Ethic Theory Moral Prac (2016) 19:1275–1289
DOI 10.1007/s10677-016-9745-2

The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?

Sven Nyholm[1] · **Jilles Smids**[1]

Accepted: 6 July 2016 / Published online: 28 July 2016

Abstract Self-driving cars hold out the promise of being safer than manually driven cars. Yet they cannot be a 100 % safe. Collisions are sometimes unavoidable. So self-driving cars need to be programmed for how they should respond to scenarios where collisions are highly likely or unavoidable. The accident-scenarios self-driving cars might face have recently been likened to the key examples and dilemmas associated with the trolley problem. In this article, we critically examine this tempting analogy. We identify three important ways in which the ethics of accident-algorithms for self-driving cars and the philosophy of the trolley problem differ from each other. These concern: (i) the basic decision-making situation faced by those who decide how self-driving cars should be programmed to deal with accidents; (ii) moral and legal responsibility; and (iii) decision-making in the face of risks and uncertainty. In discussing these three areas of disanalogy, we isolate and identify a number of basic issues and complexities that arise within the ethics of the programming of self-driving cars.

Keywords Self-driving cars · The trolley problem · Decision-making · Moral and legal responsibility · Risks and uncertainty

Self-driving cars hold out the promise of being much safer than our current manually driven cars. This is one of the reasons why many are excited about the development and introduction of self-driving cars. Yet, self-driving cars cannot be a 100 % safe. This is because they will drive with high speed in the midst of unpredictable pedestrians, bicyclists, and human drivers (Goodall 2014a-b) So there is a need to think about how they should be programmed to react in different scenarios in which accidents are highly likely or unavoidable. This raises important ethical questions. For instance, should autonomous vehicles be programmed to always minimize the number of deaths? Or should they perhaps be programmed to save their passengers at all costs? What moral principles should serve as the basis for these "accident-algorithms"? Philosophers are

✉ Sven Nyholm
s.r.nyholm@tue.nl

[1] Philosophy and Ethics, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

slowly but surely beginning to think about this general issue, and it is already being discussed in the media and in various different online forums.

Some philosophers have recently likened accident-management in autonomous vehicles to the so-called trolley problem. Several journalists and opinion piece writers have also done so.[1] The trolley problem is the much-discussed set of philosophical thought experiments in which there is a runaway trolley and the only way to save five people on the tracks is to sacrifice one person. (Thomson 1985) Different versions of these trolley cases vary with respect to how the one will need to be sacrificed in order for the five to be saved. It is the most basic versions that are said to foreshadow the topic of how to program autonomous vehicles.

For example, Patrick Lin writes:

> 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 (viz. driverless trains), 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)

Nor are philosophers alone in making this comparison. Economists and psychologists Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan write:

> 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)

According to these various writers, then, the problem of how to program self-driving cars and other autonomous vehicles for different accident-scenarios is very similar to the trolley problem. If true, this would have important implications concerning how to best approach the ethics of self-driving cars. It suggests that when we approach the ethics of accident-algorithms for autonomous vehicles, the ever growing literature on the trolley problem is a good, if not the best, place to start. Moreover, it suggests that that literature treats the key issues we need to focus on when we try to formulate an ethical framework for sound moral reasoning about how autonomous vehicles should be programmed to deal with risky situations and unavoidable accidents.[2]

In this paper, we critically examine this tempting analogy between the trolley problem and the problem of accident-algorithms for self-driving cars. We do so with a skeptical eye. Specifically, we argue that there are three very important respects in which these two topics are not analogous. We think, therefore, that it is important to resist the temptation to draw a very strong analogy between the ethics of accident-algorithms for self-driving cars and the philosophy of the trolley problem.

[1] E.g. Achenbach 2015; Doctorow 2015; Lin 2013; Windsor 2015; Worstall 2014

[2] Reasoning in just that way, Bonnefon et al. (2015) propose that the methods developed within experimental ethics to investigate judgments about the trolley problem should be used to investigate ordinary people's intuitions about accident-algorithms for self-driving cars. Bonnefon et al. further think that once these intuitions have been carefully surveyed and systematically analyzed, they should then serve as starting points for normative discussions of how self-driving cars ought to be programmed.

