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

Fiona Woollard

University of Southampton

[f.woollard@soton.ac.uk](mailto:f.woollard@soton.ac.uk)

***This is the peer reviewed version of the following article: The New Trolley Problem: Driverless Cars and Deontological Distinctions, which is, as of date of deposit, is forthcoming in final form at the Journal of Applied Philosophy. As the paper will be published open access, copyright remains with the author.***

***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 depend on (a) our conception of the behaviour of driverless cars; (b) the forms of driverless cars that are developed and used; (c) the background expectations of programmers/manufacturers/ owners of driverless cars and the conditions of being able to put those cars on the roads. These are as yet unsettled – and may even be undetermined. Nonetheless, trolley problems may still be useful in thinking about the ethics of driverless cars.

**Keywords: Driverless cars, autonomous vehicles, trolley problems, deontological distinctions, doctrine of doing and allowing.**

***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 the most famous trolley problem case, the agent must choose whether to turn a runaway trolley that is currently speeding toward five innocent people onto a side-track where it will hit one innocent person. In the supposed real-life trolley 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 real driverless cars will never encounter trolley problem cases; that driverless cars will not be programmed with top-down ethical principles; and that there are other, more pressing ethical issues surrounding driverless cars.[[2]](#footnote-3)

Geoff Keeling provides convincing responses to two of these objections.[[3]](#footnote-4) As Keeling argues, consideration of trolley problem cases may be relevant to the ethics of driverless cars even if driverless cars will never encounter trolley problem cases and even if they will not be programmed top down from ethical principles. Trolley problem cases 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 “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, there is room for discussion of trolley problems and other ethical issues.

I will identify another significant disanalogy between the human trolley problem cases and situations faced by driverless cars. In the human case, the outcome depends upon the direct action (or inaction) of a human agent.  In the driverless car case, the direct action (or inaction) is from an autonomous machine. On what we might call the traditional view, driverless cars are not moral agents[[6]](#footnote-7) (Purves et al 2015; Talbot et al 2017). On other views, at least some driverless cars are moral agents in at least some sense, but still differ in significant ways from human moral agents (List 2021; Tiggard 2021a; 2021b). I will argue that this disanalogy has important implications for the use of trolley problem cases in the ethics of driverless cars.

The term ‘trolley problem’ is used in various ways in the literature. It may refer to the question of what to do in the famous trolley cases: should the agent turn the runaway trolley away from five innocent people towards one innocent person? It may also refer to the problem of explaining why it is permissible to turn the trolley in the famous trolley cases given that it is often impermissible to kill one to save five: for example, given that it is impermissible to stop a runaway trolley from hitting five by pushing one large person into the path of the trolley (Thomson 1976). However, we can also use the term ‘trolley problem’ more widely to refer to the use of idealised cases to identify morally relevant features. [[7]](#footnote-8) Keeling’s defence of the use of trolley problems supports this wide understanding of what a trolley problem is. Keeling argues that trolley problems are likely to be helpful in identifying which features of situations involving driverless cars matter morally and how (2020, 296). There is no reason to think that the only morally relevant feature is whatever explains the permissibility of turning the trolley in the original trolley case. I will be using the wide understanding of ‘trolley problem’.

An important group of potentially morally relevant features are *deontological distinctions*. I think of a deontological distinction as a distinction between how agents, victims and harms are related, which appears to matter morally even though it does not affect the severity or type of harm suffered.[[8]](#footnote-9) I’ll focus on the distinction between doing and allowing harm.[[9]](#footnote-10)

Although human agents are involved at various points in the design and use of driverless cars, the autonomous machine is the immediate agent. It is the autonomous machine that implements the decision to e.g. swerve or not to swerve. So, considering a deontological distinction in driverless car cases leads to a new question: what is the moral relevance of this distinction applying to the behaviour of an autonomous machine? In the case of the doing/allowing distinction, the new question is: what is the moral relevance of the fact that *an autonomous machine* does harm rather than merely allowing harm?

The relationship of the human agents who are involved to the harm is different in the human trolley case and the driverless car trolley case. So, another new question might be: what is the moral relevance of the fact that a human agent *designs an autonomous machine* to do harm rather than merely allow harm? We cannot simply assume without argument that this has same moral relevance as the fact that a human agent does harm rather than merely allowing harm. I take this second new question to be an aspect of the first new question, which encompasses how the difference between a driverless car’s doing vs allowing harm may have moral significance not just for designers, but also for owners, users, and regulators.

