This is the accepted version. The final/published version appears on pages 211-230 in Lillehammer, H. (ed.), *The Trolley Problem*. Cambridge: Cambridge University Press, 2023

Ethical Accident Algorithms for Autonomous Vehicles and the Trolley Problem: Three Philosophical Disputes

Sven Nyholm

In this chapter I discuss whether it is helpful for those interested in the real-world ethics of crashes involving self-driving cars to compare that set of ethical issues with the trolley problem. What the phrase "the trolley problem" refers to should be clear to readers of this book, but it is something we will have occasion to return to below.[1] The usefulness of comparing real-world ethical issues concerning self-driving cars to the standard cases and philosophical issues associated with the trolley problem is controversial among philosophers who have written on this topic (e.g. Nyholm & Smids 2016; JafariNaimi 2018; Himmelreich 2018; Keeling 2019; Kamm 2020). In this chapter, however, my most general thesis is that it *is* instructive to reflect on the comparison between crashes with self-driving cars and the trolley problem and the examples associated with it. Indeed, as I see things, it is almost impossible for it not to be useful to compare the ethics of

[1] I follow Judith Thomson (2015) and Frances Kamm (2015) in taking "the trolley problem" to refer, not to any particular dilemma case involving a runaway trolley, but rather to the philosophical question(s) raised by such cases. According to Kamm (2015), the basic philosophical problem is this: why are certain people, using certain methods, morally permitted to kill a smaller number of people to save a greater number, whereas others, using other methods, are not morally permitted to kill the same smaller number to save the same greater number of people? For example, it is thought to be permissible for the bystander to save five people by sacrificing one person in the standard "switch" trolley case that we will get to below. But why is it not permissible for a medical doctor to save five patients in need of organ-transplants by "harvesting" five organs from a perfectly healthy patient who just came into the hospital for a routine check-up? The transplant case doesn't mention trolleys. But Kamm thinks that it nevertheless falls under the wide umbrella of the trolley problem.

crashing self-driving cars with the trolley problem. It is either directly or indirectly useful. It can be directly useful because the trolley problem itself brings up ethical issues that are of immediate importance for the ethics of self-driving cars. Or it is at the very least indirectly useful because the process of highlighting key differences between the real-life ethics of self-driving cars and the philosophy of the trolley problem is a good way of clarifying what matters most for the ethics of self-driving cars.

This chapter divides into the following sections. First, I provide some more background and context for my discussion (section 1). Next, I divide up my discussion into three main segments, each of which considers what I will call a particular "philosophical dispute" that one finds in the literature regarding whether it is useful to discuss the trolley problem within the ethics of self-driving cars. The first dispute is about whether there is something flippant, or perhaps even downright immoral, about comparing real-world ethics with the trolley problem (section 2). The second dispute is about whether the disanalogies between trolley problem cases and crashes involving self-driving cars in the real world are significant enough to make this comparison unhelpful (section 3). The third dispute is about whether the academic literature on the trolley problem has been about rather different issues than those that matter the most in relation to the real-world ethics of crashes involving self-driving cars (section 4). Having introduced these three disputes, I turn to the question of what we should make of all of this (section 5). I argue that it partly depends on whether we think one should be a monist or a pluralist about methodology in ethics: i.e., whether one thinks only one method should be used or whether one thinks it is best to combine insights from multiple methods. My view is that we should be pluralists about

methodology, and that this is part of why the self-driving cars/trolley problem comparison cannot help but being either directly or indirectly relevant to the real-world ethics of self-driving cars.

The last thing I will mention in this introduction is that whether or not there will ever be many real-life crash scenarios involving self-driving cars that are very similar to the cases associated with the trolley problem, there have already been many real-world crashes involving self-driving cars. And some of these have been fatal, both to people in- and outside of the self-driving cars. Already in 2015, there were about 20 minor crashes involving self-driving cars. Nobody was seriously injured, and what happened was simply that people driving regular cars bumped into self-driving cars (Schoettle & Sivak, 2015). In 2016, however, a crash involving a self-driving car and a bus was clearly caused by an experimental self-driving car operated by Google (LeBeau 2016). Later in that same year, the first person died in a self-driving car when his Tesla car operating in "autopilot" mode crashed into a white truck that the car's sensors did not properly distinguish from the bright sky (Tesla 2016). In 2018, a pedestrian was, for the first time, struck and killed by a self-driving car. Elaine Herzberg was crossing the street in Tempe, Arizona, when an experimental self-driving car operated by the ride-hailing service company Uber drove into her, leaving her with fatal injuries (Levin & Wong 2018). There have also been other serious accidents involving cars with different levels and kinds of automation. So, in other words, ethical questions about crashes involving self-driving cars are real-world issues. They are very serious issues; human lives are at stake. It can be hard to discuss the trolley problem without discussing comical and absurd scenarios. It

can be fun, and also instructive, to do so. But as we do, we should keep in mind that real crashes involving self-driving cars are no joke.

1: Background

It is easy to think of examples involving crashing self-driving cars that bear at least a superficial resemblance to the thought experiments commonly associated with the trolley problem. Notably, we can distinguish between (a) examples with crashing self-driving cars that are rather loosely based on or similar to the examples associated with the trolley problem and (b) examples that are closely modelled on the trolley problem examples. The former are more common than the latter. More on this below.

In the most famous trolley problem case[2], a trolley is about to drive into five people on a train track, and it is possible for a bystander to save the five by pulling a switch and redirecting the trolley onto a side track. On that other track, there is one person, who will be hit and killed if the train is redirected. In the second most well-known variation of the example, there is no side track. But there is a large and heavy person up on a footbridge who could be pushed off the bridge down onto the tracks. His weight is hefty enough to set off the automated breaks of the trolley. This would save the five, but kill the large person (Kamm 2015). Similar to these kinds of trolley examples, we can imagine cases in which, for

[2] I am using the phrase "trolley problem cases" to refer to the examples discussed in the literature about the trolley problem, such as the cases above. For a review of many of the most relevant cases – and the perhaps most thorough discussion of the trolley problem to date – see Kamm 2015. For the history of the trolley problem, see Edmonds 2013.

instance, self-driving cars are about to drive into five people on the road, but where, say, another person would be hit and killed if the self-driving car turned onto a side road. Just as philosophers have imagined numerous forced dilemmas where people are killed by out-of-control trolleys, we can also imagine different variations of bad outcomes involving people being hit and killed by self-driving cars facing forced dilemmas (Davnall 2019: 432-433).

