Utilitas (2023), **35**, 70–87
doi:10.1017/S0953820822000395

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Of trolleys and self-driving cars: What machine ethicists can and cannot learn from trolleyology

Peter Königs

Universiteit Utrecht, Utrecht, the Netherlands
Email: p.j.konigs@uu.nl

Abstract

Crashes involving self-driving cars at least superficially resemble trolley dilemmas. This article discusses what lessons machine ethicists working on the ethics of self-driving cars can learn from trolleyology. The article proceeds by providing an account of the trolley problem as a paradox and by distinguishing two types of solutions to the trolley problem. According to an optimistic solution, our case intuitions about trolley dilemmas are responding to morally relevant differences. The pessimistic solution denies that this is the case. An optimistic solution would yield first-order moral insights for the ethics of self-driving cars, but such a solution is difficult to come by. More plausible is the pessimistic solution, and it teaches us a methodological lesson. The lesson is that machine ethicists should discount case intuitions and instead rely on intuitions and judgments at a higher level of generality.

1. Introduction

An old joke about fusion energy is that it is some thirty years away – and always will be. A similar joke could be made about self-driving cars. Self-driving cars have been 'just around the corner' for quite some time now, with their large-scale introduction constantly being postponed. At this point in time, it is unclear when, or indeed whether, there will be autonomous cars, with manufacturers, investors, and AI experts making wildly different predictions.[1] This uncertainty reflects a more general uncertainty about the future of AI. Some predict that the AI bubble is going to burst, citing insurmountable technical obstacles or pointing to past AI booms that failed to deliver on their promises. Others are imagining how highly sophisticated AI might soon radically transform how we live and work, or even destroy human civilization.[2] Time will tell

[1]For instance, in early 2022, Tesla's Elon Musk said he 'would be shocked if we do not achieve full self-driving safer than human this year' (Jin and Balu 2022), having made similar predictions in previous years. More pessimistically, the CEO of the Toyota Research Institute said in 2017 that 'we are not even close' to achieving fully autonomous driving (Muoio 2017), and the CEO of Volkswagen Autonomy assesses that 'maybe it will never happen' (Henry 2020).

[2]For one sceptical take on AI, see Larson 2021. See, for example, Danaher 2019 and Susskind 2020 for explorations of how AI could radically transform our lives, and Bostrom 2014 for how it might destroy human civilization.

who is right. One thing that is safe to say, given undeniable advances in AI and considerable corporate efforts to develop self-driving cars, is that there exists a *non-negligible chance* of self-driving cars roaming our streets sometime in the near future. This scenario raises a host of ethical issues and it is sufficiently likely for these issues to merit our attention.

One of these issues is that self-driving cars are bound to crash, giving rise to dilemmatic situations in which the death of an individual is inevitable. For instance, a self-driving car may have to decide between either staying on course and crashing into another vehicle, killing the passengers in this car, or swerving to the side, killing a pedestrian on the sidewalk. One intensely debated question in machine ethics is therefore how algorithms must be designed to ensure that self-driving cars encountering such dilemmas crash ethically.[3]

This question has led to renewed interest in the trolley problem, with many magazines and news outlets running stories on how dilemmas involving self-driving cars ('SDC dilemmas' hereafter) resemble trolley dilemmas. The academic literature on the topic has been somewhat more cautious.[4] The dominant view in that literature seems to be that the appearance of similarity between trolley dilemmas and SDC dilemmas is deceptive. There are significant differences between trolley dilemmas and SDC dilemmas, and many have taken this to mean that there is not too much that machine ethicists can learn from trolleyologists. In this article, I seek to offer clarification on the significance of the trolley problem for the ethics of self-driving cars. In so doing, I also hope to remedy the oft-lamented dearth of literature on the trolley problem.

To the question whether the trolley problem is relevant for the ethics of self-driving cars, this article offers a resounding 'No, but yes'. Disagreeing with those who have defended its relevance, I deny that it yields any *first-order* moral insights as to how self-driving cars ought to be programmed. Reflection on the trolley problem does not offer any substantive moral guidance as to how we should deal with seemingly similar dilemmas involving self-driving cars. But there is nonetheless a valuable *methodological* lesson to be gained from studying the trolley problem. What we can learn from the trolley problem is that we should distrust our case intuitions about SDC dilemmas and instead rely on intuitions and judgments at a higher level of generality in our theorizing about the ethics of self-driving cars. Although this is not the sort of philosophical gain that some have hoped for, this finding would be significant. If sound, this finding should play a critical role in the way we theorize about the ethics of self-driving cars and, indirectly, in their legal regulation. It would also mean that the much-maligned trolleyological research programme has some philosophical value after all.[5]

[3]By assuming that autonomous driving gives rise to such dilemmatic situations, I am echoing the dominant view in the debate. For one dissenting opinion, see Himmelreich (2018: 673–74). Note that we do not necessarily need to reach level-5 autonomy (full automation) for the question to become relevant how we should deal with such dilemmatic situations. What we need is AI that is capable of making sophisticated 'crash decisions' for the driver. This may already happen at level 3 (conditional automation) or 4 (high automation).

[4]Discussions of the relevance of the trolley problem for the ethics of self-driving cars and other autonomous vehicles include Brändle and Schmidt 2021; Davnall 2020; Etzioni and Etzioni 2017; Gogoll and Müller 2017; Goodall 2016; Himmelreich 2018; Hübner and White 2018; Kamm 2020; Keeling 2020; Leben 2017; Lin 2016; Lundgren 2021; Nyholm and Smids 2016; Paulo ms.; Savulescu et al. 2021; Schmidt 2022; Wallach and Allen 2009: 13–16; Wolkenstein 2018. For helpful reviews of the literature on the ethics of self-driving cars, see Hansson et al. 2021; Nyholm 2018a, 2018b.

[5]For criticisms of trolleyology, see Fried 2012; Wood 2011.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

To anticipate, the reason why the trolley problem has methodological rather than first-order moral implications for the ethics of self-driving cars has to do with what is the most plausible solution to the trolley problem. I will distinguish between an optimistic solution to the trolley problem and the pessimistic solution. An optimistic solution to the trolley problem would yield first-order moral insights, but no such solution seems in prospect. The most plausible solution seems to be the pessimistic one, and it teaches us the methodological lesson that we should distrust our case intuitions.

I begin with a brief characterization of the trolley problem, exploring especially how the trolley problem is a *problem*. This will be helpful for understanding how there can be different types of solutions to the trolley problem (section 2). I then consider how an optimistic solution would yield first-order moral recommendations for the ethics of self-driving cars. I concur with those who have defended this idea against the claim that SDC dilemmas and trolley dilemmas are too dissimilar for the trolley problem to be relevant. However, as no optimistic solution seems available, this way of extracting insights from trolleyology remains counterfactual (section 3). Next, I discuss the pessimistic solution to the trolley problem and how its plausibility has the methodological implication that we should distrust case intuitions about SDC dilemmas. I also consider two objections to this attempt at drawing a methodological lesson from trolleyology for the ethics of self-driving cars (section 4).