Why is this an important topic to investigate? Firstly, the issue of how to program self-driving cars is a pressing ethical issue, given the rapid development of this technology and the serious risks involved. We therefore need to identify the best sources of ethical theory that could help us to deal with this part of moral practice. At this stage, we are only beginning to grapple with this problem. So it is crucial to thoroughly investigate any initial "leads" we have about where best to start as we open up the discussion of this general topic. The similarity between accident-planning for self-driving cars and the trolley problem that some writers claim to have identified is one such lead. That's one reason why it is important to investigate whether (or not) the literature on the trolley problem is indeed the best place to primarily turn to as we approach this ethical issue. Secondly, in investigating how similar or dissimilar these two topics are, we in effect isolate and identify a number of basic key issues that further work on the ethics of accident-algorithms for self-driving cars needs to deal with. In conducting this positive part of our inquiry, we investigate what types or categories of considerations are at issue here. And we make it very clear that the problem of how to program autonomous vehicles to respond to accident-scenarios is a highly complex ethical issue, under which there are various sub-issues that on their own also exhibit a lot of complexity.

We proceed as follows. We first say a little more about why an ethical framework for risk-management for autonomous vehicles needs to be developed (section 1). We then say more about the trolley problem and the main issues discussed in the literature on the trolley problem (section 2). After that, we explain the three main differences we see between the ethics of accident-management in self-driving cars, on the one hand, and the trolley problem and trolleyology, on the other hand (sections 3–5). To anticipate, these differences have to do with (i) prospective planning by groups that takes large numbers of situational features into account vs. imagined split-second decisions by individuals that abstract away all but a few features of the situation directly at hand; (ii) taking seriously pressing issues of moral and legal responsibility vs. setting such issues aside as irrelevant as a matter of stipulation; and (iii) reasoning about probabilities, uncertainties and risk-management vs. intuitive judgments about what are stipulated to be known and fully certain facts. Lastly, we end by briefly summarizing our main conclusions (section 6).

1 Programming Self-Driving Cars for how to React in the Event of Accidents

As we noted above, even self-driving cars will inevitably sometimes crash. Noah Goodall (2014a-b) and Patrick Lin (2015) convincingly argue for this claim, and in addition explain why programming self-driving cars for crashes involves ethical choices. In explaining what's at issue, we here draw on Goodall's and Lin's work.

A self-driving car uses advanced sensor-technology to detect its surroundings and sophisticated algorithms to subsequently predict the trajectory of nearby (moving) objects. Self-driving cars can also use information technology to communicate with each other, thus achieving better coordination among different vehicles on the road. However, since cars are heavy and move with high speed, physics informs us that they have limited maneuverability, and that they often cannot simply stop. Therefore, even if the car-to-car communication and the sensors and algorithms are all functioning properly (and would be better than current technology), self-driving cars will not always have sufficient time to avoid collisions with objects that suddenly change direction. (Goodall 2014a) Self-driving cars will sometimes collide with each other. But there are also other moving objects to worry about. Pedestrians,

cyclists, and wildlife naturally come to mind here. (Lin 2015) However, we must also take into account human-driven cars. Because, as is generally acknowledged by the experts, self-driving cars will for a long period drive alongside human-driven cars (so-called "mixed traffic") (van Loon and Martens 2015).

For these reasons, automated vehicles need to be programmed for how to respond to situations where a collision is unavoidable; they need, as we might put it, to be programmed for how to crash. At first blush, it might seem like a good idea to always transfer control to the people in the car in any and all situations where accidents are likely or unavoidable. However, human reaction-times are slow. It takes a relatively long time for us to switch from focusing on one thing to focusing on another. So handing over control to the human passengers will often not be a good option for the autonomous vehicle. Hence the car itself needs to be prepared, viz. programmed, for how to handle crashes.

This has certain advantages. A self-driving car will not react in the panicky and disorganized ways a human being is apt to react to accident-scenarios. Even in situations of unavoidable collisions, the car's technology enables significant choices and control-levels regarding how to crash. Based on its sensor inputs and the other information it has access to, the car can calculate the most likely consequences of different trajectories that involve different combinations of braking and swerving.

