unclear, but as yet to be determined. They will depend upon yet-to-be-negotiated conventions about (a) our conception of the behaviour of driverless cars; (b) the background expectations of programmers/manufacturers/ owners of driverless cars and the conditions of being able to put those cars on the roads. Although my focus in this paper is on the doing/allowing distinction, I would expect similar issues with other deontological distinctions in the context of driverless cars.

Thus even if we had a clear uncontroversial account of which deontological distinctions matter and how in human trolley cases, we could not simply transfer over these deontological distinctions to situations involving driverless cars.[[7]](#footnote-8) Nonetheless, I will show that thinking about trolleys in the context of driverless cars may still be useful. But Trolley Problems must be used with appropriate awareness of their limitations and of the fact that in many ways we are not so much discovering how existing deontological distinctions apply to the behaviour of driverless cars as negotiating new deontological distinctions.

**The doing and allowing distinction and driverless cars**

In order to understand how the application and moral significance of the doing/allowing distinction may be different in the context of the behaviour of a driverless car, we first need to understand the application and moral significance of the distinction in standard cases.

According to common-sense morality, it can matter morally whether an agent has done harm or merely allowed harm, even if we hold equal all other factors, including the harm to the victim and how much it would have cost the agent to avoid the harm. For example, suppose that Bob has been bitten by a poisonous snake and will die without immediate treatment. Compare the two following scenarios:

Push: A boulder is in the middle of the road, blocking Bob’s route to the hospital. The boulder can only be moved out of the way by pushing it down the adjacent slope. Victor is trapped on the slope, directly in the boulder’s path. Freeing Victor or detouring around the boulder will delay Bob so he will not reach the hospital in time. Bob pushes the boulder. The boulder rolls down the slope, hitting Victor and crushing him to death.

Non-Interpose: The boulder is already rolling towards Victor. Bob could drive his car into the boulder’s path, bringing it to a halt. Doing so would delay him so he would not reach the hospital in time. He does not drive into the boulder’s path and the boulder hits Victor. [[8]](#footnote-9)

Intuitively, it is not morally permissible for Bob to push the boulder towards Victor in Push, but it is morally permissible for Bob to refuse to stop and interpose his car in Non-Interpose. [[9]](#footnote-10) There is a difference in permissibility even though the harm to Victor is the same in each case (he will be crushed to death by the boulder if it rolls down the hill) and the cost to Bob is the same in each case (he will die from the snakebite if he does not get to hospital in time). The difference seems to be because in Push Bob does harm to Victor while in Non-Interpose, Bob merely allows Victor to die. Doing harm is harder to justify than merely allowing harm.

Of course, the claim that there is a difference between doing and allowing harm needs some defence. Even if we intuitively judge cases involving doing and allowing harm differently, more argument is needed to show that we are right to do so. Such argument should explain what makes an agent count as doing harm rather than merely allowing harm and why this matters morally.

There are three different types of defence of the doing allowing distinction: (1) *Victim-centred defences* locate the moral difference between doing and allowing solely in the features of the victim; (2) *Agent-centred defences* locate the difference solely in the features of the agent; (3) *Dual-centred approaches* draw on both features of the agent and features of the victim to defend the moral relevance of the distinction.[[10]](#footnote-11)

My own defence of the moral relevance of the distinction between doing and allowing[[11]](#footnote-12) is a dual-centred approach. Partly for this reason, I’m going to use it as a model for discussion of how the differences between humans and driverless cars might affect the moral significance of the distinction between doing and allowing. When considering how the doing/allowing distinction will apply to cases involving driverless cars, for victim-centred accounts, we need to consider whether the relevant features of the victim are present in cases involving driverless cars; for agent-centred accounts, we need to consider whether the relevant features of the agent are present in cases involving driverless cars. For dual-centred accounts, we need to consider both these issues. So considering a dual-centred account helps us to understand how all three types of account may be affected by differences between human cases and driverless car cases.