This observation has sparked the imagination of academics and non-academics alike. It was picked up by the media around 2015, and has since then often been revisited in mass media headlines and articles. By simply picking some of these headlines, one can tell a sort of "rise and fall" story or discern a "hype cycle" with respect to how publicly accessible debates about this topic have developed in recent years:

> *Driverless Cars Are Colliding with the Creepy Trolley Problem* (Washington Post 2015)
>
> *Why Mercedes Plans to Let Its Self-Driving Cars Kill Pedestrians in Dicey Situations* (Business Insider 2016)
>
> *Google's Chief of Self-Driving Cars Downplays 'The Trolley Problem'* (Washington Post 2016)
>
> *MIT Study Explores the 'Trolley Problem' and Self-Driving Cars* (Venture Beat 2018)

> *Should a Self-Driving Car Kill the Baby or The Grandma? Depends on Where You're From. The Infamous "Trolley Problem" was Put to Millions of People in A Global Study, Revealing How Much Ethics Diverge Across Cultures*
> (Technology Review 2018)

> *Trolley Dilemmas Shouldn't Influence Self-Driving Policies, Experts Argue*
> (Robotics Business Review 2019)

A lot of these articles were prompted by the MIT study referred to in two of these just-quoted headlines: the so-called "moral machine experiment" (Bonnefon et al. 2016; Awad et al. 2018). This was an enormous survey of intuitions about dilemma cases that was run – not by moral philosophers – but by a team of psychologists and behavioral economists. It was a study inspired by the extensive empirical investigations of ordinary people's intuitions about trolley problem cases that had been carried out by psychologists, philosophers, and legal researchers in the previous few years (e.g., Greene 2013). As the authors of the moral machine study put it in one of their first publications on this topic:

> situations of unavoidable harms, as illustrated in [our examples of crashes with self-driving cars], bear a striking resemblance with the flagship dilemmas of experimental ethics – that is, the so-called 'trolley problem'. (Bonnefon et al. 2015: 3)

What Bonnefon and colleagues did was to create a large set of vignettes (basically, cartoon-like images showing two options, where different driving paths of self-

driving cars involved killing different people), which were made available on the moral machine website[3]. People were asked to judge what the self-driving cars in the vignettes should do (e.g., kill three grandfathers on the left, or go right and kill four toddlers). The responses were collected and patterns in people's intuitions identified.[4]

One interesting finding of this study – as one of the headlines above indicates – is that there are differences among countries in the patterns of people's intuitions, for example with respect to whether the young should be prioritized over the old, or whether breaking traffic laws should make one more liable to being killed by the self-driving car (Awad et al. 2018). Another striking finding from the same lab is that people have different attitudes regarding how their own self-driving cars should be programmed (they should always save the person riding in the car, even if this does not minimize overall harm) and how other people's cars should be programmed (they should minimize overall harm even if it is detrimental to the person riding in the car) (Bonnefon et al. 2015; 2016). This research team has also offered a normative argument for why they think we should engage in this kind of research. But let us save that until later.

The next thing I will note in this section is that it is not only the mass media and empirical researchers who make this comparison between crashes with self-driving cars and the trolley problem. Perhaps more important for our current purposes – since we are looking at three philosophical disputes about this comparison – is that many philosophers have also made and endorsed this

[3] https://www.moralmachine.net

[4] Sütfeld et al. (2017) have pursued a fascinating related line of experimental ethics by examining how people respond to AV collisions presented in virtual reality. Their idea is that we can model people's moral preferences by fitting predictive models for their decisions based on relevant features of the collision, e.g. whether it involves a person or a non-human animal.

comparison, again in more or less loose ways. Here are two quick quotes to illustrate this. Patrick Lin, in one of the earliest philosophical papers about the ethics of self-driving cars, wrote the following:

> One of the most iconic thought-experiments in ethics is the trolley problem. …and this is one that may now occur in the real world, if autonomous vehicles come to be. (Lin 2015: 78)

Similarly, when discussing another kind of autonomous vehicles (namely, driverless trains), the authors of *Moral Machines: Teaching Robots Right from Wrong,* Wendell Wallach and Colin Allen write:

> …could trolley cases be one of the first frontiers for artificial morality? Driverless systems put machines in the position of making split-second decisions that could have life or death implications. As the complexity [of the traffic] increases, the likelihood of dilemmas that are similar to the basic trolley case also goes up. (Wallach and Allen 2009: 14)

Geoff Keeling (2019), in turn, has published a powerful article-length defense of this comparison entitled "Why Trolley Problems Matter for the Ethics of Automated Vehicles". So, while most philosophers who make this comparison usually do so in a quick and underdeveloped way, there are also those who have devoted whole articles to defending this idea (see also Wolkenstein 2018).

Moreover, it is perhaps also interesting to note here that going back even further in time, one of the most important contributors to the philosophy of the

trolley problem, Frances Kamm (1996), once imagined a self-driving ambulance. This ambulance had to be preprogrammed to either always prioritize getting dying patients to the hospital as quickly as possible, even if this would mean crashing into pedestrians, or to never do so. The question in that example was whether the former kind of programming could ever be justified. As we will see below, Kamm has since then joined the discussion of real self-driving cars and the trolley problem, and she thinks there are important differences between what she was interested in while discussing her ambulance case and the real-world ethics of self-driving cars. But it is nevertheless striking that Kamm already came up with an example in her influential 1996 book *Morality, Mortality* that had some similarity to the topic of this chapter.

In short, the idea of comparing ethical questions regarding crashing self-driving cars with the trolley problem is an idea that has not only fascinated philosophers interested in the ethics of self-driving cars. It is also an idea that has resonated with many people outside of philosophy. And millions of people all over the world have been surveyed about their attitudes regarding these kinds of examples. This makes it important to reflect on whether it is helpful to the ethics of self-driving cars to make this comparison. Let us now turn to the three philosophical disputes about this issue that I want to discuss in this chapter.

2: First Philosophical Dispute: Is Making This Comparison Perhaps Morally Required, or Is There Something Morally Problematic About the Trolley Problem/Self-Driving Cars Comparison?