Consider now the following scenario.[3] A self-driving car with five passengers approaches a conventional car (e.g. a heavy truck) that for some reason suddenly departs from its lane and heads directly towards the self-driving car. In a split-second, the self-driving car senses the trajectory and the likely weight of the oncoming truck. It calculates that a high-impact collision is inevitable, which would kill the five passengers, unless the car swerves towards the pavement on its right-hand side. There, unfortunately, an elderly pedestrian happens to be walking, and he will die as a result if the self-driving car swerves to the right and hits him. This is the sort of situation in which the human passengers of a self-driving car cannot take control quickly enough. So the car itself needs to respond to the situation at hand. And in order for the five passengers in the self-driving car to be saved, as they are likely to be if the head-on collision with the heavy truck is avoided, the car here needs to make a maneuver that will most likely kill one person.

It is evident that scenarios like this one involve significant ethical dilemmas. Among other things, they raise questions about what the self-driving car's pre-set priorities should be. Should it here swerve to the sidewalk and save the greatest number, or should it rather protect the innocent pedestrian and crash into the oncoming truck? In general, should the car be programmed to always prioritize the safety of its passengers, or should it sometimes instead prioritize other considerations, such as fairness or the overall good, impartially considered? Crucially, unless the self-driving car is programmed to respond in determinate ways to morally loaded situations like the one we just described, there is an unacceptable omission in its readiness to deal with the realities and contingencies of actual traffic. (Goodall 2014a-b; Lin 2015) Not programming the car for how to respond to situations like this and others like it

[3] Our illustration here is a "mixed traffic"-case. Self-driving cars will inevitably sometimes collide with each other, for example if one of them is malfunctioning. But the risks are even greater within mixed traffic involving both self-driving cars and conventional cars, since human drivers and self-driving cars have a harder time communicating with each other. (van Loon and Martens 2015) Still, self-driving cars need to be programmed for how to handle collisions with both other self-driving cars and conventional cars (in addition to any other objects that might suddenly appear in their paths). (Lin 2015) We discuss the ethics of compatibility-problems within mixed traffic at greater length in (Nyholm and Smids in progress).

amounts to knowingly relinquishing the important responsibility we have to try to control what happens in traffic. It amounts to unjustifiably ignoring the moral duty to try to make sure that things happen in good and justifiable ways. We should not do that. Hence the need for ethical accident-algorithms.[4]

At first glance, the accident-scenario we just described above looks similar to the examples most commonly discussed in relation to the trolley problem. But suppose that we probe a little deeper beneath the immediate surface, and that we home in on the more substantial ethical issues raised by the choice of accident-algorithms for self-driving cars. Are Lin, Wallach and Allen, and Bonnefon et al. then still right that the choice of accident-algorithms for self-driving cars is like a real world version of the trolley problem, as it is usually understood and discussed in the literature?

2 "The Trolley Problem"?

The easiest way to introduce the trolley problem is to start with the two most widely discussed cases involved in these thought experiments. These are the cases the above-cited writers have in mind when they compare the ethics of accident-algorithms for self-driving cars to the trolley problem.

In the "switch" case, a driverless trolley is heading towards five people who are stuck on the tracks and who will be killed unless the trolley is redirected to a side track. You are standing next to a switch. If you pull the switch, the trolley is redirected to a side-track and diverted away from the five. The trouble is that on this side track, there is another person, and this person will be killed if you pull the switch to redirect the train. Nevertheless, a very common response to this case is that it is here permissible for you to save the five by redirecting the train, thus killing the one as a result (Greene 2013).

In a different variation on this theme, the "footbridge" case, saving the five requires different means. In this case, you are located on a footbridge over the tracks. Also present on the footbridge is a very large and heavy man. His body mass is substantial enough that it would stop the trolley if he were pushed off the footbridge and onto the tracks. But this would kill him. Is it morally permissible to push this man to his death, thereby saving the five by this means? A very common response to this case is that it is not permissible. (Greene 2013) So in this case saving the five by sacrificing the one seems wrong to most of us, whereas in the other case, saving the five by sacrificing the one seems morally permissible.