I argue that for anything to genuinely belong to us, even our own bodies, we require both constraints against doing harm and permissions to allow harm. This is because the doing/allowing distinction lines up with differences in whether we impose on others: reaching beyond our own sphere into the sphere of others. On my analysis, an agent does harm when there is an unbroken chain of substantial facts leading from their behaviour to the harm. They merely allow harm when any such chain of facts linking them to the harm is broken by a non-substantial fact about something that belongs to the victim (or has been allocated to their use by an authorised third party). Substantial facts are the sorts of facts that are suitable to be part of the sequence leading to an outcome rather than mere conditions of the sequence continuing. A fact can be substantial by being positive – telling us that something is the case rather than that something is not the case – or by being contrary to the presuppositions of dialogues – the features of the world that are so taken for granted that they go without saying. Thus Bob counts as doing harm to Victor in Push because there is a chain of substantial facts linking Bob’s behaviour to Victor’s death: Bob pushes the boulder; the boulder rolls down the hill; the boulder hits Victor. In Non-Interpose, Bob counts as merely allowing harm to Victor. In Non-Interpose, Bob is only relevant to Victor’s death through the fact that his own car was not in the path of the boulder. This is a negative fact about something belonging to Bob. It is a mere condition for the sequence leading to Victor’s death.[[12]](#footnote-13)

I argue that in doing harm, the agent *causally imposes* on the victim. Causal imposition involves the behaviour of one person reaching into what belongs to another, substantially affecting what belongs to the victim against their wishes. In merely allowing harm, we do not causally impose on the victim. Similarly, constraints against allowing harm *normatively impose* on the agent. Normative imposition involves the needs of one person reaching into what belongs to another: requiring the agent to put their body or their belongings at the use of another. Constraints against doing harm do not as such involve normative imposition.[[13]](#footnote-14)

So, constraints against doing harm protect us against other agents’ causally imposing on us, substantially affecting what belongs to us against our wishes. Permissions to allow harm protect us against others’ normatively imposing on us, requiring us to put our body or our belongings at their use. For our bodies and other belongings to genuinely belong to us requires protection from both types of imposition. My laptop is not really my laptop if others are permitted to use it whenever they want – but it is also not really my laptop if I am required to use it to type out random people’s work whenever this would benefit them more than it would harm me. These permissions and constraints are prima facie, not absolute. My body and my resources can still genuinely belong to me even if it is sometimes permissible for others to interfere with them and even if I am sometimes required to put them at the use of others. For genuine belonging, what we need is *for the most part*, *in normal circumstances* that my body and other resources are at my use and not the use of others.[[14]](#footnote-15)

On this account, in order to work out the application and moral significance of the doing/allowing distinction to the behaviour of driverless cars, we need to answer two questions. First, how do the differences between driverless cars and human agents affect the constraints against doing harm: how do the protections against causal imposition apply to the behaviour of driverless cars? Second, how do the differences between driverless cars and human agents affect the permissions to allow harm: how do the protections against normative imposition apply to the behaviour of driverless cars? I will argue that neither of these questions has a clear answer. Indeed, I will suggest that we should see these answers as to-be-determined. We, as a society, must work out how to think about the behaviour of driverless cars.

Although the terms ‘protection against causal imposition’ and ‘protection against normative imposition’ are my own and play a specific role in my defence of the moral relevance of the distinction between doing and allowing, a lot of what I say in these terms can be generalised to other accounts. We can do this by hearing both ‘protections against casual imposition’ and ‘protections against normative imposition’ in a suitably wide way. ‘Protections against casual imposition’ can be heard as picking out whatever protections for potential victims are provided by the victim-based constraints against doing harm included in the moral significance of the doing/allowing distinction. Similarly, ‘protections against normative imposition’ can be heard as picking out whatever agent-based protections against being required to prevent harm are included.

**Victim-based Constraints Against Doings Harm**

Constraints against doing harm are needed to protect potential victims against harmful causal imposition. Such protection against causal imposition is needed to respect that the potential victims’ bodies and other belongings genuinely belong to them. In cases involving driverless cars, we do not have a human agent. However, we are still assuming that the potential victims are human. The bodies and belongings of these human potential victims belong to them. They will require protection against causal imposition to respect this.

However, it is not straightforward whether this protection against causal imposition involves constraints against having harm done to one by a driverless car. After all, the distinction between doing and allowing harm and its protections against imposition are primarily designed to protect us against the behaviour of human agents. These protections are not normally triggered when *objects* do harm to us: if I am drenched by a heavy rainfall or struck by lightning, this does not infringe my authority over my body. So we need to think about the moral significance of having an object like a driverless car do harm.