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. When we do not have a human agent in the relevant role, either the classification of behaviour as doing or allowing or the normative implications of that classification may be different.

I will argue that the application and moral significance of the doing/allowing distinction in the context of the behaviour of an autonomous machine will depend (a) our conception of the behaviour of driverless cars; (b) the forms of driverless cars that are developed and used; (c) the background expectations of programmers/manufacturers/ owners of driverless cars and the conditions of being able to put those cars on the roads. There are a number of relevant discussions that are ongoing here, for example the aforementioned discussion about what, if any, agency can be attributed to autonomous machines. At the very least, the application of the doing/allowing distinction to driverless cars depends on which views one holds in these as-yet-unsettled debates. But it may be that truths in action theory and/or ethics leave how to apply the doing/allowing distinction underdetermined. A definitive answer may require yet-to-be-negotiated conventions. Although my focus here 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.[[10]](#footnote-11)

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, of the dependence on how other debates turn out, and of the fact it may end up that 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 a human 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 human 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 (Woollard 2015, 3). [[11]](#footnote-12)

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. [[12]](#footnote-13) 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.[[13]](#footnote-14)

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.[[14]](#footnote-15)

I will base my discussion on Fiona Woollard’s defence of the moral relevance of the distinction between doing and allowing (2013, 2015). This is the only existing book length analysis and defence of this distinction. Partly for this reason, and partly because it is a dual-centred approach, 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. As will be seen, Woollard’s account provides us with a framework for thinking about the application of the doing/allowing distinction that can be generalised to other accounts. Considering Woollard’s 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.

Woollard argues 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 Woollard’s 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 (Woollard 2015, chapters 1-5).

Woollard argues 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. (Woollard 2015, 98-105.)[[15]](#footnote-16)

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 (Woollard 2015, 105-116).

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?

Although the terms ‘protection against causal imposition’ and ‘protection against normative imposition’ are Woollard’s own and play a specific role in her 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 protect potential victims against harmful causal imposition. On Woollard’s account, such protection against causal imposition is needed to respect that the potential victims’ bodies and other resources genuinely belong to them. In cases involving driverless cars, we do not have a human agent. However, often the potential victims are human.[[16]](#footnote-17) The bodies and belongings of these human potential victims belong to them. On Woollard’s account, they require protection against causal imposition to respect this.

Other dual-centred or victim-centred accounts of the doing/allowing distinction may appeal to other features of the human potential victim to ground constraints against doing harm. It seems likely that whatever these features are, the human potential victim will still possess these features and still require protection against casual imposition in cases involving driverless cars. What I say in this section will apply to both Woollard’s account and to most dual-centred or victim-centred accounts.

Human potential victims require protection against causal imposition. 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 mere objectsdo harm to us: if I am drenched by a heavy rainfall or struck by lightning, this does not infringe my rights against casual imposition. So we need to think about the moral significance of having something like a driverless car – which is neither precisely like a human agent nor precisely like a heavy rainfall- do harm.

Talbot et al (2017, 258) describe a “continuum of agency” running from events “caused by the intentional action of some robust moral agent, like a human being” to natural disasters like earthquakes. They themselves hold that autonomous machines “fall closer to the proper natural disaster end of the continuum than the agential end” (258). They argue that autonomous machines are not agents on the following grounds: agency requires intentional action; intentional action requires acting for a reason; acting for a reason requires an attitude of belief or desire (or some further propositional attitude); these attitudes require phenomenological consciousness which autonomous machines – at least for the immediate future – lack (259).

Talbot et al argue that as autonomous machines are not agents, they cannot be subject to praise or blame, but the consequences of their actions can be morally good or bad. As the only moral claims that apply to them are about the goodness or badness of the outcomes they cause, the best state of the world is one in which autonomous machines act like maximising consequentialists. Nonetheless, they argue that it does not immediately follow that we should make autonomous machines that act as consequentialists. They seem to suggest that making an autonomous machine behave in a way that violates a right is equivalent to violating that right yourself.[[17]](#footnote-18)

Talbot et al’s discussion focuses on cases where a human deliberately programmes a machine to violate rights. It is not clear whether their views would be the same about *failure* to programme a machine *not* to violate rights. It does seem as if constraints against doing harm can apply when a human agent is seen as acting through an object when they fail to prevent the object’s action. Warren 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 (Quinn 1989, 298).