2. The trolley problem as a paradox

We are all familiar with the trolley problem, so I will keep my summary of the basics brief. The trolley problem is a thought experiment featuring a runaway trolley that is headed for five workers on the track who can only be saved from being run over by the trolley if an innocent bystander is sacrificed. Some versions of the dilemma tend to elicit the intuition that it is permissible to sacrifice one person to save five. In SWITCH, for instance, the trolley can be redirected to another track, killing only one person on the side-track instead of the five on the main track. Most people intuitively judge this permissible. Other versions trigger the opposite response. In FOOTBRIDGE one can prevent the trolley from running over the five people on the track by shoving a heavy person off a bridge onto the tracks below to block the trolley, thus killing the heavy person but saving the five. Most people's intuitive response is that this is morally impermissible. SWITCH and FOOTBRIDGE are only two of the many variants of the dilemma, which vary in all sorts of ways, such as how many people can be saved, who these people are, and how exactly they can be saved.[6]

While the basics of trolleyology are all too well known, what is less obvious is what the 'problem' part of the trolley problem is. How is the fact that people have the intuitions they have about trolley dilemmas a philosophical problem? Usually, our having certain moral intuitions about ethically interesting cases is simply treated as useful information, not as a philosophical problem. To be sure, it is obvious that there is a *descriptive* problem, namely that of identifying the principle that underlies our responses to trolley dilemmas. Philosophers and psychologists have tried hard to solve this problem empirically (by actually conducting empirical studies) and pseudo-empirically (by collecting, reporting, and systematizing intuitions from the armchair). But the descriptive problem, though relevant for the philosophical problem, is

[6]Some of the dilemmas discussed in this context do not involve a trolley. The founding papers of trolleyology are Foot 1967; Thomson 1976; 1985.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

not identical with the philosophical problem.[7] The reason why philosophers are interested in the solution to the descriptive trolley problem is not that they are interested in people's moral psychology for its own sake. They are interested in it because they take it to be relevant to solving the philosophical problem. But what is the philosophical problem? Having a good grasp of what the problem is is useful for understanding the different types of solutions to the problem and how they have different implications for the ethics of self-driving cars.

I believe that the trolley problem is best conceived of as a paradox. Following Michael Huemer, we can understand a paradox 'as a piece of reasoning, or a situation about which such reasoning is constructed, that has widespread and robust appeal, but that leads to a contradictory or absurd conclusion'. Huemer goes on to explain that '[p]aradoxes trade on intellectual illusions: cases in which something that is actually wrong seems correct when considered intellectually' (2018: 5).

The trolley problem fits this account of what a paradox is. We have different intuitions about dilemmas that, at least on the face of it, do not appear to differ from a moral point of view.[8] In particular, trolley dilemmas tempt us to assent to the following six propositions, which generate a contradiction:

(P1) The one-for-many trade-off is permissible in *D*, a set of dilemmas containing, for example, SWITCH. [based on our case intuitions about *D*]
(P2) The one-for-many trade-off is impermissible in *D**, a set of dilemmas containing, for example, FOOTBRIDGE. [based on our case intuitions about *D**]
(P3) The permissibility of the one-for-many trade-off is not the same for dilemmas in *D* as for dilemmas in *D**. [from P1 and P2]
(P4) The dilemmas in *D* and the dilemmas in *D** do not differ in any morally relevant respect. [based on the seeming that they do not thus differ]
(P5) If the dilemmas in *D* and the dilemmas in *D** do not differ in any morally relevant respect, the permissibility of the one-for-many trade-off is the same for dilemmas in *D* as for dilemmas in *D**. [very plausible assumption]
(P6) The permissibility of the one-for-many trade-off is the same for dilemmas in *D* as for dilemmas in *D**. [from P4 and P5]

P3 and P6 contradict each other. The contradiction arises when we take our intuitions about different cases at face value (P1 and P2) as well as the impression that these cases do not differ in any morally relevant respects (P4). The reason, then, why our responses to trolley dilemmas are not just a useful piece of information but a real philosophical problem is that they generate a paradox.

Solving the trolley problem requires showing how we can plausibly refuse to accept at least some of the propositions that jointly generate the paradox. Identifying the principle that underlies our responses – solving the *descriptive* trolley problem – may help us solve the paradox, but, as will become clear later, it is not necessary for solving it.

[7]I am here adapting a distinction drawn by Greene 2016.

[8]Using the generic 'we', I am here assuming that people's intuitions mostly converge. This is a bit of an idealization. Despite considerable convergence, there is also some variation in people's intuitions. I leave this problem to one side, assuming that there is sufficient convergence to permit talk of 'our' intuitions. Also, the problem of intuitional disagreement is a general problem in philosophy and not specific to the topic of this article.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

Sometimes the trolley problem is characterized along the above lines.[9] Usually, however, it is characterized differently. The standard conception of the trolley problem rests explicitly or implicitly on the assumption that our intuitions about trolley dilemmas can be trusted to be responding to morally relevant factors. Consider for instance how Bruers and Braeckmann characterize the trolley problem in their helpful review of the debate: 'The basic question is the following: What is the *morally relevant* difference between Dilemma A and Dilemma B, such that it is morally allowable to act in A, but not to act in B?' (2014: 252, my emphasis).[10] This framing of the problem rests on the assumption – highlighted by my italicization – that our responses to the dilemmas are responding to a morally relevant difference. It dismisses the impression, captured by P4, that the dilemmas do not seem to be relevantly different. On this view, the problem part of the trolley problem is not that it generates a paradox. Rather, the problem is that there is an important moral principle, detected by our case intuitions, that keeps eluding us. Solving the trolley problem is all about identifying this principle.

One problem with this standard account is that it rules out solutions to the trolley problem that do not simply identify the principle that accounts for our case intuitions. Joshua Greene (2008), for instance, suggests that we dismiss our case intuitions about the impermissibility of the one-for-many trade-off in FOOTBRIDGE and similar cases. Greene's and other solutions that dismiss some of our case intuitions should surely qualify as attempts at solving the trolley problem. The standard account would also rule out the pessimistic solution, which will be discussed in section 4. Another problem with the standard account is that even those who have a lot of confidence in our case intuitions have acknowledged that it is a possibility that these intuitions are not in fact responding to morally relevant factors. For instance, Frances Kamm, champion of the 'method of cases' that privileges case intuitions, notes that once a principle is identified we must 'consider the principle on its own, to see if it expresses some plausible value or conception of the person or relations between persons. This is necessary to justify it as a *correct* principle, one that has normative weight, not merely one that makes all of the case judgments cohere' (Kamm 2007: 5; see also 1992: 7–8; 2016: 60). Bruers and Braeckmann, too, acknowledge that some of our intuitions might be 'moral illusions' (2014: passim). I will discuss this possibility in more detail shortly. At this point, it matters because it means that our account of the trolley problem should not presuppose that our intuitions are responding to morally relevant factors.

Construing the trolley problem as a paradox, which can be solved other than by identifying a presumed morally relevant difference, is thus more inclusive than the standard conception. With this account of the trolley problem in hand, we can now turn to exploring how different solutions to the trolley problem yield different lessons for the ethics of self-driving cars.