As Warren Quinn notes, constraints against doing harm can apply when a human being is seen as acting through an object. Quinn describes the following case:

Rescue III: a train is carrying the agent to rescue five people in imminent danger of death. The agent spots a person on the train tracks. He can pull the cord and stop the train, but by the time the man is freed it will be too late to rescue the five.[[15]](#footnote-16)

Quinn argues that letting the train continue in this case violates constraints against doing harm. Quinn states: “the combination of control and intention in Rescue III makes for a kind of complicity. Your choice to let the train continue forward is strategic and deliberate. Since you clearly would have it continue for the sake of the five, there is a sense in which, by deliberately not stopping it, you do have it continue.” [[16]](#footnote-17) According to Quinn, an agent will count as doing harm through an object when (a) the object is under the agent’s control; (b) the agent intends the harmful action of the object. [[17]](#footnote-18)

Quinn’s key insight that we can count as doing harm through objects is important, but I have some concerns about his analysis. First, it does not seem to make room for intentionally allowing an object to harm someone. Consider the following case:

Rolling Log: I see my enemy at the bottom of a steep hill. A log is rolling towards him, gathering speed as it goes. Although I could stop the log by placing a large stone in its path, I want my enemy to be hurt. I do nothing and the log hits my enemy, hurting him badly.

My behaviour in this case is certainly not as it should be. My malicious intentions reveal an immoral character. Depending on the seriousness of the harm and the amount of effort that would be required to interpose the rock, it may be morally impermissible for me to act as I do. However, it does not seem as if I am doing harm. If my behaviour is impermissible, it does not seem as if the Doctrine of Doing and Allowing has anything to do with this impermissibility.

The possibility of intentionally allowing an object to harm someone suggests that having an object under one’s control and intending it to harm the victim is not *sufficient* to count as doing harm through an object. My second concern with Quinn’s account is that this combination of control and intention doesn’t seem to be *necessary* for doing harm through an object either. Suppose that I am driving a car and it skids towards a pedestrian and I do not swerve. I do not intend the car to continue on its path, I am simply distracted or too scared to swerve. I suggest that we would still treat my behaviour as doing harm – even if I have merely refrained from acting. I am sufficiently identified with my car that it is treated as an extension of my body in this situation – I am seen as doing what it does. Constraints against doing harm protect us against what is done by objects like these. The mere fact that it is the car, and not the human agent, who acts doesn’t undermine the applicability of constraints against doing harm.

Vehicles designed to be driven by human drivers have become familiar. We are very used to assessing situations involving such vehicles. We have developed a clear – if not always clearly articulated - sense of how to understand the behaviour of such vehicles: we can recognise who is responsible for a vehicle’s behaviour and whether they should count as doing what it does.

Driverless cars are not familiar. These cars are not under the direct control of a human agent. It is not clear if we should understand the action of objects that can act autonomously – but where the agent has some control over the programming of the autonomous action – in the same way as the actions of objects which are directly under the agent’s control. Another complication is that, partly because the car is not under any agent’s direct control, it is not quite clear who should count as doing harm when the car does harm. Should it be the programmers of the car? The user? The owner? All or some of the above?[[18]](#footnote-19)

I suggest that how we should understand the behaviour of driverless cars is not merely unclear, but not-yet-determined. There may be no single correct way to conceive of the actions of a driverless car. Instead, we may need conventions which specify when an agent counts as identified with a driverless car and whether we treat the identified agent (or agents) as doing what the car does. There is an additional option in the driverless car case: we might treat the driverless car itself as an agent, whose behaviour triggers the constraints against doing harm *in its own right* even if there is no human agent who is thought of as doing harm when the car does harm.

These decisions may involve an extension or modification of the original victim-based constraint against doing harm. I originally described these constraints as protecting victims from other agents’ casually imposing on us. Suppose we see the driverless car as identified with some agent and treat the identified agent as doing what the driverless car does (in at least some cases). Arguably, we could see this as the agent causally imposing on the victim through the car. So we might think that this is covered by the original victim-based constraint against doing harm. But we might equally think that this is an extension of the original victim-based constraint against doing harm: it now protects us against causal imposition by other agents and by objects which other agents are seen as acting through. Similarly, if we treat the driverless car as an agent in its own right, we may well think this involves an extension of the original victim-based constraint against doing harm: it now protects us against causal imposition by other agents and by special objects which are conventionally treated as agents.