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” (1989, 300). 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 (1989, 298-301).

Quinn’s key insight that we can count as doing harm through objects, even when inactive ourselves, is important. 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.

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 this impermissibility is due to a constraint against doing harm.

The possibility of intentionally allowing an object to harm someone suggests that having control over an object and intending it to harm the victim is not *sufficient* to count as doing harm through an object. 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, that acts doesn’t undermine the applicability of constraints against doing harm.[[18]](#footnote-19)

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. No one has precisely the same relationship to the behaviour of a driverless car as the driver of a standard vehicle has to its behaviour. We do not yet have a settled understanding of how we should understand the behaviour of driverless cars.[[19]](#footnote-20)

Indeed, if Talbot et al are right that driverless cars are not agents, it may be that there is no single correct way to conceive of the actions of a driverless car. 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 hold that the driverless car is not an agent, but nonetheless decide to 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. Woollard originally describes 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 Woollard’s 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 extend the victims’ authority beyond their own body to the bodies of others.

Importantly, even if driverless cars are not moral agents, that does not mean that we are mistaken if we decide to treat them as if they were and to apply the constraint against doing harm to their behaviour on that basis. This is a legitimate extension of our current moral practices. In my view, once such a practice is in place, then the deontological constraint does apply. As I write, the constraint against doing harm applies to the behaviour of traditional cars under the control of human drivers. I count as doing harm if my car does harm *even if I perform no action with my body*. This may be an extension of ‘natural’ applications of the doing/allowing distinction, albeit one that is now so well embedded that we do not notice it. If it is based partly on convention, this does not undermine the correctness of our application of the constraint.[[20]](#footnote-21)

Even if there is no single correct way to conceive of the actions of driverless cars, this does not imply all possible conceptions of the behaviour of the driverless car are equally good. There are reasons to conceive of the car in one way rather than another. Some of these reasons are to do with the practical consequences and moral implications of treating cars in this way. Others are 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?[[21]](#footnote-22) 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.[[22]](#footnote-23) Nonetheless, because driverless cars are significantly different from the objects which we are used to dealing with, we should not necessarily expect our existing concepts to give us a definitive answer about how to classify their behaviour.[[23]](#footnote-24)

So far, I have argued that, even if we agree with Talbot et al that driverless cars are not agents, this does not settle how we should understand their behaviour. But not everyone agrees with Talbot et al. Some hold that driverless cars might count as agents and even in some sense moral agents. For example, Christian List (2021) uses a functionalist account of moral agency used in philosophy of group agency to explain how e.g. corporations can be morally responsibly. Contra Talbot et al, List argues not only that driverless cars have beliefs and desires and count as intentional agents (1219), but that driverless cars could feasibly be designed to make moral judgments and count as morally responsible agents (1230). List does recognise some important differences between such artificial intelligent systems and human moral agents (1234). We will discuss these later in our consideration of agent-based permissions. If List’s view is right, it seems that victim-based constraints against doing harm will apply straightforwardly to the behaviour of driverless cars.

Floridi and Sanders (2004) argue that autonomous machines can be moral agents and moral patients at a relevant level of abstraction. Theirs is an extremely deflationary understanding of moral agent, where an agent is said to be a moral agent if and only if it is capable of morally qualifiable action (364) and an action is morally quantifiable so long as it has outcomes we care about. They hold that a Web-bot’s actions are morally quantified ‘since we value our email’(370). They hold that as moral agents, autonomous machines are accountable (371). But they do not claim that AI can be morally responsible: “To be also morally responsible for x, the

agent needs to show the right intentional states…” (371). I take it that holding that a driverless car is an agent in this very deflationary sense does not settle whether constraints against doing harm should apply to the behaviour of driverless cars.