[9]See especially Baumann 2022: 1738. Consider also how Kamm introduces the problem in one of her papers: 'Some moral philosophers with nonconsequentialist leanings have puzzled over how to reconcile their apparently conflicting intuitions that it is morally permissible to redirect fatal threats away from a greater number of people, but that it is not morally permissible to kill one person in order to transplant his organs into a greater number of people merely because this alone will save their lives' (Kamm 1989: 227). The intuitions appear to conflict because there does not appear to be a morally relevant difference between these two cases. The 'puzzle' is this apparent conflict. Similarly, Baltzly (2021) characterizes the trolley problem as a puzzle generated by a (seeming) clash of intuitions.

[10]Similar characterizations have been offered, for example by Kamm 2016: 47; Otsuka 2008: 93; Paulo ms.; Thomson 1976: 206; 1985: 1401; 2008: 362–63; 2016: 115–16.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

3. An optimistic solution would yield first-order moral insights

The traditional approach to solving the trolley problem is by showing that there exists a plausible moral principle that accounts for our puzzling intuitions about trolley dilemmas. Those pursuing this approach seek to solve the paradox by rejecting P4. Once the principle underlying our case intuitions is identified, it will become apparent that there is a morally relevant difference between dilemmas in *D* and *D** after all. Trolley dilemmas may *appear* morally indistinguishable, but the discovery of the principle that underlies our intuitions will make us see that, contrary to appearance, they really do differ in morally relevant respects. I call such a solution to the trolley problem an 'optimistic' solution.

Many who have worked on the trolley problem have sought to offer an optimistic solution. Those seeking to provide an optimistic solution must achieve two things. First, they must identify the principle that underlies our intuitions about trolley dilemmas. In other words, they must solve the descriptive trolley problem. This explains why a large proportion of trolleyological efforts have been expended on identifying this principle. In a second step, they must show that this principle is not just some morally arbitrary psychological principle but one that possesses at least some independent normative plausibility.

This second step requires some elaboration. Understanding it will also become relevant for appreciating why the pessimistic solution to the trolley problem, to be discussed shortly, is in fact the more plausible one. To understand why the second step is necessary, it is important to consider that intuitions and judgments at different levels of generality play a role in the trolley problem. First, there are our case intuitions, for example, that the one-for-many trade-off is permissible in SWITCH but impermissible in FOOTBRIDGE. P1 and P2 are judgments based on these case intuitions. But we also form moral judgments at a higher level of generality. For instance, everybody would agree that the hair colours of the persons in the dilemmas are morally irrelevant, whereas whether harm is intended or merely foreseen might well be morally relevant. These are not judgments about individual cases. They are judgments at a higher level of generality, which are often themselves based on more general intuitions. P4 expresses such more general judgments. The trolley problem is thus about an apparent tension between case intuitions and judgments at a higher level of generality.[11]

This tension requires us to say something about the relationship between case intuitions on the one hand and intuitions and judgments at a higher level of generality. It is a plausible assumption that our case intuitions have evidential force. In fact, if we did not take case intuitions to be evidentially relevant in moral theory, philosophers would have no reason to be interested in the trolley problem to begin with. We are interested in our responses to trolley dilemmas because we take these responses to have at least some evidential force. For instance, if our responses turned out to be sensitive to whether harm is intended or merely foreseen, this would be a good reason to conclude that this factor *is* morally relevant. If we were previously unsure about the moral significance of this distinction, the trolleyological findings might tip the balance in favour of accepting its significance.[12]

[11]I take intuitions to be 'seemings' or 'appearances' (see Huemer 2005: 102). Judgments can be based directly on intuitions that are taken at face value. For instance, judgments at a higher level of generality can be directly based on intuitions at a higher level of generality. But they can also be the result of more complex inferential processes.

[12]As a side note, this is something that Kagan (2016), whose take on the trolley problem I am mostly endorsing in this article, in my view does not sufficiently appreciate.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

At the same time, case intuitions have only so much evidential force. When case intuitions clash with a judgment at a higher level of generality, this might prompt us to dismiss the judgment at a higher level of generality. But it might as well indicate that there is something wrong with our case intuitions. For instance, if our case intuitions about trolley dilemmas turned out to be sensitive to people's hair colours, it is arguable that this should not prompt us to revise our prior belief that hair colour is morally irrelevant. Rather, it should make us conclude that there is something wrong with our case intuitions. It would defeat their evidential force.

This is why the second step in the traditional approach to solving the trolley problem is necessary. Once the factor that our intuitions are responding to is identified, it must be shown to possess sufficient independent plausibility. This requirement is an acknowledgment of the fact that our case intuitions possess some but not unlimited evidential force. As noted above, even those who assign a privileged role to case intuitions, such as Kamm, agree that this second step is indispensable.[13]

This being said, let us assume that an optimistic solution to the trolley problem can be offered. The descriptive trolley problem is solved (step 1), and the identified principle turns out to possess sufficient independent normative plausibility (step 2). P4 can plausibly be rejected, and the paradox is solved. Would such an optimistic solution to the trolley problem help us in dealing with SDC dilemmas? The principal objection to attempts at bringing trolleyology to bear on the ethics of self-driving cars has been that SDC dilemmas are too dissimilar from trolley dilemmas. Differences between SDC dilemmas and trolley dilemmas highlighted in the literature include the following:

1. In trolley dilemmas, the decision has to be made in a split second, whereas those deciding how self-driving cars ought to behave make the decision long in advance.
2. In trolley dilemmas, the decision is made by one individual, whereas the decision how self-driving cars ought to behave is made by multiple stakeholders (ethicists, politicians, lawyers, etc.).
3. In trolley dilemmas, the set of potentially relevant considerations is artificially restricted, whereas real-life SDC dilemmas require taking into account a much larger number of considerations, including all sorts of contextual and situational factors.
4. Trolley dilemmas bracket questions of moral and legal responsibility, which are crucial to the ethics of SDC dilemmas.
5. Unlike trolley dilemmas, SDC dilemmas involve risk or uncertainty.
6. Unlike trolley dilemmas, SDC dilemmas involve strategic interaction in that the outcome depends in part on other agents' behaviour.
7. While trolley dilemmas are isolated one-shot games, SDC dilemmas must be construed as iterated decision problems.
8. Trolley dilemmas are about how a bystander, who is not herself at risk, should act. SDC dilemmas are about how cars should behave that can both cause harm and be at the receiving end of crashes.[14]

These significant differences between trolley dilemmas and SDC dilemmas seem to make it impossible to derive first-order moral insights from trolleyology for the ethics

[13]This point is discussed by Hurka 2016: 142; Kagan 2016. For a related discussion, see Königs 2020.

[14]I am here summarizing differences pointed out in Gogoll and Müller 2017; Goodall 2016; Himmelreich 2018; Nyholm and Smids 2016. Further potential differences are discussed in Kamm 2020.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

of self-driving cars. Siding, however, with those who have defended the relevance of the trolley problem, I believe that *if* an optimistic solution were found, it would yield first-order moral insights for the ethics of self-driving cars. I do not think that the antecedent of the conditional will be satisfied, as I will shortly explain. But it is worth exploring how such an optimistic solution *would* prove useful for machine ethicists, contrary to what some have suggested in light of the differences between SDC dilemmas and trolley dilemmas (see again Gogoll and Müller 2017; Goodall 2016; Himmelreich 2018; Nyholm and Smids 2016).