One might wonder whether such extensions to the protection against causal imposition are possible. This will vary depending on the details of the particular defence, but in the case of my defence, such extensions are possible. The ground of the protection against causal imposition is the need to respect that the victim’s body and other belongings genuinely belong to them. The original defence provides protection against other agent’s causally imposing on the victim, rather than from all objects’ causally imposing on them, because this wider protection would either give the victims a meaningless claim against objects like trees and rain clouds, or give them a claim that other agents protect them against imposition by trees and clouds. A claim for protection against other agents would make the victim’s authority extend beyond the victim’s body and belongings and onto the body and belongings of others. The proposed extensions do not result in meaningless claims or to extend the victims’ authority beyond their own body to the bodies of others. [[19]](#footnote-20)

When I say that there may be no single correct way to conceive of the actions of driverless cars, I am not suggesting that all possible conceptions of the behaviour of the driverless car are equally good. There will be reasons to conceive of the car in one way rather than another. Some of these reasons will be to do with the practical consequences and moral implications of treating cars in this way. Others will be conceptual reasons: Does this way of understanding the behaviour of a driverless car make sense? Does it stretch our concepts of action and agency too far? The existing understanding of the distinction between doing and allowing and of action-through-objects will place constraints on how we can classify cases involving the behaviour of driverless cars. Nonetheless, because driverless cars are significantly different from the objects which we are used to dealing with, we should not expect our existing concepts to give us a definitive answer about how to classify their behaviour.

I’ll finish off this discussion by discussing an obvious way of understanding the behaviour of driverless cars which I think should be rejected. I do not think we should resolve these complications simply by deciding that constraints against doing harm do not apply to the behaviour of driverless cars. Consider the following cases:

Hospital Trolley: A trolley is taking a passenger to hospital. There is one innocent person trapped in the track in front of the trolley. If the trolley driver stops the trolley this will delay the trip to hospital resulting in the passenger’s death. If the trolley driver does not stop the trolley, the innocent person in the track will be crushed to death.

Hospital Driverless Car: A driverless car is taking a passenger to hospital down a very narrow road. There is one innocent pedestrian trapped in the road in front of the driverless car. The driverless car can continue, driving over the pedestrian, or stop before hitting the pedestrian but cannot swerve to avoid the pedestrian. If the driverless car stops to avoid the innocent pedestrian, this will delay the trip to hospital resulting in the passenger’s death. If the driverless car does not stop, the innocent pedestrian will be crushed to death.

Stopping should be required in both cases. Indeed, any ethics of driverless cars that permitted crushing the innocent pedestrian would be utterly implausible. The best way of explanation of the requirement to stop in the Hospital Trolley case seems to be that continuing would be doing harm and that the innocent person on the track has protection against being harmed. Similarly, the simplest way for an account of the ethics of driverless cars to imply that it is impermissible to continue in the Hospital Driverless Car case is to see continuing as a kind of doing harm covered by victim-based constraints against causal imposition. Moreover, it seems as if the victim-based constraints against being harmed *should* protect the pedestrian in this case. We should not be happy at the idea of a driverless car that sees no constraint against running over pedestrians.[[20]](#footnote-21)

**Agent-Based Permissions to allow harm**

On my account, human agents need (prima facie) permissions to allow harm to respect that their bodies belong to them. Do constraints against driverless cars’ allowing harm put any agents’ body at the use of another? There are two really quite different cases here: in the first type of case, we have a driverless car with no passengers; in the second type of case, we have a driverless car containing passengers – who may or may not own the car.

Constraints against driverless cars without passengers allowing harm do not (in themselves) put any agents’ body at the use of another.

True, the programmers must *do* something to make the car act: they must ensure that the car is programmed to respond in the appropriate way. However, requiring the driverless car to do something does not translate to a requirement for additional agency on the part of the programmers. The programmers are already required to programme the car to react to situations of this kind in an ethical manner. Requiring them to programme the car to do X in a given situation rather than requiring them to programme the car not to do X in a given situation does not require additional agency. Of course, it might be that in a specific situation, it *is* technically much harder to programme the car to do something. But it could just as easily be true that it is harder to programme the car not to do a specific thing. There’s no conceptual connection between requirements to make the car do something and additional requirements for the programmers to act.