Tigard (2021a and 2021b) argues that “[Artificial moral agents] represent an outlying specimen, a deviation from our standard conceptual map of moral agency and patiency” (438) but emphasise the “priority and adaptability of our moral attitudes and practices” (435). Add to this - Nyholm: Some very useful material on how our ethical systems evolved for situations without robots so we shouldn’t expect (p. 35), existential choice (36) and flexible concept of agency (36), Nyholm (2020, 31) describes deciding how to understand the agency of robots is an ‘existential problem’: there is not always any clear and necessary answer to whether we should think of certain robots as agent. Both these views are similar to the view I discussed above, on which we may decide to extend our current moral practices to recognise driverless cars as agents – and thus also, presumably to recognise their behaviour as covered by the constraint against doing harm.

How we should understand the agency of driverless cars may depend upon the forms of autonomous machines that are developed and used. Discussion of the ‘responsibility gap’ and the need for ‘meaningful control’ are relevant here. The use of autonomous systems, particularly autonomous weapons, is said to lead to an objectionable ‘responsibility gap’: it is claimed that the autonomous machine is not the kind of thing that can be morally responsible while the human agents lack the kind of control required for responsibility (Matthias 2004; Sparrow 2007). Nyholm argues that we can – and should - avoid the responsibility gap by designing autonomous machines so that we retain the right kind of control over them (70-71). If humans retain the ability adjust the settings and to switch off the autonomous machine, then such machines should be thought of as exercising supervised and deferential collaborative agency rather than individual agency (62-64).[[24]](#footnote-25) Nyholm suggests we draw on hierarchical models of collaborative agency to hold humans responsible in such cases (65-71). If Nyholm is right, then we may be able to draw on our understanding of how constraints against doing harm apply in similar models of collaborative agency to inform how we apply constraints against doing harm to the behaviour of driverless cars. However, this will depend upon whether all autonomous machines are in fact designed to fulfil the conditions Nyholm identifies. [[25]](#footnote-26)

I’ll finish 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.[[26]](#footnote-27)

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 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.

**Agent-Based Permissions to allow harm**

Permissions to allow harm are needed to protect the agent against normative imposition. On Woollard’s account normative impositions are requirements for the agent to put their body or belongings at the use of another in a way that undermines their authority over what belongs to them. However, as before, ‘protections against normative imposition’ can be heard as picking out whatever agent-based protections against being required to prevent harm are included.

There are two really quite different cases where we can ask if a driverless car should be permitted to allow harm: 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. Let us start with the first type of case.

Consider the following case described by Geoff Keeling:

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 (Keeling 2020, 297).

Let us assume that the driverless car has no passengers.

Keeling 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.” (Keeling 2020, 297.)

Keeling only discusses this case very briefly. His main point is that properties other than harm seem to make a difference to the permissibility of the driverless car’s behaviour. I agree with this point: the motorcycle case is clearly morally distinct from cases where the driverless car could either swerve into one pedestrian or hit five pedestrians. Nonetheless, I think Keeling is wrong to conclude too quickly that it would be too demanding to require the driverless car to intervene.

Keeling’s framing of the requirement as too *demanding* suggests that our concern is agent-centred. Suppose we thought that it would be wrong for the driverless car to intervene because of constraints against harming the motorcyclist. It would follow that we could not require the driverless car to kill the motorcyclist. However, it would be strange to convey this by saying that it would be too demanding to suppose the driverless car must intervene. In this section my focus is on what can be required with respect to the driverless car compared to what can be required of a human agent. I am going to put aside the complex question of whether it would be permissible to kill the motorcyclist. It does seem that even if it were permissible to kill the motorcyclist, it would be too demanding to require a human agent to do so. Would it also be too demanding to require a driverless car to intervene if it were permissible to do so?

Unless driverless cars are themselves moral agents, they are not strictly speaking subject to moral demands. So, unless driverless car are moral agents, when we talk about what can be required of the driverless car, strictly speaking we are talking about what requirements ethical guidelines or legal frameworks on the design of driverless cars could reasonably place on how driverless cars are designed to behave.

There are some situations where a requirement for a driverless car to intervene may be too demanding on relevant human agents. Suppose that it is highly contested whether intervening is morally or legally permissible. The programmer, owner, or user of a driverless car that intervened may be held morally or legally responsible for an unjust killing. This risk may be enough to make requirements for the driverless car to intervene too demanding on the programmers, owners, or users. Similarly, such requirements might be too demanding for the programmers because it might be too technically difficult to make a driverless car intervene in this case without unwanted behaviour in other cases.