Bonnefon et al. (2015: 2016) and Awad et al (2018)[5] argue that their moral machine experiment testing people's intuitions about trolley problem-inspired cases is not only in itself interesting and motivated by the academic interest in studying patterns in ordinary people's intuitions about trolley-like scenarios involving self-driving cars. It might also be morally required. As they see things, given the potential that self-driving cars might (eventually) become much safer than regular cars, it is important that the general public should accept – and be willing to use – self-driving cars. For this goal to be achieved, self-driving cars should be programmed to handle accident scenarios in ways that fit with how the general public finds it acceptable for autonomously operating cars to handle such scenarios. Hence the need for the moral machine experiment and its surveys of people's intuitions about how self-driving cars should crash when crashes are unavoidable. We should not leave this to ethical reflection by academics alone. As they themselves put it:

> even if ethicists were to agree on how autonomous vehicles should solve moral dilemmas, their work would be useless if citizens were to disagree with their solution, and thus opt out of the future that autonomous vehicles promise in lieu of the status quo. Any attempt to devise artificial intelligence ethics must be at least cognizant of public morality (Awad et al. 2018: 59).

[5] To be clear: Bonnefon et al. and Awad et al. are members of the same research team. So, these references refer to work by the same group of researchers, not to two separate teams.

Some philosophers who have responded to this idea, however, have argued that there is something inherently ethically problematic about this whole approach and the mindset behind it. John Harris (2020), in particular, responds directly to Bonnefon et al. and their "moral machine experiment" in very sharp terms. Nassim JafariNaimi (2018) responds more generally to the idea of considering trolley problem-inspired moral dilemmas in the context of ethical reflection on self-driving cars, reflecting concerns similar to those expressed by Allen Wood (2011) and others in more general critical discussions of trolley problem-like approaches to moral reasoning.

As I understand it, Harris's (2020) response to the way that Bonnefon et al. compare the ethics of self-driving cars to the experimental approach to the trolley problem has two main parts. Firstly, how issues of life and death are handled in society should not be based on people's gut reactions to cartoon-like vignettes, but on careful deliberations and legal processes. Secondly, the idea that self-driving cars should "target" some people rather than others in accident scenarios seems to suggest that they should make judgments about who lives or dies, which Harris finds highly problematic (cf. Purves et al. 2015). Life and death decisions, Harris argues, are serious matters. Decisions about how society should deal with life and death decisions should be the outcomes of slow and careful legal and moral deliberations, which are allowed to take time. And we should not arrive at a situation where the AI in machines like self-driving cars are allowed to "punish" certain people and condemn them to death.

JafariNaimi, in turn, argues that the idea of comparing ethical questions about life and death decisions involving self-driving cars to trolley dilemmas involves an objectionable "utilitarian" framing, reducing all ethical issues to

numbers and quantities, whereas in real life, the ethics of self-driving cars is much more complicated (JafariNaimi 2017: 303). She summarizes her overall criticism as follows:

> First, ethical situations are marked by a deep sense of uncertainty and an organic character. Second, our place within ethical situations matters greatly. Third, the impact of our actions in response to ethical situations is not limited to immediate outcomes[;] consequences are broad and long ranging. Therefore … principles that appear to solve the scenarios of experimental ethics may or may not serve similar ethical situations encountered in real life. (JafariNaimi, 2018: 306).

These criticisms anticipate some of the other disputes about the trolley problem/self-driving cars ethics comparison that we will get to below. But what I in particular want to highlight in this section is the overall message from JafariNaimi that there is something reductive/over-simplifying, and therefore insensitive, about thinking that the ethics of crashes and risks involving self-driving cars could be adequately accounted for if we try to tackle these real-world issues by consulting either our own or the general public's intuitions about stylized dilemmas where self-driving cars have to "choose" whom to crash into.

These worries about there potentially being something morally problematic about comparing the ethics of crashing self-driving cars with the trolley problem echo more general ethical worries about the trolley problem that some critics have expressed in more general discussions. Wood, for example, approvingly cites a Tanner Lecture on Human Values from 2001, where a novelist,

Dorothy Allison, had commented on the trolley problem, which she said she had heard about from some philosophers she knew. In her lecture, Allison said that her reaction was to reject the problem itself and to refuse to form an opinion about it. She focused on what she called "lifeboat cases" (cf. Gibbard 2008, chapter two). Commenting on choices regarding whether to save five people or one person if there is only one lifeboat and one can only go to the five or to the one, Allison said that there was something immoral about thinking about the problem in this crass way. The only ethically appropriate question was why provision had not been made to make available more lifeboats to begin with. Wood approvingly remarks that this reaction from Allison can be applied to the cases usually discussed in relation to the trolley problem, the novelist's reaction being "far more sensible and right-minded than what we usually get from most of the philosophers who make use of such examples." (Wood 2011: 67)

Wood writes, furthermore, that in relation to many trolley cases, "the right reaction is to regard it as simply indeterminate what the agent should do, and the only real moral issue raised by the problem is . . . how the situation in question was permitted to arise in the first place." (ibid.: 72) Because "even if some choices do inevitably have the consequence that either one will die or five will die, there is nearly always something wrong with looking at the choice only in that way." (ibid.: 73) Given that this is what Wood has to say about the traditional trolley problem, one can only imagine what he might say about the trolley problem-inspired moral dilemmas that Bonnefon et al. depict when they present the general public with moral dilemmas involving self-driving cars.

What's at issue here, in other words, is whether there is something frivolous, inherently insensitive, misguided, or perhaps downright immoral about

reflecting on these kinds of forced dilemma scenarios – whether they involve crashing self-driving cars or runaway trolleys. Harris and JafariNaimi take this view with respect to the self-driving cars issue. Allison and Wood take this view in relation to trolley cases. This clashes starkly with Bonnefon et al.'s view that it would be morally problematic not to engage in trolley problem-like research about patterns in people's intuitions regarding different cases involving crashing and life-threatening self-driving cars, since making self-driving cars acceptable to people requires programming them in ways that fit with how ordinary people think that self-driving cars should be programmed to handle accident scenarios.

3: Second Dispute: Are Any Real-Life Crashes Involving Self-Driving Cars Relevantly Similar to The Examples Associated with the Trolley Problem?