These differences do rule out one way of extracting insights from the trolley problem for the ethics of self-driving cars, but they leave open two other ways.[15] What I think the differences do rule out is case-based analogies. An optimistic solution to the trolley problem would confirm that our case intuitions are picking up on an important moral difference. It would vindicate our case intuitions about trolley dilemmas. One might be led to think that this would allow relying on judgments about particular trolley dilemmas as starting points for deciding difficult SDC dilemmas. In a nutshell, an argument along these lines would look like this: 'The one-for-many trade-off is impermissible in trolley dilemma *T*. There is no morally relevant difference between *T* and the target SDC dilemma *S*. Therefore, a one-for-many trade-off is also impermissible in *S*.' It is such case-based analogies that the many differences between trolley dilemmas and SDC cases render virtually impossible. For many of these differences cannot be assumed to be irrelevant from a moral point of view. It matters morally whether a dilemma involves risk or uncertainty, whether it involves strategic interaction, and so forth. What these differences therefore imply is that, apart maybe from very few exceptions, SDC dilemmas will not resemble trolley dilemmas in all morally relevant respects. SDC cases almost always involve additional moral complexity that prevents us from drawing such analogies.[16]

While this approach is almost always blocked, there are two other ways in which an optimistic solution to the trolley problem would yield valuable insights for the ethics of self-driving cars. First, an optimistic solution to the trolley problem would vindicate the status of our case intuitions about trolley dilemmas as important moral data points that ethical theorizing should strive to accommodate. These intuitions could thus serve as touchstones for moral theorizing in general and for theorizing about SDC dilemmas in particular. If the ethical principles we intend to use to regulate self-driving cars cannot account for our intuitions about trolley dilemmas, this is a good, if defeasible, reason to reject them.[17]

Second, as also pointed out by Geoff Keeling (2020: 297–300), an optimistic solution would involve the identification of some valid moral principle that is likely also to have validity in SDC dilemmas. If our case intuitions about trolley dilemmas vindicate, say,

[15]My understanding of the first two approaches has benefited from discussions with Norbert Paulo and is informed by Paulo ms.

[16]I am here disagreeing with Paulo (ms.), who has explored and defended this approach in much more detail. Our disagreement seems to be mostly a matter of degree, though, as he agrees that drawing such analogies works only in rare situations.

[17]In this I am now following Paulo, who has, again, explored and defended this approach in much more detail (Paulo ms.; see also Goodall 2016: 811–12; Keeling 2020: 297). A disagreement between Paulo and me is that I believe that this approach presupposes an optimistic solution to the trolley problem. Such a solution would provide the necessary vindication of our case intuitions about trolley dilemmas, thereby establishing them as suitable touchstones for moral theorizing. Paulo, as I read him, thinks that our case intuitions can serve as such touchstones even in the absence of a 'clear solution to the trolley problem' (Paulo ms.: 31).

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

the moral difference between intending and foreseeing harm, it is plausible that this difference matters in SDC dilemmas too. To be sure, SDC dilemmas differ from trolley dilemmas along a range of dimensions. But this should not lead us to conclude that the intending/foreseeing distinction *loses* its relevance in SDC dilemmas. While this is a theoretical possibility, it is not clear why we should assume this to happen. What we should conclude is that there are various morally relevant factors *in addition to* the intending/foreseeing distinction that need to be accounted for in SDC dilemmas. The differences between trolley dilemmas and SDC dilemmas thus do matter, but they do not render the solution to the trolley problem irrelevant.

The first and the second approach are two sides of the same coin. The moral principle that features in the second approach is gained from looking at the case intuitions that feature in the first approach. In a way, the first approach focuses on the input (the case intuitions), whereas the second approach focuses on the output (the moral principle that our case intuitions reveal to us). Still, it is useful to distinguish the two approaches as they are effectively two different ways of connecting the trolley problem with the ethics of self-driving cars. The two approaches are not mutually exclusive and can be used in combination.

All this is not to say that the differences between SDC dilemmas and trolley dilemmas do not matter. They do, and those who have highlighted these differences deserve credit for doing so. But I agree with those who have defended the significance of the trolley problem that these differences do not mean that an optimistic solution to the trolley problem would not yield first-order moral insights for the ethics of self-driving cars. *If* such a solution were found, it would yield such insights. My disagreement with them concerns the antecedent of the above conditional. While an optimistic solution to the trolley problem *would* be helpful, trolleyologists have struggled to produce such a solution.

Recall, producing such a solution requires, first, solving the descriptive trolley problem, and second, showing the principle that explains our intuitions to have sufficient prima facie plausibility. The first step already has proved notoriously difficult. As Guy Kahane has noted, trolleyologists have 'failed to identify [the] principle after over 40 years of reflection on the trolley problem!' (2013: 434). Similarly, Peter Baumann observed as recently as 2022 that '[n]o single proposed solution has convinced a significant majority of authors' (2022: 1738). One could say that the descriptive trolley problem is the Gettier problem of moral philosophy. Whenever some solution to the descriptive trolley problem is put forth, it does not take long for some clever trolleyologist to come up with counterexamples. The descriptive trolley problem is famous for its intractability.

What is more, it is becoming increasingly difficult to see how the second criterion might be satisfied. Many independently plausible distinctions – positive versus negative duties, doing versus allowing, intending versus foreseeing (the doctrine of double effect), etc. – have been tried as solutions to the descriptive trolley problem, but such attempts have tended to fail. To see the difficulty of meeting the second criterion more clearly, consider that some individual response combinations are particularly odd from a moral standpoint. Joshua Greene has found that while the one-for-many trade-off in Footbridge is typically judged impermissible, a majority of people think the trade-off is permissible when the heavy person is not pushed off the bridge but dropped onto the tracks through a switch-operated trapdoor (Trapdoor) (Greene 2014: 709; see also Greene et al. 2009).[18] Not only is this a data point that existing

[18]Greene tested two different versions of Trapdoor, which were judged roughly equally permissible. Greene's own theory why Footbridge and Trapdoor elicit different responses is criticized in Berker 2009: 323 n. 73; Railton 2014: 854–55.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

proposed solutions to the trolley problem, including more recent ones, fail to accommodate, which renders these solutions inadequate at the descriptive level.[19] It is also extremely difficult to conceive how there could *possibly* exist a morally relevant difference between FOOTBRIDGE and TRAPDOOR. Although we currently do not know which factor explains this difference in intuitions, we seem to be in a position to predict with some confidence that this factor would not strike us as morally relevant upon discovery.

In the light of such findings, and given how intractable the problem has proven, it seems reasonable to be sceptical that the factors our case intuitions are responding to will turn out morally relevant. Rather, we should assume that they are erratically responding to factors that, upon discovery, prove morally arbitrary. This is the pessimistic solution. In defending the pessimistic solution, I am concurring with Shelley Kagan, who 'suspect[s] that any principle actually capable of matching our intuitions across the entire range of trolley problems will be a messy, complicated affair, full of distinctions that strike us as morally irrelevant'. This means that 'we cannot fit [our case intuitions] into a larger moral discourse that we should be prepared to embrace' (Kagan 2016: 154 and 164, respectively; similarly, Unger 1997).