However, suppose that there is general agreement that intervening is permissible, protection against legal liability, and there are not too many technical challenges. It does not seem to be overly demanding on the programmers, owners, or users to require that, as a condition of the driverless car being allowed onto the road, driverless cars are designed so that they minimise traffic deaths whenever possible.

Woollard’s account predicts that there will be a difference between human cases and driverless car cases. On Woollard’s account, human agents need (prima facie) permissions to allow harm to respect that their bodies belong to them. Woollard does not hold that agents must be permitted to allow harm no matter what. She argues that we are required to make substantial sacrifices to prevent harm when personally involved in an emergency. Nonetheless, these requirements must be strictly limited and must leave our bodies for the most part at our own use and not at the use of others (Woollard 2015, Chapters 7&8).

It does not seem as if a driverless car’s body belongs to it in the same way that a human agent’s body belongs to them. List argues that autonomous machines can have legal and moral rights in the same way that a university may have legal and moral rights (1234-5). However, these rights are strictly limited. Autonomous machines do not matter intrinsically because they lack the capacity for phenomenal consciousness (1236). They can therefore have only derivative moral rights rather than rights for their own sake. What rights they have is determined by whether their having these rights is in the interests of those who do matter intrinsically (1235). List’s position is highly plausible. If it is correct, whether a requirement to aid would conflict with the body-ownership rights of a driverless car depends on whether – and what kind – of body-ownership rights the driverless car has. This in turn depends on what kind of body-ownership rights it is best for driverless cars to have – in the sense of being in the best interests of those who matter intrinsically. It is not clear why it would be in the interests of those who matter intrinsically for a driverless car to have body-ownership rights that conflict with a requirement to minimise traffic death where possible.

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

So, on Woollard’s account, in the cases of the driverless car without passengers, we do not have permissions to allow harm based on the need to protect human 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 human 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 include a system such as that described above, which required the driverless car without passengers to intervene in cases like Keeling’s.

Let’s turn to the second type of case: driverless car cases with passengers. We can adjust Keeling’s example so that preventing the crash would not require the car to kill a motorcyclist, but to interpose itself, placing the passenger that it is carrying at risk. Here I think there is a strong intuition that it would be too demanding to require the passengers to sacrifice themselves in this way. But we need to take care with our intuitions. Reflection shows that we are not in fact asking the passengers to sacrifice themselves. The passengers are not required to do anything – at least not during the interaction itself. The case is different from the human case where the agent is required to interpose the car (which they are inside) in the path of the crash.[[27]](#footnote-28)

To apply Woollard’s account: 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 actor 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 narrow trolley problem).[[28]](#footnote-29)

The above discussion appealed to specifics of Woollard’s account of agent-based permissions to do harm and List’s account of the rights of autonomous machines. Other agent-centred or dual-centred accounts may give a different defence of agent-based permissions to allow harm. Alternative positions may be taken on the rights of autonomous machines. However, it seems that any plausible account will recognise differences between: (1) requirements for a human agent to perform a costly action for the sake of others and requirements for resources belonging to a human agent to be used for the sake of others; (2) human body-ownership and any body-ownership of a driverless car; (3) sacrificing oneself to save others and being harmed when contained within a resource that is used to save others.

**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 depends on as-yet-unsettled issues in our conception of driverless cars and may not even be 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 the conventions in place including 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. It may also depend on what rights it is best (i.e. in the best interests of those who matter intrinsically) for driverless cars to have. These still seem to be open questions and there may be different reasonable ways of answering them.

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.