The second and third philosophical disputes regarding the comparison of the ethics of self-driving cars and the philosophy/psychology of the trolley problem are closely related. In fact, back in 2015 when it became very popular to compare crashing self-driving cars with runaway trolleys in what was sometimes fast and loose ways, it struck me and my collaborator Jilles Smids that somebody should write a philosophical article examining skeptically whether the analogy between the ethics of self-driving cars and the trolley problem is as close as many academics and others were making it out to be. We learned shortly thereafter that Noah Goodall had had the same thought, and the year after, we – as well as Goodall

– had papers out on this (Nyholm & Smids 2016; Goodall 2016).[6] When we wrote our pieces, we didn't draw a sharp distinction between the issue of whether real-world crashes involving self-driving cars are interestingly similar to trolley problem examples, on the one hand, and the issue of whether the trolley problem literature has treated the ethical issues most relevant to the ethics of crashing self-driving cars, on the other hand. The two issues are clearly closely related. But here I am focusing on the former issue, and in the next section on the latter.

When it comes to whether any real-world crashes involving self-driving cars are sufficiently analogous with the examples associated with the trolley problem for it to be worth making this comparison, there are those both in- and outside of philosophy who deny this. Outside of philosophy, as one might expect, representatives from the car and technology industries have bemoaned philosophers' and psychologists' comparisons between the trolley problem and crashes with self-driving cars, claiming that they are very unlikely to happen, for which reason they say that it is stifling innovation to make such comparisons. (Recall the "*Google Chief of Self-Driving Cars Downplays the 'Trolley Problem'*" headline quoted above!) Inside of philosophy, interventions from Johannes Himmelreich (2018) and Rebecca Davnall (2019) stand out when it comes to what Keeling (2019) calls the "not going to happen" objection to comparing crashing self-driving cars to the trolley problem. According to Himmelreich, if self-driving cars drive so fast that crashes become unavoidable and tragic choices have to be made, then the cars will not be able to make meaningful choices quickly enough

[6] Alexander Hevelke and Julian Nida-Rümelin had already briefly discussed whether we should compare the ethics of self-driving cars to the trolley problem, and expressed skepticism about this, in their 2015 article about who should be held responsible when self-driving cars crash. But they made some brief remarks about this in passing, and the main focus on their article lay elsewhere.

that it is possible to program in any particular forms of responses into how the cars function. According to Davnall, if we are considering a self-driving car that is otherwise operating normally (e.g., there is nothing wrong with the breaks or anything like that), and a crash is unavoidable because there are cars or people within the car's breaking distance, then it will always be safest to simply break very hard rather than to try to veer off in any other direction.

Keeling (2019: 295-296), in contrast, responds to this supposed lack of realism by calling the examples we use when we imagine self-driving cars facing trolley problem-like scenarios theoretical "idealizations". The idea, according to Keeling, is not that trolley problem-like scenarios involving self-driving cars are likely to happen. It is rather that we can think about such cases in order to get clear on what is important to us – in order, in other words, to get clear on what our priorities are. Just like the ideal gas law in physics describes an idealization used for theoretical purposes – and does not describe the behavior of all actual gases in the real world – so do trolley problem cases involving self-driving cars describe idealized scenarios that we consider for theoretical, not practical, purposes, according to Keeling. We can think, in other words, that our moral ideas about what values should guide the programming of self-driving cars could be sharpened by, or benefit from, considering trolley problem cases even if we do not think that these are likely to occur in real life (cf. Goodall 2016). The comparison between imagined cases involving crashing self-driving cars and the cases associated with the trolley problem is not important, on this view, because of the realism of these cases, but because of the role(s) they can play in theorizing about important ethical issues.

Setting aside the realism issue as it relates to the likelihood of real-world instances of trolley problem-like crashes with self-driving cars, though, we can also ask whether there are other important disanalogies between cases involving self-driving cars in traffic and out-of-control trolleys in philosophical thought experiments.[7] One thing that Jilles Smids and I in our above-mentioned article – and also Goodall in his above-mentioned piece – highlighted and presented as a key difference here concerns the distinction between decision-making in the face of uncertainty and on the basis of assessments of risk, on the one hand, and decision-making about certain and known outcomes, on the other hand (Nyholm & Smids 2016; Goodall 2016). This also resonates with JafariNafari's above-cited claim that real-world ethical choices are often marked by "a deep sense of uncertainty."

In the trolley problem cases, we imagine that we know for certain that the five can be saved if the trolley is redirected to the side track where the one is standing, or if the large and heavy person is pushed onto the track, and so on. So, the question is simply what is right to do, given those certain and known facts. In stark contrast, when choices are made about how self-driving cars should be programmed to deal with accident-scenarios, we are dealing with the real world, which means dealing with uncertainty and making risk assessments about what might happen with some degree of probability and some unknown magnitude of harm. In a more general discussion of the relevance of the trolley problem, Sven Ove Hansson (2012: 44) complains about the trolley problem as a way of modelling real-world ethical decision making because it does not involve any

[7] For further apparent disanalogies, see also the discussion in Gogoll and Müller 2017, especially p. 690.

uncertainty nor any assessments of risks. If those are aspects of most real-world moral decision making, and we want our theorizing about the ethics of crashes involving self-driving cars to involve the sort of considerations we have to take into consideration in real-world decision making, we have an important disanalogy here between the ethics of crashing self-driving cars and the thought experiments associated with the trolley problem (Nyholm & Smids 2016: 1284-1286).

Another thing – Smids and I also argued – that distinguishes the real-world ethics of crashing self-driving cars from the philosophy of the trolley problem concerns issues of moral and legal responsibility (ibid., 1282-1284). Think about how it is when philosophy teachers present the trolley problem to their students during ethics courses. Often, what happens is that one of the students will raise their hand and ask whether it wouldn't be the case that one would go to jail if one pushed a large person off a bridge to his death in order to save five people on the tracks, or even if one redirected a train onto a side track where one person is hit and killed by the trolley. What philosophy teachers usually do when they get this very good question is to tell the student to set any such issues about legal or moral responsibility aside and simply focus on what the right or best choice to make is in the circumstances, here and now – wholly independent of any further consequences or any worries about who is responsible for what. This is another thing that makes the philosophy of the trolley problem very different from real-world cases involving crashes with self-driving cars. When it comes to the latter, issues related to legal and moral responsibility are inescapable.