The apparent failure of the traditional approach to solving the trolley problem thwarts both of the above outlined attempts at extracting insights from the trolley problem. For one thing, as we do not know as of yet what principle is governing our responses to trolley dilemmas, this principle *cannot* possibly inform our theorizing about SDC dilemmas. This already rules out the second approach. For another thing, it is increasingly likely that this principle, whatever it is, will turn out arbitrary from a moral point of view. This means that even if we were to identify the principle, this principle *should not* inform our theorizing about SDC dilemmas. This is another reason why the second approach fails. It also means that our case intuitions, which seem to be erratically responding to morally irrelevant factors, should not serve as touchstones for our theorizing about SDC dilemmas. This rules out the first approach.

It is true, then, that an optimistic solution to the trolley problem *would* yield first-order moral insights for the ethics of self-driving cars. The problem is that such a solution is extremely difficult to come by, as our case intuitions seem to be misleading us. It is therefore doubtful that such first-order moral insights might realistically be gained from the trolley problem.

All this does not mean that the trolley problem does not hold useful lessons for the ethics of self-driving cars. Nor does it mean that the trolley problem cannot be solved. On the contrary, the above discussion suggests that there is an alternative, pessimistic solution to the trolley problem, and it holds a valuable methodological lesson for those interested in the ethics of self-driving cars. It is to the discussion of this pessimistic solution and the methodological lesson it holds that I now turn.

4. The pessimistic solution teaches us a methodological lesson

Solving the trolley problem requires resolving the tension between our case intuitions, which suggest that the various trolley dilemmas differ in morally relevant respects, and the initial impression that they are morally indistinguishable. The optimistic solution is

[19]Recent approaches that seem to fall victim to this finding include Baumann 2022; Graham 2017; Haslett 2011; Kamm 2016; Kleingeld 2020. Further convincing critiques of some of these accounts have been put forth by Hurka 2016; Kagan 2016 (both criticizing Kamm); Schmidt 2022: 205–6 (criticizing Kleingeld); Graham 2017: 182 n. 20 (criticizing Haslett). See also Königs 2020.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

to show that this initial impression is mistaken and that there really exist morally relevant differences, revealed to us by our case intuitions. The above discussion, which has cast doubt on the prospects of this traditional approach, suggests that an alternative solution is preferable. The most plausible solution, it seems, is to resolve the tension by dismissing our case intuitions. It is becoming increasingly inconceivable that our case intuitions are responding to morally relevant factors. Whatever obscure principle our case intuitions are governed by, it does not seem to be a principle that possesses sufficient independent moral plausibility. If we dismiss our case intuitions on this ground, we no longer need to accept P3, which is derived from these case intuitions. The tension that constitutes the trolley paradox dissolves.

Note that this solution to the trolley problem does not require solving the descriptive trolley problem. Although we do not know which factors our intuitions are responding to, we seem to know enough to be fairly confident that they are not responding to factors that are morally relevant. To be sure, the possibility that our case intuitions are responding to morally relevant factors cannot be ruled out with complete certainty. Theoretically, it is still possible that our responses turn out to be governed by some – perhaps extremely messy and non-obvious – principle that we have not been able to think of and that, upon discovery, strikes us as possessing considerable independent normative plausibility. But at this point in the history of trolleyology, we should be very reluctant to put stock in our case intuitions.[20]

Here is how this pessimistic take on the trolley problem holds an important methodological lesson for the ethics of self-driving cars. The pessimistic solution suggests that our case intuitions about trolley dilemmas are not to be trusted. Given that our case intuitions about trolley dilemmas are not to be trusted, it is plausible to conclude that our case intuitions about SDC dilemmas should not be trusted either. The reasoning is simple: if our case intuitions about sacrificial dilemmas of one kind (trolley dilemmas) are leading us astray, so presumably are our case intuitions about sacrificial dilemmas of a very similar kind (SDC dilemmas). To be sure, the fact that our case intuitions are unreliable in *one* moral subdomain – trolley dilemmas – should not lead us to conclude that they are unreliable in *all* moral domains. No such radical conclusion can be drawn from an assessment of a limited number of case intuitions about a set of very specific scenarios. If the domain we are considering is sufficiently remote from the trolley problem, it is arguable that our case intuitions should still be given a lot of credit. However, the problem of self-driving cars does not seem all that remote from trolley dilemmas. Despite some undeniable differences, SDC dilemmas are still *a lot* like trolley dilemmas. It is no coincidence that so many have likened SDC dilemmas to trolley dilemmas. In fact, one could say that real-life SDC dilemmas are just more complex variations of the sort of dilemmas that feature in standard trolley thought experiments. Given then that our case intuitions about standard trolley dilemmas seem to be unreliable, we should not rely on such intuitions when we consider SDC

[20]Two side notes on the significance of the pessimistic solution: (1) Although the pessimistic solution should qualify as a solution, it does not answer *all* our questions. We still do not know when it is permissible to sacrifice one person to save others. But it transforms the trolley problem from a genuine puzzle or paradox into an ordinary tractable philosophical question like any other. There is no real trolley problem anymore. (2) By casting doubt on our case intuitions, the pessimistic solution supports P4, the impression that there is no moral difference between the various trolley dilemmas. It does not, however, completely rule out that there are such differences. For there may be reasons other than our case intuitions to assume that such differences exist.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

dilemmas. Our theorizing about self-driving cars should instead be based on intuitions and judgments at a higher level of generality. Case intuitions about how self-driving cars ought to behave are to be distrusted and should not be allowed to play a prominent role in our theorizing about this issue. This is the methodological lesson that I believe trolleyology holds for machine ethicists.[21]

Some such judgments at a higher level of generality include that it is more permissible to kill a small number of people than a large number of people, that human life has greater value than the life of a non-human animal, or that killing a person is worse than injuring a person. Less trivially, it seems plausible that it is more permissible to kill someone who has violated traffic regulations (e.g. by crossing a red light) than someone who has not, and that those travelling by car must not externalize risks onto cyclists and pedestrians, who have chosen a means of transportation that is less dangerous to other traffic participants. As Hübner and White have plausibly suggested (2018: 694–95), there is also an important moral difference between whether a person is 'involved' or 'uninvolved' in a traffic situation. It seems more acceptable to kill someone who is voluntarily participating in traffic (e.g. another passenger in a car) than someone who is not even a traffic participant (e.g. someone sitting in a sidewalk café).

This brief review of some principles that might be relevant is of course woefully incomplete. But it conveys a rough idea of how ethical theorizing about self-driving cars should proceed. It is not the goal of this article to put forth anything resembling a fully worked-out ethical framework for dealing with self-driving cars. Nor do I mean to suggest that developing such a framework will become an easy task once we have internalized the methodological lesson that I think trolleyology teaches us. Like so many real-life ethical problems, the ethical problems that self-driving cars give rise to are challenging and complex. Things are further complicated by the fact that many principles are subject to disagreement among ethicists. Ethicists disagree about the relevance of the doing versus allowing distinction, about the validity of the doctrine of double effect, about whether the age of the potential dilemma victim is a morally relevant factor, and so forth. Also, with such principles, as so often, the devil is in the details. For instance, while 'involvement' in a traffic situation seems morally relevant, it is less clear who exactly should count as 'involved', as Hübner and White readily acknowledge. Finally, even when the applicable principles have been identified and understood, their *pro tanto* and *ceteris paribus* nature still requires a good deal of skill and judgment in application.