Many people casually use the phrase "the trolley problem" to refer to one or both of these examples. But some influential philosophers use this phrase to mean something more distinct. According to Judith Jarvis Thomson, for example, the basic trolley problem is to explain the above-described asymmetry in our judgments. (Thomson 2008) That is, why is it permissible in one case to save the five by sacrificing the one, whereas it is not permissible to save the five by sacrificing the one in the other case? Others favor a wider interpretation of the trolley problem, holding that this problem also arises in cases that don't involve any trolleys at all.

[4] The design of ethical decision-making software immediately presents two major challenges. First, what moral principles should be employed to solve this sort of ethical dilemmas? Second, even if we were to reach agreement, it turns out to be a formidable challenge to design a car capable of acting fully autonomously on the basis of these moral principles (cf. Goodall 2014a). We will not go into this latter question, but will instead here simply note that this is an important and pressing issue, which is studied in the field of machine morality or robot ethics (e.g. Wallach and Allen 2009).

According to Frances Kamm, 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? (Kamm 2015) For example, why is it is not permissible for a medical doctor to save five patients in need to organ-transplants by "harvesting" five organs from a perfectly healthy patient who just came into the hospital for a routine check-up? This case doesn't mention trolleys, but Kamm thinks it nevertheless falls under the wide umbrella of the trolley problem.

These various thought-experiments have been used to investigate a number of different normative issues. For example, they have been used to investigate the difference between: (i) "positive" and "negative" duties, that is, duties to do certain things vs. duties to abstain from certain things; (ii) killing and letting die; and (iii) consequentialism and non-consequentialism in moral theory, that is, the difference between moral theories only concerned with promoting the overall good vs. moral theories that also take other kinds of considerations into account.(Foot 1967; Thomson 1985; Kamm 2015) And in recent years, they have also been used to empirically investigate the psychology and neuroscience of different types of moral judgments. (Greene 2013; Mikhail 2013).

We agree that trolley cases can certainly be useful in the discussion of these various topics. But how helpful are these thought-experiments and the large literature based on them for the topic of how self-driving cars ought to be programmed to respond to accident-scenarios where dangerous collisions are highly likely or unavoidable? We will now argue that there are three crucial areas of disanalogy that should lead us to resist the temptation to draw a strong analogy between the trolley problem and the ethics of accident-algorithms for self-driving cars.

3 Two Very Different Decision-Making Situations

To explain the first noteworthy difference between the ethics of accident-algorithms for autonomous vehicles and the trolley dilemmas we wish to bring attention to, we will start by returning to the quote from Wallach and Allen in the introduction above. Specifically, we would like to zoom in on the following part of that quote:

> Driverless systems put machines in the position of making split-second decisions that could have life or death implications.

It is tempting to put things like this if one wishes to quickly explain the ethical issues involved in having driverless systems in traffic.[5] But it is also somewhat misleading. Wallach and Allen are surely right that there is some sense in which the driverless systems themselves need to make "split-second decisions" when they are in traffic. And these can indeed have life or death implications. However, strictly speaking, the morally most important decision-making is made at an earlier stage. It is made at the planning stage when it is decided how the autonomous vehicles are going to be programmed to respond to accident-scenarios. The "decisions" made by the self-driving cars implement these earlier decisions.

[5] In a similar way, Lin writes that if "motor vehicles are to be truly autonomous and be able to operate responsibly on our roads, they will need to replicate [. . .] the human decision-making process." (Lin 2015, 69). Cf. also Purves et al.'s remark that "[d]riverless cars [. . .] would likely be required to make life and death decisions in the course of operation." (Purves et al. 2015, 855)