Just think of the real-world crashes mentioned in the introduction: the Google car that collided with a bus; the Tesla car that crashed into a truck and

killed the person in the car; and the experimental Uber car that hit and killed a pedestrian. In all of these cases, questions of responsibility were immediately raised. Google admitted partial responsibility for the crash involving their car (LeBeau 2016). Tesla released a statement denying all responsibility, but expressed their sympathy for the victim (Tesla 2016). Uber tried to evade legal accountability by proposing a financial settlement to the family of the woman who was hit and killed by their self-driving car (Wakabayashi & Conger 2018). In the Tesla and Uber cases, many commentators felt that those companies had not properly been held legally responsible for what happened. The question of how to move forward in a responsible way was raised in all of these cases. Google promised to update the software in their cars to make them better able to predict the behavior of buses. Tesla promised to update their sensors to make them better able to detect white trucks on sunny days. And Uber temporarily ceased their testing of self-driving cars in Tempe, where the deadly accident happened. (For further discussion, see Nyholm 2020, chapter 3.)

The general point here is that in the real world, ethically salient decisions and incidents causing harm and potentially death are always intimately tied to questions of responsibilities, duties of care, and other issues pertaining to how we are related to those around us, not just in the immediate present, but also over time, and as members of a shared society (JafariNaimi 2016). In trolley problem reasoning intended to pump intuitions about moral principles, we are asked to set such contextual considerations and responsibility-related issues aside (Kauppinen 2020). This can be seen as a rather stark contrast. The second philosophical dispute about the self-driving cars/trolley problem comparison is about whether these differences between real-world ethical issues related to crashing self-driving

cars are stark enough that comparing crashes with self-driving cars with trolley problem cases is perhaps interesting in the abstract, but not obviously and clearly relevant to the real-world ethics of our future with self-driving cars. Here too one could bring up Keeling's (2019) point about idealizations for purely theoretical purposes versus attempts to depict realistic situations lining up perfectly with real-life case studies. But I will set that point aside for now. Let us now instead turn to the third philosophical dispute I want to bring up.

4: Third Philosophical Dispute: Is the Literature about the Trolley Problem Relevant to the Ethics of Crashes involving Self-Driving Cars?

When Smids and I wrote our 2016 article, one of the things we were asking ourselves was whether the philosophical and psychological literature about the trolley problem has been concerned with the sorts of issues that are most relevant to the ethics of crashes with self-driving cars. We wrote:

> the key issues . . . of great importance for the ethics of accident-algorithms for self-driving cars are typically not discussed in the main literature on the trolley problem. For example, this literature is not about the risks or the legal and moral responsibilities we face in traffic. On the other hand, the main issues that the literature on the trolley problem does engage directly with have to do with rather different things than those . . . most pressing for the ethics of accident-algorithms for self-driving cars. . . [T]his literature discusses things such as: the ethical differences between positive and negative duties and killing and letting die, and psychological and neuro-

> scientific theories about how different types of moral judgments are generated by our minds and brains. (Nyholm & Smids 2016: 1276)

Taking those considerations together, we concluded that the literature on the trolley problem is not the best, nor perhaps even a particularly good, place to turn to for source materials and precedents directly useful for the ethics of accident-algorithms for self-driving cars (cf. Cunneen et al. 2019). Others have chimed in with similar conclusions. Antti Kauppinen (2020), for example, argues that an important difference between what is (or is not!) discussed in the trolley problem literature and what we should discuss when we think about how self-driving cars should handle crash scenarios has to do with whether people are liable in relation to risky situations that they are part of.[8] If an accident scenario is caused by the reckless behavior of one party – which can often happen in real traffic – it can seem morally fitting that they bear a greater risk in the resolution of the dangerous situation than somebody who was taking all appropriate precautions in their traffic behavior.

Kamm (2020) makes virtually the same point in her recent paper about what she calls "uses and abuses of the trolley problem". She notes that in the trolley problem as she and others have discussed it, the people at risk (e.g., the five on the tracks or the large and heavy man on the footbridge) are no different from each other in terms of whether their being at risk is their own fault. Kamm, like Kauppinen, thinks that in cases of real-world car crashes, we cannot similarly

[8] For more general discussions of that type of reasoning concerning risks people are (partly) responsible for and what their responsibility does to their liability to be harmed, see, e.g, McMahan 2005; 2009 and Frowe 2015.

assume that everyone is equally innocent in this way. We must instead count on its being possible that some are more liable to be harmed than others. In addition, then, to not being about crucial topics such as risk and uncertainty, and legal and moral responsibility – all of which are highly relevant to the ethics of self-driving cars – the trolley problem literature has also not been about the important issue of greater liability to be harmed because one bears more responsibility than others in creating a risky situation.

Kamm also turns this on its head by noting that the kinds of cases some philosophers and psychologists compare to the trolley problem fail to track the philosophical concerns that she and others who have discussed the trolley problem have been particularly interested in (ibid.). The trolley problem, Kamm argues, involves examples that have been very carefully engineered to serve certain illustrative purposes, e.g., teasing out certain ethical distinctions. And many of the envisioned cases involving crashing self-driving cars or other forced dilemmas fail to track the sorts of issues that those interested in the philosophy of the trolley problem have been concerned with. For example, consider this question: when different people under immediate threat are all equally innocent, and a bystander could save some of them but not others, but this would involve killing some of those people, what ethical considerations should that bystander take into account? Many self-driving car cases – such as those in the moral machine vignettes – are not about that issue. But it is a central question in the trolley problem literature (Kamm 2015; Thomson 2015). Accordingly, just as we might not learn anything about the ethics of self-driving cars by considering some parts of the literature about the trolley problem, it might also be that we do not learn anything about the key issues engaging those interested in the trolley

problem by considering the sorts of cases that are sometimes compared with the trolley problem in a fast and loose way.