I do not therefore pretend that the methodological suggestion that we avoid relying on case intuitions about SDC cases yields straightforward answers. But it is still an important lesson that we need to take seriously if we are to come to grips with the problem of self-driving cars. It means that machine ethicists should deviate from what is arguably the standard method in ethical theory. The standard method is to consider intuitions and judgments at *all* levels of generality, including case intuitions, and try to make them fit into a coherent whole. This is for instance how advocates of the method of reflective equilibrium suggest we proceed. Here is how John Rawls describes this method:

> People have considered judgments at all levels of generality, from those about particular situations and institutions up through broad standards and first principles to formal and abstract conditions on moral conceptions. One tries to see how people

[21]A similar methodological approach has been championed by Unger 1997.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

> would fit their various convictions into one coherent scheme, each considered conviction whatever its level having a certain initial credibility. (Rawls 1975: 8)

If the above analysis is correct, those working on the ethics of self-driving cars should reject this model by altogether ignoring or at least significantly discounting case intuitions.[22] A fortiori, they should reject methods that assign a privileged status to case intuitions, such as Kamm's method of cases. One more concrete implication of this concerns the normative significance of the much-discussed Moral Machine Experiment (Awad et al. 2018). Its authors collected and analysed responses to different SDC dilemmas from millions of people around the world, and the results are fascinating. However, by collecting people's case intuitions about SDC dilemmas, the experimenters focused on the sort of intuitions that we have reason to believe to be confused. The normative relevance of the results of this experiment is therefore doubtful.[23]

Before concluding, I wish to consider two possible objections to this way of drawing a methodological lesson from trolleyology. The first objection challenges the claim that the unreliability of our case intuitions about trolley dilemmas should undermine our trust in our case intuitions about SDC dilemmas. I suggested that we draw this conclusion because these two types of dilemmas are so similar. My underlying assumption is that the unreliability of our case intuitions about trolley dilemmas has to do with the subject matter. We should distrust case intuitions about SDC dilemmas because they are about the same subject matter as the intuitions about trolley dilemmas, which seem to be confused. They concern the same moral subdomain. Against this, the first objection holds that the unreliability of our case intuitions about trolley dilemmas has nothing to do with the subject matter itself but with how trolley dilemmas tend to be designed. It is a frequent objection to trolley dilemmas and other thought experiments that the scenarios we are invited to consider are overly simplified. They fail to elicit valid intuitions because they fail to reflect the complexity of real-life moral problems. If we make sure to consider SDC dilemmas in all their complexity when gathering case intuitions about them, these intuitions might be more trustworthy than our case intuitions about highly stylized and simplified trolley dilemmas, or so the objection goes. The additional complexity that is characteristic of SDC dilemmas might thus enhance the reliability of our case intuitions.[24]

There are three issues with this objection. To begin with, it is unclear whether it really is the simplified nature of standard trolley thought experiments, rather than the subject matter, that is the root cause of the apparent unreliability of our case intuitions about them. Until this hypothesis has been confirmed, it seems unwise to bank on it. Moreover, while it is true that *real-life* SDC dilemmas involve a whole lot of additional complexity compared to standard trolley dilemmas, it is likely, and perhaps

[22]Technically, my methodological suggestion could be compatible with sophisticated versions of the method of reflective equilibrium, as they might be able to accommodate findings regarding the reliability of certain intuitions (for related discussions, see Paulo 2020; Tersman 2008). I am therefore not specifically taking issue with Savulescu et al.'s (2021) or Brändle and Schmidt's (2021) defence of the method of reflective equilibrium in machine ethics. The essential point I wish to make is that we should discount case intuitions when considering the ethics of self-driving cars – whether this is also implied by the method of reflective equilibrium or not.

[23]I am here disagreeing with Savulescu et al. (2021), who are more positive about the significance of the results.

[24]Thanks to an anonymous reviewer and Norbert Paulo for raising this objection.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

inevitable, that when gathering case intuitions about SDC dilemmas ethicists will again employ simplified thought experiments. The Moral Machine Experiment, for instance, used highly stylized dilemmas to collect people's case intuitions about self-driving cars. Therefore, if simplification is the root cause of our case intuitions' unreliability, this problem is bound to reappear when we consider SDC dilemmas, even though *real-life* SDC dilemmas are more complex. To be sure, machine ethicists could design thought experiments that capture SDC dilemmas in their full complexity, breaking with the philosophical tradition of using highly stylized thought experiments. But this leads to the third problem: the additional complexity that real-life SDC dilemmas involve compared to simple trolley cases does not seem to be the sort of complexity that one would expect to enhance our case intuitions' reliability. In particular, it is difficult to see how adding risk or uncertainty to the mix should have this effect, given that human cognition is notoriously challenged by risk and uncertainty. By the same token, the additional element of strategic interaction clearly makes things a lot more intricate. Given how confused our case intuitions about trolley dilemmas apparently are, relying on case intuitions about the *very similar but significantly more intricate and trickier* SDC dilemmas seems like a risky move.

This is not to rule out completely the reliability of our case intuitions about (complex or simplified versions of) SDC dilemmas. It is a theoretical possibility that our intuitions about SDC dilemmas are spot-on while our case intuitions about trolley dilemmas are completely confused. But it should no longer be the default assumption that our case intuitions about SDC dilemmas are reliable. Usually, there is a presumption in favour of treating intuitions as possessing (defeasible) evidential force. I submit that this presumption is no longer plausible for our case intuitions about SDC cases. It is arguable that their reliability is now something that needs to be positively established. Given how confused our case intuitions about trolley dilemmas seem to be and how similar SDC dilemmas are, it would be methodologically reckless to presuppose their reliability about SDC cases.

Unless a positive reason to trust our case intuitions about SDC dilemmas is provided, I suggest that we rely on intuitions and judgments at a higher level of generality in our theorizing about the ethics of self-driving cars. There is of course the option of re-running the trolleyological program for SDC dilemmas. One could explore which principle (or principles) underlies our case intuitions about SDC dilemmas and try to determine whether it possesses sufficient independent normative plausibility. This could in theory lead to a vindication of our case intuitions about SDC dilemmas. But I suspect that few machine ethicists will be tempted to embark on such a project.

The second and perhaps less obvious objection to the methodological lesson I draw from trolleyology is that relying on more general intuitions does not in fact allow us to sidestep unreliable case intuitions, because our more general intuitions (and judgments based on them) are themselves derived from case intuitions. Therefore, if our case intuitions about sacrificial dilemmas are not to be trusted, our more general intuitions about sacrificial dilemmas are not to be trusted either. The idea here is that the more general principles we hold to be true are themselves derived from case intuitions. They are case intuitions in condensed form, as it were. For instance, the reason why we think that hair colour is morally irrelevant in sacrificial dilemmas might be based on a large number of case intuitions about moral scenarios featuring people with different hair colours. All these case intuitions combined have led us to accept the more general principle that hair colour is a morally irrelevant factor. If this model of the origins of more general moral intuitions is correct, the methodological suggestion to rely on general intuitions

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

and to discount case intuitions would not make much sense, or so the objection goes. For, although it might not *look* like it, we would be relying on case intuitions after all. Call this the objection from condensation, because it is based on the assumption that general intuitions are case intuitions in condensed form.