This means that in the cases of the driverless car without passengers, we do not have permissions to allow harm based on the need to protect agents from having to put their *bodies* at the use of others. In these cases, constraints against allowing harm require the use of resources other than the agent’s body. The permissions to allow harm by refusing the use of non-bodily resources are much weaker than the permissions to allow harm by refusing the use of one’s body. In addition, they tend to be much more vulnerable to conventions about the precise conditions under which certain types of resources are allowed to be owned and used in given society. Such conventions could easily include a system under which ownership of driverless cars depended on an expectation that driverless cars without passengers will be programmed to prevent road accidents whenever possible.

The most obvious place we’d expect the differences between human driven cars and driverless cars without passengers to make a difference are in cases where preventing harm is onerous or dangerous. So a human driver would not be expected to interpose themselves between two colliding cars. A driverless car without passengers might be. However, I think there might be a difference even in cases where the acting is not itself onerous. Compare the standard trolley problem to a case described by Geoff Keeling.

Standard Trolley Case: A runaway trolley, carrying no passengers, is headed towards five innocent persons. A bystander can switch the trolley to a side-track where another innocent person is trapped. Whoever is hit by the trolley will be killed.[[21]](#footnote-22)

Keeling’s Motorcycle Case: Suppose that a motorcyclist is skidding across the road towards a crowd of pedestrians on the pavement. The [driverless car] can brake, in which case the motorcyclist will skid into the pedestrians and cause their deaths. The [driverless car] could also accelerate into the motorcyclist, in which case the motorcyclist would be killed, but the skid would be deflected and the pedestrians would be unharmed.[[22]](#footnote-23)

Plausibly in the Standard Trolley Case, switching is permissible but not required. The constraint against doing harm to the single person is defeated by the potential to save five other lives. However, the agent retains the permission to allow harm.

In the human case, the permissibility of refusing to switch the trolley is related to the fact that switching requires the use of the agent’s body. Requirements to switch impose on the agent by demanding that they put their body at the use of another. In this case the use of the agent’s body is relatively minor – they are simply required to push a button or pull a lever. Indeed, if it were not for the presence of the one person on the track, then it would seem that the agent *would* be required to switch: if the trolley is headed towards five people on the main track and I can switch the trolley to an empty side track by pulling a lever, then it seems that I am required to do it.[[23]](#footnote-24) I suggest that in the standard trolley case, the permissibility of not switching stems from the fact that switching requires the use of the agent’s body *and* switching is both psychologically difficult and requires the agent to bear a potentially significant moral cost.

In contrast to the human case, I think that, in Keeling’s Motorcycle Case, it would be reasonable to require a driverless car without passengers to accelerate whenever it is permissible to do so. This may require certain other conventions to be in place i.e. general agreement that accelerating is permissible in such cases, and that the programmers/owners/users of driverless cars will not face prohibitive legal liability. Nonetheless, if these conditions are in place, then it does not seem as if the protections against having one’s non-bodily resources put at the use of others would necessarily block a requirement to accelerate.[[24]](#footnote-25)

A few notes about the above discussion. I have not discussed whether it is permissible to accelerate and kill the motorcyclist. That is a very complicated question. With respect to permissibility, it is worth noting that, like many thought experiments, this case assumes that we can know with certainty what the outcome of accelerating / not accelerating would be. In real life, it will not be certain either that the motorcyclist would kill the pedestrians if the driverless car does not interfere or that all pedestrians will be saved if the driverless car does interfere. So even if we did conclude that it would be permissible for the driverless car to accelerate in the case as described, real life uncertainties may mean it is never permissible to do so in real life. Nonetheless, as in this section my focus is on what can be required of the driverless car compared to what can be required of a human agent, I think we can use the thought experiment without addressing such complications.

I say that it would be reasonable to require a driverless car to accelerate whenever it is permissible to do so. Why do I say this rather than simply that the driverless car is required to accelerate? Here we come back to two features of the ethics of driverless cars that have already been noted. First, I have argued that requirements for driverless cars to prevent harm to others are in many ways as yet to be determined. What is required of a driverless car will depend on conventions surrounding the conditions under which driverless cars are permitted to be on the roads. Secondly, driverless cars are not themselves moral agents and are not strictly speaking subject to moral demands. Thus when I say that it could be reasonable to require a driverless car to accelerate, what I mean is that it is could be reasonable for the ethical guidelines or legal frameworks on the design of driverless cars to require driverless cars to be designed to behave in such a way that they would accelerate in the case described. I noted that such restrictions would depend upon other conventions beings in place. They might also depend on how technically difficult it is to design driverless cars to behave in this way. The key point, however, is that neither the owner nor the programmers of the driverless car could appeal to a permission to allow harm to block such restrictions. In contrast, the permissions to allow harm built into the distinction between doing and allowing do mean that ethical guidelines or legal frameworks requiring a human to turn the trolley (or to accelerate into the motorcyclist) would be illegitimate.[[25]](#footnote-26)