Not everyone agrees that the trolley problem literature has primarily focused on issues that are unhelpful for the ethics of crashes with self-driving cars, however. I have already mentioned Keeling above, but there are others as well. A paper by Dietmar Hübner and Lucie White (2018) makes a strong case in favor of the idea that the trolley problem literature – in particular the early papers by Philippa Foot and Judith Jarvis Thomson – contains important moral distinctions that matter to the issue of how self-driving cars should handle accident scenarios. Hübner and White think that the classic trolley problem discussions about the difference between negative rights and positive rights (Foot 1967) and differences between people's moral claims (Thomson 1976) are useful when we think about how self-driving cars should respond to situations involving unavoidable crashes. Specifically, they argue that various suggestions in the early trolley literature about how to draw the ethical difference between "involved" and "uninvolved" parties are highly relevant to the real-world ethics of self-driving cars (Hübner & White 2018).[9] In short, whether the literature on the trolley problem – be it the early contributions to it or more recent ones – is relevant to the ethics of how people should behave around self-driving cars and how self-driving cars should be made to behave around people is a matter of philosophical dispute.

[9] Hübner and White think that going back to the early trolley literature is a way of "clearly transcending the restricted horizon of purely utilitarian optimization, and providing important frameworks for taking people's individual responsibilities and mutual obligations into account" (Hübner, personal correspondence). There is a striking difference here between that view and JafariNaimi's above-cited view that trolley problem-reasoning inevitably leads to a "utilitarian framing" of ethical reflection. Foot and Thomson, it can be noted, used the trolley problem examples to illustrate what they regard as crucial deontological distinctions; and Kamm (2015) also uses the trolley problem in her defense of a "nonconsequentialist" view of ethics.

5: What Should We Make of All of This?

During an auto-show in Paris in 2016, a representative of Mercedes, named Christoph von Hugo, was interviewed about the company's self-driving car prototype that was being showcased at the event. When asked about how their self-driving cars would be programmed to respond to accident scenarios, Mr. von Hugo answered that Mercedes' cars would always prioritize their owners (Taylor, 2016). He even presented some off-the-cuff arguments for why this would be a good policy (which prompted the headline quoted above about "*Why Mercedes Plans to Let Its Self-Driving Cars Kill Pedestrians in Dicey Situations"* [10]*.*

Given many people's above-discussed attitudes about the moral machine thought experiments suggesting that they would prefer buying a car that would be programmed to always save them, one might have predicted that this would go over well with people. However, there was an outcry. And von Hugo had to later retract his previous statements. He ended up claiming that his previous statements—which included his arguments for why it would be a good idea to always prioritize the owner of the car—were taken out of context. Mercedes had certainly not made up their minds to program their cars to always prioritize their owners (Orlove 2016).

[10] For example, von Hugo said "Save the one in the car. If all you know for sure is that one death can be prevented, then that's our first priority" (Taylor 2016). For more on this Mercedes controversy and the issue of whether to always put the passenger first, see Katherine Evans's interesting discussion in her contribution to Keeling et al. 2019.

The person who interviewed von Hugo had clearly heard about ethical discussions inspired by the trolley problem about how self-driving cars should be programmed to handle crash scenarios. And Mr. von Hugo seemed to also have heard about this – moreover, he also seemed to potentially have heard about the empirical finding that most people would want to buy or use a car programmed to always save them. Presumably he may even have thought that his answers would appeal to potential buyers and users of self-driving Mercedes cars, but seems to have not predicted the reactions from others, who might not be comfortable with the idea of self-driving Mercedes cars that would drive around and do everything to save their passengers if any crash scenario should arise where different people's lives would be at stake. Safer in the end, then, to take everything back and assure the general public that the company had not made up its mind and that they would leave it to others – e.g., regulators or other public officials – to make decisions about these things.

Given all the above-discussed disagreements about whether comparing the ethics of crashes with self-driving cars with the trolley problem is a good idea, one might think that a similar conclusion would be what would make most sense with respect to the three philosophical disputes discussed above as well. In other words, one might think that ethical issues about how self-driving cars should behave in risky situations should not be discussed and argued about by philosophers and other academics interested in the trolley problem. In fact, this is the suggestion that Himmelreich ends up making in his above-mentioned article criticizing the self-driving cars/trolley problem comparison. Himmelreich suggests that people will have so many disagreements that the best thing to do is to treat the issue of how self-driving cars should handle risky situations as a

"social choice" issue that should have some sort of political solution.[11] It should not be seen as an ethical problem at all, but a political one (see also Rodríguez-Alcázar et al. 2020). Filippo Santoni de Sio (2017), in turn, suggests something that has some similarity to Himmelreich's approach. Rather than basing reasoning about how self-driving cars should handle accident scenarios on ethical theorizing, it might be better, Santoni de Sio thinks, to turn to legal arguments. In particular, the suggestion is to consider legal reasoning related to emergency situations and specifically the so-called "doctrine of necessity" that is found in Anglo-American jurisprudence and elsewhere. This has something in common with Harris's (2020) suggestion that rather than ordinary people's gut reactions to the sorts of dilemma scenarios the MIT moral machine experimenters have confronted people with, it is better to base reasoning about the ethics of self-driving cars on precedents from the legal context.

Should we follow the lead of these writers, and conclude that it is best to not make comparing the ethics of self-driving cars to the trolley problem part of the tool box we use for thinking about real-world ethical issues concerning self-driving cars and risky traffic situations? When we think about this issue, it is useful to distinguish among three different methodological approaches we could take. The first would be to approach the ethics of self-driving cars by only considering cases similar to those associated with the trolley problem (while

[11] When Himmelreich (2018) suggests that we should take a "social choice" approach, he uses that phrase in a slightly looser sense than it is sometimes otherwise used. Standardly, social choice theory is understood along fairly narrow lines, namely, as the sub-discipline of economics that looks at aggregating individual judgements to determine a collective judgement. But in his article, Himmelreich has in mind a more deliberative democratic approach, like that of John Rawls (1993) in *Political Liberalism*, which aims for an "over-lapping consensus".

perhaps also consulting the literature about the trolley problem) and doing nothing else than this. A second approach would be to do what I just described as the first approach, but to also do other things when we think about the ethics of self-driving cars – e.g. make use of arguments inspired by legal reasoning like Santoni de Sio suggests, or any other type of ethical methodology we might find helpful in this context. This second approach would be what one might call a methodological pluralism approach. The third approach would be to not at all do anything associated with the first approach, and to only use methods wholly divorced from anything resembling the trolley problem when thinking about the ethics of dangerous situations involving self-driving cars.