Whether this assumption is correct is an interesting question in its own right, which has methodological implications beyond the problem considered in this article. I am not aware of an explicit defence of this assumption, and the dominant view is arguably that intuitions at a higher level of generality are largely independent from case intuitions. The objection thus relies on a not widely accepted speculation. Still, I find the speculation sufficiently plausible for the objection from condensation to deserve being taken seriously.[25]

I believe the objection fails even if all or many of our more general intuitions really just are case intuitions in condensed form. Let us assume, then, for the sake of argument, that the condensation assumption is correct, and let us call the case intuitions from which the general intuitions are condensed the input intuitions. For the objection from condensation to be sound, all or the bulk of the input intuitions must have been shown to be unreliable. We would then have a 'garbage in/garbage out'-problem. Our more general intuitions would be 'garbage', as they would be condensations of input intuitions that are mostly 'garbage'. However, the above-raised concerns about the reliability of our case intuitions concerned only a relatively small set of case judgments. The conclusion that I suggested we draw is that our case intuitions *about trolley dilemmas and SDC dilemmas* are not to be trusted. I do not think that the findings from trolleyology warrant concluding that our case intuitions are *generally* unreliable. Since our case intuitions about trolley dilemmas and SDC dilemmas probably make up at most a very small fraction of the input intuitions, a lot of the input is in fact not 'garbage'. The few confused case intuitions about sacrificial dilemmas would be neutralized, as it were, by a large set of case intuitions that we have no particular reason to believe to be 'garbage'.[26] Even if the condensation assumption were correct, drawing on more general intuitions would be a way of leveraging epistemic resources that are independent from the case intuitions that have been identified as problematic.

Note also that we observed that our case intuitions about trolley dilemmas *clash* with our general intuitions about which factors are relevant in trolley dilemmas. This means that the latter cannot possibly be mere condensations of the former. For if they were, our general intuitions and our case intuitions would align. One of two things must therefore be the case. Either the condensation assumption is false, meaning that our more general intuitions about trolley dilemmas are independent from our case intuitions. Or they are derived from case intuitions, but our case intuitions about trolley dilemmas make up only a very small fraction of the input intuitions. Both possibilities would explain why our case intuitions about trolley dilemmas and our general intuitions about trolley dilemmas do not align. In either case, focusing on our more general judgments would be an effective way of avoiding or significantly reducing reliance on case intuitions that must be assumed to be confused.

[25]For a related discussion, see Daniels 1979: 259.

[26]These other case intuitions *could* be false too. But such a more general scepticism about the reliability of moral intuitions is not a dialectically relevant problem here. What matters here is that the reliability problem identified in this article concerns only a subset of intuitions.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

5. Conclusion

I conclude with a quick summary of the argument of this article. If an optimistic solution to the trolley problem could successfully be offered, this would have interesting first-order moral implications for the ethics of self-driving cars, despite differences between trolley dilemmas and SDC dilemmas. But such a solution is extremely difficult to come by. It therefore seems that no first-order moral insights can be gained. The most plausible solution to the trolley problem is the pessimistic one. The pessimistic solution holds the valuable lesson that we should distrust our case intuitions about SDC dilemmas in our theorizing about the ethics of self-driving cars. The lesson machine ethicists can learn from trolleyology is thus a negative and methodological one, but it is no less important for that.

Acknowledgements. Thanks to Norbert Paulo and two anonymous reviewers for helpful comments on earlier drafts of this article. While working on this project, the author was a member of the project group 'Regulatory theories of Artificial Intelligence' at the Centre Responsible Digitality (ZEVEDI). This work is part of the research programme Ethics of Socially Disruptive Technologies, which is funded through the Gravitation programme of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

Competing interests. The author declares none.

References

Awad, Edmond, et al. 2018. The Moral Machine Experiment. *Nature*, **563**: 59–78.

Baltzly, Vaughn Bryan. 2021. Trolleyology as First Philosophy: A Puzzle-Centered Approach to Introducing the Discipline. *Teaching Philosophy*, **44**.4: 407–48.

Baumann, Peter. 2022. Trolleys, Transplants and Inequality: An Egalitarian Proposal. *Erkenntnis*, **87**.4: 1737–51.

Berker, Selim. 2009. The Normative Insignificance of Neuroscience. *Philosophy and Public Affairs*, **37**.4: 293–329.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press).

Brändle, Claudia, and Schmidt, Michael W. 2021. Autonomous Driving and Public Reason: A Rawlsian Approach. *Philosophy & Technology*, **34**.4: 1475–99.

Bruers, Stijn, and Braeckmann, Johan. 2014. A Review and Systematization of the Trolley Problem. *Philosophia*, **42**.2: 251–69.

Danaher, John. 2019. *Automation and Utopia: Human Flourishing in a World without Work* (Cambridge, MA: Harvard University Press).

Daniels, Norman. 1979. Wide Reflective Equilibrium and Theory Acceptance in Ethics. *The Journal of Philosophy*, **76**.5: 256–82.

Davnall, Rebecca. 2020. Solving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics. *Science and Engineering Ethics*, **26**.1: 431–49.

Etzioni, Amitai, and Etzioni, Oren. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, **21**.4: 403–18.

Foot, Philippa. 1967. The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*, **5**: 5–15.

Fried, Barbara H. 2012. What *Does* Matter? The Case for Killing the Trolley Problem (Or Letting It Die). *The Philosophical Quarterly*, **62**.248: 505–29.

Gogoll, Jan, and Müller, Julian F. 2017. Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Science and Engineering Ethics*, **23**.3: 681–700.

Goodall, Noah J. 2016. Away from Trolley Problems and Toward Risk Management. *Applied Artificial Intelligence*, **30**.8: 810–21.

Graham, Peter A. 2017. Thomson's Trolley Problem. *Journal of Ethics and Social Philosophy*, **12**.2: 168–90.

Greene, Joshua. 2008. The Secret Joke of Kant's Soul. In *Moral Psychology: Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, ed. by Walter Sinnott-Armstrong (Cambridge, MA: MIT Press): 35–80.

Greene, Joshua. 2014. Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics. *Ethics*, **124**.4: 695–726.

Greene, Joshua. 2016. Solving the Trolley Problem. In *A Companion to Experimental Philosophy*, ed. by Justin Sytsma and Walter Buckwalter (Malden, MA: Wiley Blackwell): 175–89.

Greene, Joshua, et al. 2009. Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment. *Cognition*, **111**.3: 364–71.

Hansson, Sven Ove, Belin, Matts-Åke, and Lundgren, Björn. 2021. Self-Driving Vehicles: An Ethical Overview. *Philosophy and Technology*, **34**.4: 1383–1408.

Haslett, D.W. 2011. Boulders and Trolleys. *Utilitas*, **23**.3: 268–87.