Things are further complicated in the second type of case – driverless cars with passengers. When a driverless car with passengers prevents harm, the bodies of the passengers are not themselves put at the use of others. But the passengers are encased within the non-human agent which is being put at the use of others. As it moves, it moves them. Instead of the familiar protections against being required to put one’s body at the use of others, we have protections against having one’s body put at risk of harm when contained within a resource that is being put at the use of others. What were protections against normative imposition are replaced by protections against a specific type of causal imposition. What were permissions to allow harm become constraints against others doing harm to one in particular ways. The driverless car version of the doing/allowing distinction appears to take on elements of the Doctrine of Double Effect (or whatever should replace that doctrine in explaining the difference between the Standard Trolley case and some variations of the Bridge Case).[[26]](#footnote-27)

**Implications for the Use of Trolley Problems in the Ethics of Autonomous Vehicles**

Deontological distinctions like the doing/allowing distinction do not apply in quite the same way to the behaviour of driverless cars as to human behaviour. Because of important differences in the driverless car situation, the normative significance of these deontological distinctions is not the same. Moreover, how these deontological distinctions apply is not fully determined by our existing concepts and practices. How victim-based constraints against doing harm apply to the behaviour of driverless cars depends upon how we conceive of the agency of the driverless cars. How agent-based permissions to allow harm apply depends on what costs and responsibilities programmers/manufacturers/ owners of driverless cars are required to take on as a condition of being able to put those cars on the roads. These still seem to be open questions and there may be different reasonable ways of answering them. It may also depend upon other yet to be determined factors such as whether we have a model in which cars are owned by individuals or whether individuals use a fleet of driverless cars all owned by a company or by society.

This means that we need to be very careful when applying deontological distinctions to driverless cars – and in considering our intuitions about cases involving driverless cars. More generally, recognising that deontological distinctions might not have the same normative significance in cases involving driverless cars raises an additional concern with the use of trolley problems in the ethics of driverless cars. Even if we had perfect solutions to all human trolley problems, we could not just transfer these solutions to cases involving driverless cars.

One option in working out how deontological distinctions function in driverless car cases would be to look at Truly Equivalent Trolley Problems. A Truly Equivalent Trolley Problem is a trolley problem which is adapted to match the ways in which driverless cars differ from classic trolley problems. In the literature on deontological distinctions, we sometimes speak of Equalised Contrast Cases.[[27]](#footnote-28) These are pairs of cases which are equalised so the only difference between them is that they fall on opposite side of the candidate distinction which is being tested for relevance, for example, if we were testing the relevance of the doing/allowing distinction, we would pick one doing harm and another merely allowing harm and make everything else equal. Truly Equivalent Trolley Problems are not equalised contrast cases. Indeed to produce a pair of equalised contrast cases for driverless cars, we would need to identify a candidate distinction that we wanted to test, then produce two Truly Equivalent Trolley Problems, one on each side of the distinction, each of which had been designed to match the ways in which the driverless car case differs from the standard human agent case. [[28]](#footnote-29)

However, it may be difficult to produce Truly Equivalent Trolley Problems. The cases may also end up being so complicated and unfamiliar that we no longer have firm intuitions about them. Moreover, our responses to those individual cases may depend on our assumptions about the conception of the agency of driverless cars and the normative background conditions that haven’t been settled yet. Even if I am wrong and there is some single reasonable answer to each of the questions raised above, these background conditions may well be different enough that we should not expect our intuitions to reflect them accurately.

Yet, I think Trolley Problems still have some role to play in the ethics of driverless cars. Some cases that may be helpful appear in this paper. I use the Hospital Cases to try to show that the doing/allowing distinction must play at least some role in the ethics of driverless cars. Versions of Geoff Keeling’s Motorcycle Case may be used to help us to think about what requirements for driverless cars to aid other road vehicles may be reasonable under different background conditions.