It seems to me that the sections above have reviewed enough critical arguments raising skeptical worries about comparing the ethics of self-driving cars with the trolley problem that what I call the first approach in the paragraph above does not seem like a very satisfying approach. Clearly, there is potentially something morally suspect about drawing a very close analogy between crashing self-driving cars and the philosophy of the trolley problem (see the first dispute above). There are clearly also important disanalogies between real-life crashes involving self-driving cars and the examples associated with the trolley problem (see the second dispute above). And, lastly, there is clearly a question of whether the literature about the trolley problem has consistently been about issues of crucial importance for the real-world ethics of self-driving cars (see the third dispute). But that there are reasons – encapsulated in these three philosophical disputes – for shying away from ethical theorizing about self-driving cars that is primarily or exclusively about trolley problem-like cases does not mean that the

self-driving cars/trolley problem comparison has no value or that we should not pay any attention to it.

After all, by considering reasons for being skeptical about drawing a close analogy between the ethics of self-driving cars and the philosophy of the trolley problem, we are in effect creating an account of what issues are most important for the real-world ethics of crashes involving self-driving cars. In other words, comparing the ethics of self-driving cars with the trolley problem is at the very least indirectly important. It helps us to highlight what is and what is not important for the ethics of self-driving cars. And, furthermore, while many philosophers who have written about the self-driving cars/trolley problem comparison have been highly skeptical, there are also those who see great value in this comparison, such as Keeling, Hübner and White, and others. They have presented interesting and important arguments for making this comparison. If they are right – and surely some of their arguments are sound – then the self-driving cars/trolley problem comparison is also directly useful to the real-world ethics of self-driving cars.

Accordingly, it seems to me that just as the first methodological approach mentioned a few paragraphs above is problematic, so is the third methodological approach. In other words, we do best to take the second approach. We should neither rely too heavily (or indeed exclusively) on the comparison between the ethics of self-driving cars and the trolley problem, nor wholly ignore and pay no attention to the comparison between the ethics of self-driving cars and the trolley problem. Rather, we do best to make this one – but not the only – thing we do when we think about the ethics of self-driving cars. With what is still a relatively new issue for philosophical ethics to work with, and indeed also regarding older

ethical issues that have been around much longer, using a mixed and pluralistic method that approaches the moral issues we are considering from many different angles is surely the best way to go. In this instance, that includes reflecting on – and reflecting critically on – how the ethics of crashes involving self-driving cars is both similar to and different from the philosophy of the trolley problem.

At this point, somebody might say, "what if I am somebody who really dislikes the self-driving cars/trolley problem comparison, and I would really prefer reflecting on the ethics of self-driving cars without spending any time on thinking about the similarities and differences between the ethics of self-driving cars and the trolley problem?" In other words, should everyone working on the ethics of self-driving cars spend at least some of their time reflecting on the comparison with the trolley problem? Luckily for those who are reluctant to spend any of their time reflecting on the self-driving cars/trolley problem comparison, there are others who are willing and able to devote at least some of their energies to this comparison.

In general, I think we should view the community that works on the ethics of this issue as being one in which there can be a division of labor, whereby different members of this field can partly focus on different things, and thereby together cover all of the different aspects that are relevant and important to investigate regarding the ethics of self-driving cars. As it happens, there has been a remarkable variety in the methods and approaches people have used to address the ethics of self-driving cars (see Nyholm 2018 a-b). So, while it is my own view that anybody who wants to form a complete overview of the ethics of self-driving cars should, among other things, devote some of their time to studying the comparison with the trolley problem, it is ultimately no big problem if not

everyone wishes to do so. There are others who have been studying, and who will most likely continue to reflect on, this comparison.[12]

References:

Awad, Edward, Dsouza, Sohan, Kim, Richard, Shulz, Jonathan, Henrich, Joseph, Shariff, Azim, Bonnefon, Jean-Francois, Rahwan, Iyad (2018): 'The Moral Machine Experiment', *Nature* 563**:** 59–64

Bonnefon J-F, Shariff A, Rahwan I (2015) Autonomous vehicles need experimental ethics: are we ready for utilitarian cars? arXiv:1510.03346 [cs]. Retrieved from http://arxiv.org/abs/1510.03346

Bonnefon, J.-F., Sharrif, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science, 352(6293), 1573–1576.

Business Insider 2016: "Why Mercedes Plans to Let Its Self-Driving Cars Kill Pedestrians in Dicey Situations": https://www.businessinsider.nl/mercedes-benz-self-driving-cars-programmed-save-driver-2016-10/

Cunneen, M., Mullins, M., Murphy, F., Shannon, D. Furxhi, I., & Ryan, C. (2019): "Autonomous Vehicles and Avoiding the Trolley (Dilemma): Vehicle Perception, Classification, and the Challenges of Framing Decision Ethics", *Cybernetics and Systems* 51(1): 59-80.

Davnall, R. (2019): "Solving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics", *Science and Engineering Ethics* 26(1): 431-449

Edmonds, D. (2013). *Would you kill the fat man?* Princeton: Princeton University Press

Foot, P. (1967): The problem of abortion and the doctrine of double effect. The Oxford Review 5:

Frowe, H. (2015): *Defensive Killing*. Oxford: Oxford University Press.

Gogoll J, Müller JF (2017) Autonomous Cars: In Favor of a Mandatory Ethics Setting. Sci Eng Ethics 23:681-700

Goodall, N. (2016): "Away from Trolley Problems and Toward Risk Management", *Applied Artificial Intelligence* 30(8): 810-821

[12] For helpful feedback on this chapter, I am thankful to Geoff Keeling, Lucie White, Dietmar Hübner, and the participants of Fleur Jongepier and my "Moral Theory and Real Life" PhD course. My work on this chapter is part of the research program "Ethics of Socially Disruptive Technologies", which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York: Penguin.

Gibbard, A. (2008). *Reconciling Our Aims.* Oxford: Oxford University Press.

Hansson, S.O. (2012): "A Panorama of the Philosophy of Risk" in S. Roeser, R. Hillebrand, & M. Peterson (eds.), *Handbook of Risk Theory*, Dordrecht: Springer: 27-54

Harris, J. (2020): "The Immoral Machine", *Cambridge Quarterly of Healthcare Ethics* 29(1): 71-79

Himmelreich, J. (2018): "Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations", *Science and Engineering Ethics* 21: 669-684

Hübner, D. & White, L. (2018): "Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us Beyond Harm Minimization", *Ethical Theory and Moral Practice* 21: 685-698

JafariNaimi, N (2017): "Our Bodies in the Trolley's Path, or Why Self-Driving Cars Must *Not* Be Programmed to Kill", *Science, Technology, & Human Values* 43(2): 302-323

Kamm, F. (1996): *Morality, Mortality*. Oxford: Oxford University Press.