Henry, Jim. 2020. VW Exec: Level 4 Self-Driver May Be as Good as It Gets, <https://www.wardsauto.com/ces/vw-exec-level-4-self-driver-may-be-good-it-gets>.

Himmelreich, Johannes. 2018. Responsibility for Killer Robots. *Ethical Theory and Moral Practice*, **22**.3: 731–47.

Hübner, Dietmar, and White, Lucie. 2018. Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us Beyond Harm Minimisation. *Ethical Theory and Moral Practice*, **21**.3: 685–98.

Huemer, Michael. 2005. *Ethical Intuitionism* (London: Palgrave Macmillan).

Huemer, Michael. 2018. *Paradox Lost: Logical Solutions to Ten Puzzles of Philosophy* (London: Palgrave Macmillan).

Hurka, Thomas. 2016. Trolleys and Permissible Harm. In *The Trolley Problem Mysteries*, ed. by Eric Rakowski (Oxford: Oxford University Press): 135–50.

Jin, Hyunjoo, and Balu, Nivedita. 2022. Musk's Bets on Tesla: Human-like Robots and Self-driving Cars, <https://www.reuters.com/technology/musks-bets-tesla-no-human-drivers-this-year-robots-next-2022–01–27/>.

Kagan, Shelly. 2016. Solving the Trolley Problem. In *The Trolley Problem Mysteries*, ed. by Eric Rakowski (Oxford: Oxford University Press): 151–68.

Kahane, Guy. 2013. The Armchair and the Trolley: An Argument for Experimental Ethics. *Philosophical Studies*, **162**.2: 421–45.

Kamm, Frances. 1989. Harming Some to Save Others. *Philosophical Studies*, **57**.3: 227–60.

Kamm, Frances. 1992. *Creation and Abortion: A Study in Moral and Legal Philosophy* (Oxford: Oxford University Press).

Kamm, Frances. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm* (Oxford: Oxford University Press).

Kamm, Frances. 2016. *The Trolley Problem Mysteries*, ed. by E. Rakowski (Oxford: Oxford University Press).

Kamm, Frances. 2020. The Use and Abuse of the Trolley Problem: Self-Driving Cars, Medical Treatments, and the Distribution of Harm. In *Ethics of Artificial Intelligence*, ed. by S. Matthew Liao (New York: Oxford University Press): 79–108.

Keeling, Geoff. 2020. Why Trolley Problems Matter for the Ethics of Automated Vehicles. *Science and Engineering Ethics*, **26**.1: 293–307.

Kleingeld, Pauline. 2020. A Kantian Solution to the Trolley Problem. In *Oxford Studies in Normative Ethics: Volume 10*, ed. by Mark Timmons (Oxford: Oxford University Press): 204–28.

Königs, Peter 2020. Experimental Ethics, Intuitions, and Morally Irrelevant Factors. *Philosophical Studies*, **177**.9: 2605–23.

Larson, Erik J. 2021. *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do* (Cambridge, MA: Harvard University Press).

Leben, Derek. 2017. A Rawlsian Algorithm for Autonomous Vehicles. *Ethics and Information Technology*, **19**.2: 107–15.

Lin, Patrick. 2016. Why Ethics Matters for Autonomous Cars. In *Autonomous Driving: Technical, Legal and Social Aspects*, ed. by Markus Maurer et al. (Berlin: Springer): 69–85.

Lundgren, Björn. 2021. Safety Requirements vs. Crashing Ethically: What Matters Most for Policies on Autonomous Vehicles. *AI & Society*, **36**.2: 405–15.

Muoio, Danielle. 2017. Automakers are Slowing Their Self-Driving Car Plans – and That Could be a Good Thing, <https://www.businessinsider.com/self-driving-cars-not-feasible-in-5-years-automakers-say-2017–1>.

Nyholm, Sven. 2018a. The Ethics of Crashes with Self-Driving Cars: A Roadmap, I. *Philosophy Compass*, **13**.7.

https://doi.org/10.1017/S0953820822000395 Published online by Cambridge University Press

Nyholm, Sven. 2018b. The Ethics of Crashes with Self-Driving Cars: A Roadmap, II. *Philosophy Compass*, **13**.7.

Nyholm, Sven, and Smids, Jilles. 2016. The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice*, **19**.5: 1275–89.

Otsuka, Michael. 2008. Double Effect, Triple Effect and the Trolley Problem: Squaring the Circle in Looping Cases. *Utilitas*, **20**.1: 92–110.

Paulo, Norbert. 2020. The Unreliable Intuitions Objection Against Reflective Equilibrium. *The Journal of Ethics*, **24**.3: 333–53.

Paulo, Norbert (ms.). The Trolley Problem and the Ethics of Autonomous Vehicles. Unpublished Manuscript.

Railton, Peter. 2014. The Affective Dog and Its Rational Tale: Intuition and Attunement. *Ethics*, **124**.4: 813–59.

Rawls, John. 1975. The Independence of Moral Theory. *Proceedings and Addresses of the American Philosophical Association*, **48**: 5–22.

Savulescu, Julian, Gyngell, Christopher, and Kahane, Guy. 2021. Collective Reflective Equilibrium in Practice (CREP) and Controversial Novel Technologies. *Bioethics*, **35**.7: 652–63.

Schmidt, Elke Elisabeth. 2022. Kant on Trolleys and Autonomous Driving. In *Kant and Artificial Intelligence*, ed. by Hyeongjoo Kim and Dieter Schönecker (Boston and New York: De Gruyter): 189–221.

Susskind, Daniel. 2020. *A World Without Work: Technology, Automation and How We Should Respond* (London: Penguin).

Tersman, Folke. 2008. The Reliability of Moral Intuitions: A Challenge from Neuroscience. *Australasian Journal of Philosophy*, **86**.3: 389–405.

Thomson, Judith Jarvis. 1976. Killing, Letting Die, and the Trolley Problem. *The Monist*, **59**.2: 204–17.

Thomson, Judith Jarvis. 1985. The Trolley Problem. *The Yale Law Journal*, **94**.6: 1395–1415.

Thomson, Judith Jarvis. 2008. Turning the Trolley. *Philosophy and Public Affairs*, **36**.4: 359–74.

Thomson, Judith Jarvis. 2016. Kamm on the Trolley Problem. In *The Trolley Problem Mysteries*, ed. by Eric Rakowski (Oxford: Oxford University Press): 113–34.

Unger, Peter. 1997. *Living High and Letting Die: Our Illusion of Innocence* (New York: Oxford University Press).

Wallach, Wendell, and Allen, Colin. 2009. *Moral Machines: Teaching Robots Right From Wrong* (Oxford: Oxford University Press).

Wolkenstein, Andreas. 2018. What has the Trolley Dilemma ever done for us (and what will it do in the Future)? On Some Recent Debates about the Ethics of Selfdriving Cars. *Ethics and Information Technology*, **20**.3: 163–73.

Wood, Allen. 2011. Humanity as an End in Itself. In *On What Matters: Volume 2*, ed. by Derek Parfit (Oxford: Oxford University Press): 58–82.

Cite this article: Königs P (2023). Of trolleys and self-driving cars: What machine ethicists can and cannot learn from trolleyology. *Utilitas* **35**, 70–87. https://doi.org/10.1017/S0953820822000395