Driverless cars and other autonomous machines are likely to be a key part of our future. I’ve suggested that how the distinction between doing and allowing and other deontological distinctions apply to driverless cars depends upon conceptual and normative background conditions that are not yet settled. Moreover, these deontological distinctions may end up having very different normative significance when applied to the behaviour of driverless cars. Indeed, classic deontological distinctions such as the doing/allowing distinction are likely to have to incorporate elements of other deontological distinctions. Given these considerations, we might wonder whether it makes sense to see ourselves not as asking how our human deontological distinctions apply to the behaviour of driverless cars, but as in a period of negotiation of new deontological concepts.

1. The original trolley problem was put forward by Philippa Foot (Foot, “The Problem of Abortion and the Doctrine of the Double Effect”, *Oxford Review*, 1967. No. 5) but Judith Jarvis Thomson’s modification is perhaps better known (Thomson, “Killing, Letting Die and the Trolley Problem”, *The Monist*, Volume 59, Issue 2, April 1976). [↑](#footnote-ref-2)
2. Goodall, N. (2016). Away from trolley problems and toward risk management. *Applied Artificial Intelligence, 30*(8), 810–821; Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice, 19*(5), 1275–1289; Roff, H. (2017), The Folly of Trolleys: Ethical Challenges and Autonomous Vehicles, Brooking Report, 17th December 2018, <https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/>; Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice, 21,* 669–684. [↑](#footnote-ref-3)
3. Keeling, G. Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics* (2020) 26:293–307 https://doi.org/10.1007/s11948-019-00096-1 [↑](#footnote-ref-4)
4. Keeling, Why trolley problems matter for the ethics of automated vehicles, p. 296. [↑](#footnote-ref-5)
5. Keeling, Why trolley problems matter for the ethics of automated vehicles, p. 302. [↑](#footnote-ref-6)
6. Keeling, Why trolley problems matter for the ethics of automated vehicles, p. 296. [↑](#footnote-ref-7)
7. Importantly, my argument does not assume that the driverless car itself can apply deontological distinctions. I follow Keeling and assume that if trolley problems play a role in the ethics of driverless cars this would be through the process of value-sensitive design that Keeling discusses. On Keeling’s picture, we would use trolley cases to work out which deontological distinctions matter and recognition of these distinctions would then feed into the evaluation of potential automated vehicle control systems. My claim is that this process will not work if it assumes that deontological distinctions matter in the context of driverless car in the same way as they do in trolley problems with human agents. I thank Genia Schoenbaumsfeld and Brian McElwee for raising this. [↑](#footnote-ref-8)
8. Woollard, F. 2015. *Doing and Allowing Harm*. Oxford: Oxford University Press. p. 3. As noted there, these cases are inspired by Jonathan Bennett’s Push and Stayback (Bennett, J. 1995. *The Act Itself.* Oxford: Oxford University Press. 67). [↑](#footnote-ref-9)
9. Some people may think that it is permissible for Bob to push the boulder towards Victor because it is permissible to anything necessary to save one’s own life. I suggest that such people vary the potential harm to Bob. At some point, I think they will find a harm where it is intuitively permissible for Bob to allow Victor to die to avoid suffering that harm but it would not be permissible for Bob to kill Victor to avoid suffering that harm. [↑](#footnote-ref-10)
10. Philippa Foot and Warren Quinn both defend the doing/allowing distinction by arguing that negative rights (to non-interference) are stronger than positive rights (to aid). These are clearly both victim-centred defences. (Quinn, W. 1989, Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing, *Philosophical Review*, 98 (3): 287–312. Foot, The Problem of Abortion and the Doctrine of the Double Effect, *Oxford Review*, 1967. No. 5.) Samuel Scheffler argues that the distinction between doing and allowing is needed to be able to understand oneself as an agent subject to norms. This account is agent centred (Scheffler, S. 2004, Doing and Allowing, Ethics, 114 (2): 215–239). Some other accounts are a little trickier to classify Barry and Øverland’s defend the relevance of a tripartite distinction between doing, allowing and enabling in terms of significance of whether an agent has given rise to risk. This might be thought of as agent-centred. (Barry, C & Øverland, G. 2016, Responding to Global Poverty: Harm, Responsibility and Agency, Cambridge University Press, Cambridge, UK.) Frances Kamm’s account could be classified as