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 questions that are not yet settled – and where the answers may be underdetermined by current practice. 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 is in Philippa Foot. 1967. “The Problem of Abortion and the Doctrine of the Double Effect.” *Oxford Review* 5: 3. In Foot’s case, the agent is the driver of the runaway trolley. Judith Jarvis Thomson’s modification with agent as a bystander is better known (Thomson, Judith Jarvis. 1976. “Killing, Letting Die and the Trolley Problem.” *The Monist* 59. 2: 207. [↑](#footnote-ref-2)
2. Goodall, Noah J. 2016. “Away from trolley problems and toward risk management.” *Applied Artificial Intelligence, 30*(8): 810–821; Nyholm, Sven. Smids, Jilles. 2016. “The ethics of accident-algorithms for self-driving cars: An applied trolley problem?” *Ethical Theory and Moral Practice,* 19(5): 1275–1289; Roff, Heather. 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, Johannes. 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, Geoff. 2020. “Why trolley problems matter for the ethics of automated vehicles.” *Science and Engineering Ethics* 26:293–307 <https://doi.org/10.1007/s11948-019-00096-1> [↑](#footnote-ref-4)
4. Keeling Op.Cit. 296. [↑](#footnote-ref-5)
5. Keeling Op.Cit. 302.The concept of value sensitive design was originally put forward by Batya Friedman and Peter Kahn and colleagues (Friedman, B., Kahn, P.H., and Borning, A. 2002. “Value sensitive design: Theory and methods”. Technical Report 02-12-01, University of Washington, Computer Science and Engineering; B. Friedman, P. H. Kahn, A. Borning, A. Huldtgren. ’Value Sensitive Design and Information Systems.’ In N. Doorn, D. Schuurbiers, I. van de Poel, M. E. Gorman (eds.) Early engagement and new technologies: Opening up the laboratory (Springer Dordrecht, 2013), pp. 55-95; Friedman, Batya and Hendry, David. 2019. Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA. MIT Press.) It has been further developed by philosophers of technology see for example, van den Hoven, Jeroen, Vermaas, Pieter E. , van de Poel, Ibo. 2015, Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains. Dordrecht: Springer ; I. van de Poel, “Translating values into design requirements.” In D. Michelfelder, N. McCarthy, & D. Goldberg (Eds.), Philosophy and engineering: Reflections on practice, principles and process (Dordrecht: Springer, 2013), pp. 253–266. [↑](#footnote-ref-6)
6. Purves, Duncan, Jenkins, Ryan., & Strawser, Bradley. J. 2015. “Autonomous Machines, Moral Judgment, and Acting for the Right Reasons.” Ethical Theory and Moral Practice, 18(4): 851–872. <http://www.jstor.org/stable/24478765>; B. Talbot, R. Jenkins, D. Purves, ‘When Robots Should Do The Wrong Thing.’ In P. Lin, K. Abney, R. Jenkins (eds.) Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence (Oxford: Oxford University Press, 2017), pp. 258-273. [↑](#footnote-ref-7)
7. See Kamm 2016, 12-13 for discussion of various ways of understanding what the trolley problem is. [↑](#footnote-ref-8)
8. For a similar understanding of deontological distinctions see Woollard 2021, 142. Frances Kamm describes deontology as a theory “in which what we do is not determined by what we will bring about” (Kamm 2007, 12) i.e. in which how we bring about a given consequence matters, and later uses a case in which killing is wrong, while merely letting die is permissible, despite the cost to the agent being the same to argue that “there is a fundamental moral difference between killing and letting die” (Kamm 2007, 18). [↑](#footnote-ref-9)
9. The moral relevance of the doing/allowing distinction has been widely discussed. To take just a few examples, see Foot (1967); Quinn (1989); Bennett (1995); Kamm (1996, 2007); Barry and Øverland (2016). [↑](#footnote-ref-10)
10. My argument does not assume that the driverless car itself can apply deontological distinctions. I hold that trolley problems could play a role in the ethics of driverless cars through the process of value-sensitive design described earlier. 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 human trolley problems. Basl and Behrends (forthcoming in 2022) offer a different argument that the answer to human trolley problems do not transfer directly to answers about how we should programme driverless cars to behave. They argue that given that driverless cars will use machine learning (9-12), that this will mean that choices have to be made about what to focus on when training the system (14-15), and that training to behave in the way a human should behave in a trolley cases may require us to allocate substantial resources in a way that results in costs elsewhere in the system (15-16). They offer a solution similar to Keelings, arguing that we should use the trolley problems to identify moral relevant considerations which programmers ought to take seriously when deciding on a training regime (17). [↑](#footnote-ref-11)
11. As noted by Woollard, these cases are inspired by Jonathan Bennett’s Push and Stayback (Bennett 1995, 67). [↑](#footnote-ref-12)
12. 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. Such people should vary the potential harm to Bob. At some point, they should find a harm where it is intuitively permissible for Bob to allow Victor to die to avoid suffering that harm but impermissible for Bob to kill Victor to avoid suffering that harm. [↑](#footnote-ref-13)