Kamm, F. (2015). *The Trolley Problem Mysteries*. Oxford: Oxford University Press.

Kamm, F. (2020): "Use and Abuse of the Trolley Problem: Self-Driving Cars, Medical Treatments, and the Distribution of Harm" in S.M. Liao (ed.) *Ethics of Artificial Intelligence*, Oxford: Oxford University Press: 79-108

Kauppinen, A. (2020): "Who Should Bear the Risk When Self-Driving Vehicles Crash?", *Journal of Applied Philosophy*, https://doi.org/10.1111/japp.12490

Keeling, G. (2019). Why Trolley Problems Matter for the Ethics of Automated Vehicles", *Science and Engineering Ethics* 26(1): 293-307

Keeling, G., Evans, K., Thornton, S., Mecacci, G., & Santoni de Sio, F. (2019): "Four Perspectives on What Matters for the Ethics of Automated Vehicles" in G. Meyer & S. Beiker (eds.) *Road Vehicle Automation 6*, Berlin: Springer: 49-60

LeBeau, P. (2016), "Google's Self-Driving Car Caused an Accident, So What Now?," CNBC, https://www.cnbc.com/2016/02/29/googles-self-driving- car-caused-an-accident-so-what-now.html

Levin, S. and J.C. Wong (2018), "Self-Driving Uber Kills Ari- zona Woman in First Fatal Crash involving Pedestrian," *The Guardian*, https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe

Lin, P. (2015): "Why ethics matters for autonomous cars". In: Maurer M, Gerdes J, Lenz B, Winner H (eds) Autonomous driving: technical, legal and social aspects. Springer, Berlin: 69–85

McMahan, J. (2005). "The Basis of Moral Liability for Defensive Killing", Philosophical Issues 15: 386-405

McMahan, J. (2009): *Killing in War*. Oxford: Oxford University Press

Nyholm, S. & J. Smids (2016): "The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?" *Ethical Theory and Moral Practice* 19: 1275-1289

Nyholm, S. (2018a): "The Ethics of Crashes with Self-Driving Cars: A Roadmap, I," Philosophy Compass 13(7), e12507

Nyholm. S. (2018b): "The Ethics of Crashes with Self-Driving Cars, A Roadmap, II," Philosophy Compass 13(7), e12506

Nyholm, S. (2020): *Humans and Robots: Ethics, Agency, and Anthropomorphism*, London: Rowman & Littlefield International

Orlove, R. (2016): "Now Mercedes Says Its Driverless Cars Won't Run Over Pedestrians, That Would be Illegal", *Jalopnik*: https://jalopnik.com/now-mercedes-says-its-driverless-cars-wont-run-over-ped-1787890432

Purves D, Jenkins R, Strawser BJ (2015) Autonomous machines, moral judgment, and acting for the right reasons. Ethical Theory Moral Pract 18:851–872

Rawls, J. (1993): *Political Liberalism*. New York: Columbia University Press.

Robotics Business Review (2019): "Trolley Dilemmas Shouldn't Influence Self-Driving Policies, Experts Argue" https://www.roboticsbusinessreview.com/unmanned/trolley-dilemmas-should-not-formulate-self-driving-policies/

Rodríguez-Alcázar, Javier, Lilian Bermejo-Luque, and Alberto Molina-Pérez. "Do Automated Vehicles Face Moral Dilemmas? A Plea for a Political Approach." *Philosophy & Technology* (2020): 1-22: DOI: 10.1007/s13347-020-00432-5

Santoni de Sio, F. (2017): "Killing by autonomous vehicles and the legal doctrine of necessity". *Ethical Theory and Moral Practice*, 20(2), 411– 429.

Schoettle, B., & Sivak, M. (2015). A preliminary analysis of real-world crashes involving self-driving vehicles (No. UMTRI-2015-34). Ann Arbor, MI: The University of Michigan Transportation Research Institute

Sütfeld, L.R., et al. "Using virtual reality to assess ethical decisions in road traffic scenarios: applicability of value-of-life-based models and influences of time pressure." *Frontiers in Behavioral Neuroscience* 11 (2017): 122

Taylor, M. (2016): "Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians", *Car and Driver*: https://www.caranddriver.com/news/a15344706/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/

Technology Review (2018): "Should a Self-Driving Car Kill the Baby or the Grandma? Depends on Where You're From": https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/

The Tesla Team (2016), "A Tragic Loss," Tesla Blog, https://www.tesla.com/blog/tragic-loss

Thomson, Judith Jarvis (1976): "Killing, Letting Die, and the Trolley Problem", *The Monist* 59: 204-217

Thomson, Judith (2015): "Kamm on the Trolley Problems" in F. Kamm *The Trolley Problem Mysteries*. Oxford: Oxford University Press: 113-134.

Venture Beat (2018): "MIT Study Explores the 'Trolley Problem' and Self-Driving Cars", https://venturebeat.com/2018/10/24/mit-study-explores-the-trolley-problem-and-self-driving-cars/

Wakabayashi, D. and K. Conger (2018), "Uber's Self-Driving Cars Are Set to Return in a Downsized Test," New York Times, https://www.nytimes.com/2018/12/05/technology/uber-self-driving-cars.html

Wallach, W. & C. Allen (2009): *Moral Machines: Teaching Robots Right from Wrong*, Oxford: Oxfoed University Press

Washington Post (2015): *Driverless Cars Are Colliding with the Creepy Trolley Problem,* https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/

Washington Post (2016): "Google's Chief of Self-Driving Cars Downplays 'The Trolley Problem'": https://www.washingtonpost.com/news/innovations/wp/2015/12/01/googles-leader-on-self-driving-cars-downplays-the-trolley-problem/

Wolkenstein, A. (2018): "What has the Trolley Dilemma Ever Done for Us (And What Will It Do in the Future)? On Some Recent Debates about the Ethics of Self-Driving Cars", *Ethics and Information Technology* 20: 163-173

Wood, A. (2011): "Treating Humanity as an End in Itself" in D. Parfit *On What Matters, Vol. 2*, Oxford: Oxford University Press