either victim centred (if we focus on the centrality of inviolability to her account, or as dual centred, if we focus on the fact that she argues that one of the features of letting die that makes it more acceptable than killing is that the alternative is for the agent to be interfered with. (Kamm, F. 1996, *Morality, Mortality, Volume II*, Oxford Oxford University Press; 2007, *Intricate Ethics*, New York: Oxford University Press.) The fact that it is not always easy to classify accounts from the literature does not undermine the importance of the difference between the three approaches. As I will show, which approach is taken will affect what we have to think about to work out how the doing/allowing distinction will apply to the behaviour of driverless cars. [↑](#footnote-ref-11)
11. See Woollard, Doing and Allowing Harm (OUP, 2015); Woollard, “If This Is My Body… A Defence of the Doctrine of Doing and Allowing”, *Pacific Philosophical Quarterly*, Vol.94, No.3 (September 2013): 315-341. doi: 10.1111/papq.12002. [↑](#footnote-ref-12)
12. Woollard, *Doing and Allowing Harm*, Chapters 1-5. [↑](#footnote-ref-13)
13. Woollard, *Doing and Allowing Harm*, pp. 98-105. I say that ‘*in* allowing harm, the agent does not causally impose on the victim’ and ‘constraints against doing harm do not *as such* involve normative imposition.” This is intended to acknowledge that there may be cases when an agent allows harm and causally imposes on the victim or where a constraint against doing harm involves normative imposition, but that in such cases there must be some extra feature of the situation that makes it the case that the victim/ agent is imposed on. (See Woollard, F. *Doing and Allowing Harm*, p. 102, footnote 8). [↑](#footnote-ref-14)
14. Woollard, *Doing and Allowing Harm*, pp. 105-116. [↑](#footnote-ref-15)
15. Quinn, W. 1989. Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing. In: *The Philosophical Review*, Vol.98, No.3 (Jul., 1989): p. 298. [↑](#footnote-ref-16)
16. Quinn, Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing. p. 300. [↑](#footnote-ref-17)
17. Quinn, Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing. p. 298-301. [↑](#footnote-ref-18)
18. Another familiar type of case which has some important aspects in common with cases involving driverless cars is the case of an animal that belongs to a person. [↑](#footnote-ref-19)
19. I thank Alex Gregory for pressing me on this. [↑](#footnote-ref-20)
20. I thank Elselijn Kingma and Lee Walters for pressing me to clarify this discussion. [↑](#footnote-ref-21)
21. Thomson, “Killing, Letting Die and the Trolley Problem”, *The Monist*, Volume 59, Issue 2, April 1976. [↑](#footnote-ref-22)
22. Keeling, Why trolley problems matter for the ethics of automated vehicles, 297. [↑](#footnote-ref-23)
23. For discussion of how requirements to put one’s body at the use of another in limited one-off circumstances like the trolley case can be compatible with my defence of the moral relevance of the distinction between doing and allowing see Woollard, *Doing and Allowing Harm*, Chapters 7 & 8. [↑](#footnote-ref-24)
24. Keeling’s conclusion about this case is different. He argues: “Presumably, it is morally permissible for the [driverless car] to brake here. It is too demanding to suppose that the [driverless car] is morally required to intervene and kill the motorcyclist…. The salient point here is that properties other than harm seem to make a difference to the moral permissibility of the [driverless car]’s acts in ideal cases. It seems that whether or not the harm is done or merely allowed to happen is morally relevant.” (Keeling, Why trolley problems matter for the ethics of automated vehicles, 297.) I don’t think Keeling is quite right here. I agree that it would be permissible for a human driver to brake. However, it does not follow that it is permissible for a driverless car to brake. [↑](#footnote-ref-25)
25. I thank Alexander Geddes for pressing me on this. [↑](#footnote-ref-26)
26. In the Bridge Case, the runaway trolley is heading towards five innocent victims trapped on the track and the agent can push a large and heavy innocent person from a bridge into the path of the trolley, the innocent person is large and heavy enough to bring the trolley to a halt, but he would be killed in the process (Thomson, “Killing, Letting Die and the Trolley Problem”, *The Monist*, Volume 59, Issue 2, April 1976). We might get an analogy with the driverless car case by imagining a version of Bridge where the bystander is not heavy but is inside a large metal sphere and it is the weight of the sphere not the person that will stop the trolley. [↑](#footnote-ref-27)
27. See Kagan, S. 1988, The Additive Fallacy, *Ethics*, 99: 5–31. [↑](#footnote-ref-28)
28. I thank Giulia Felappi for pressing me on this. [↑](#footnote-ref-29)