13. Hübner and White (2018) argue that the doing/allowing distinction is “highly essential to legal discourse” and must be taken into account in the programming of driverless cars. “Any rule insensitive to this differentiation is thus likely to contradict the most basic standards of jurisprudential thinking” (695). The claim that our intuitions reflect the doing/allowing distinction has been challenged by e.g. James Rachels (1975) and Michael Tooley (1994). For responses, see e.g. Frowe (2007), Kamm (2007, 17); Trammell (1975). For more recent discussion of the role of the doing/ allowing discussion in commonsense morality, see Persson (2013, 69-94) and Woollard (2015, 4). [↑](#footnote-ref-14)
14. Philippa Foot (1967) and Warren Quinn (1989) both give victim-centred defences of the doing/allowing distinction, arguing that negative rights (to non-interference) are stronger than positive rights (to aid). These are victim-centred defences. Samuel Scheffler (2004) gives an agent centred defence, arguing that the distinction is needed to be able to understand oneself as an agent subject to norms. Some other accounts are trickier to classify. Barry and Øverland (2016) 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 count as agent-centred. Frances Kamm (1996, 2007)’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 letting die is more acceptable than killing in part because the alternative is for the agent to be imposed upon. Difficulty in classifying some accounts 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-15)
15. Woollard says 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 acknowledges that an agent may allow harm *and* causally imposes on the victim or a constraint against doing harm may involve 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 (Woollard 2015, p. 102, footnote 8). [↑](#footnote-ref-16)
16. There are likely to be cases involving driverless cars and animal victims. For discussion of how Woollard’s account of the doctrine of doing and allowing applies to animals see Woollard (2021). [↑](#footnote-ref-17)
17. They say: “It is difficult to see designing and deploying [a machine that will kill Agatha] as less wrong than simply killing Agatha yourself since there is a clear connection between the intentional action of the programmer and the death of Agatha” (262). Later, they refer to a consequentialist robot as “violating duties” (263). They later argue that it is permissible to build a consequentialist robot even if deontology is true under certain types of uncertainty (264-269). [↑](#footnote-ref-18)
18. Antti Kauppinen (2021) argues that if a conventional car is heading towards an innocent bystander, a neutral third party who can influence the car’s direction is required to redirect it, even at the cost of the driver’s life (639). Kauppinen holds that this is because the driver is morally responsible for creating the risky situation (639). He argues that the users of driverless cars are similarly responsible for creating risky situations and thus driverless cars should be programmed to shift risk from innocent outsiders to users (641-642). Although I agree with Kauppinen that responsibility for risks matters, I think things may be more complex than he suggests. First, other distinctions may be relevant to liability: in this case, it seems to me that it matters that the driver would count as killing the bystander if the car hit them. Second, responsibility for risk of harm may be relevant indirectly ways: it may be partly because the user is seen as responsible for the risk of harm that we count the user as doing harm if the driverless car does harm. [↑](#footnote-ref-19)
19. REFERENCEs to literature. [↑](#footnote-ref-20)
20. I thank the referee who pressed me on this. [↑](#footnote-ref-21)
21. Peterson and Spahn (2011, 417, 420) can be understood as arguing that various ways of understanding technology as morally responsible/ possessing intentionality stretch our categories too far. [↑](#footnote-ref-22)
22. There is significant body of work about how, if at all, responsibility for AI to be attributed to human agents. For a helpful summary see Tiggard (2021c) p. 595-596. Such work can be adapted to address the question of how the behaviour of driverless cars might be attributed to human agents for the purposes of applying deontological distinctions. [↑](#footnote-ref-23)
23. Reference to Nyholm? Or whichever book it was. [↑](#footnote-ref-24)
24. See also Elena Popa (2021)’s argument that agency must be understood in relation to goals and values (1733). Because autonomous machines rely on human goals and values, humans are implicated in the agency of autonomous machines in a way that allows us to avoid the responsibility gap (1743). [↑](#footnote-ref-25)
25. For discussion on the need for humans to retain ‘meaningful control’ over autonomous machines and the conditions that the autonomous machines must meet for this to happen, see Santoni de Sio and van den Hoven 2018; Mecacci and Santoni de Sio (2020). [↑](#footnote-ref-26)
26. REF to Kamm’s ambulance case. [↑](#footnote-ref-27)
27. REF to KAmm [↑](#footnote-ref-28)
28. 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 would be killed in the process (Thomson 1976). We might imagine 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-29)