The morally relevant decisions are prospective decisions, or contingency-planning, on the part of human beings. In contrast, in the trolley cases, a person is imagined to be in the situation as it is happening. The person is forced right there and then to decide on the spot what to do: to turn the trolley by pulling the switch (switch case) or push the large man off the bridge (footbridge case). This is split-second decision-making. It is unlike the prospective decision-making, or contingency-planning, we need to engage in when we think about how autonomous cars should be programmed to respond to different types of scenarios we think may arise.[6] When it comes to the morally relevant decision-making situations, there is more similarity between accident-situations involving conventional cars and trolley-situations than between prospective programming for accident-situations involving autonomous cars and trolley-situations. For example, a driver of a conventional car might suddenly face a situation where she needs to decide, right there and then, whether to swerve into one person in order to avoid driving into five people. That is much closer to a trolley-situation than the situation faced by those who are creating contingency-plans and accident-algorithms for self-driving cars is.[7]

Nor is it plausible to think of decision-making about how self-driving cars should be programmed as being made by any single human being. That is what we imagine when we consider the predicament of somebody facing a trolley situation. This does not carry over to the case of self-driving cars. Rather, the decision-making about self-driving cars is more realistically represented as being made by multiple stakeholders – for example, ordinary citizens, lawyers, ethicists, engineers, risk-assessment experts, car-manufacturers, etc. These stakeholders need to negotiate a mutually agreed-upon solution. And the agreed-upon solution needs to be reached in light of various different interests and values that the different stakeholders want to bring to bear on the decision.[8]

The situation faced by the person in the trolley case almost has the character of being made behind a "veil of ignorance," in John Rawls' terms. (Rawls 1971) There is only a very limited number of considerations that are allowed to be taken into account. The decision-maker is permitted to know that there are five people on the tracks, and that the only way to save them is to sacrifice one other person – either by redirecting the runaway trolley towards the one (switch case) or by pushing a large person into the path of the trolley (footbridge case). These are the only situational factors that are allowed into the decision-making, as if this were a trial where the jury is only allowed to take into account an extremely limited amount of evidence in their deliberations.

This is not the ethical decision-making situation that is faced by the multiple stakeholders who together need to decide how to program self-driving cars to respond to different types of accident-scenarios. They are not in a position where it makes sense to set aside most situational

[6] As an anonymous reviewer pointed out, this difference in time-perspectives might render it plausible for different moral principles to serve as the evaluation-criteria for the programming of self-driving cars and the behavior of drivers in acute situations. For example, the aim of minimizing the statistically expected number of deaths can seem more justifiable and apt in prospective decision-making about accident-algorithms for self-driving cars than in retrospective evaluation of actual human responses to dramatic accident-scenarios. (Cf. Hansson 2013, 74–80)

[7] We owe this last observation to our colleague Auke Pols.

[8] Jason Millar argues that the accident-algorithms of self-driving cars ought to be selected by the owner of the car. (Millar 2014) This would mean that different cars could have different accident-algorithms. Two comments: firstly, this would still require a mutual decision since the basic decision to give owners of self-driving cars the right to choose their accident-algorithms would need to be agreed upon by the various different stakeholders involved. Second, this seems undesirable since different accident-algorithms in different cars would complicate coordination and compromise safety.

and contextual factors, and only focus on a small set of features of the immediate situation. Instead, they can bring any and all considerations they are able to think of as being morally relevant to bear on their decisions about how to program the cars. They can do that and should do so.

In sum, the basic features of these two different decision-making situations are radically different. In one case, the morally relevant decision-making is made by multiple stakeholders, who are making a prospective decision about how a certain kind of technology should be programmed to respond to situations it might encounter. And there are no limits on what considerations, or what numbers of considerations, might be brought to bear on this decision. In the other case, the morally relevant decision-making is done by a single agent who is responding to the immediate situation he or she is facing – and only a very limited number of considerations are taken into account. That these two decision-making situations are so radically different in their basic features in these respects is the first major disanalogy we wish to highlight.[9]

4 A Second Disanalogy: the Importance of Moral and Legal Responsibility

In order to set up the second main observation we wish to make, we will start by again returning to the following feature of the standard trolley cases. As just noted, we are asked to abstract away all other aspects of the situations at hand except for the stipulation that either five will die or one will die, where this depends on whether we (i) redirect the train away from the five and towards the one by pulling a switch (switch case) or (ii) push the one down from the bridge onto the tracks and into the line of the trolley (footbridge case). We are supposed to bracket any and all other considerations that might possibly be ethically relevant and consider what it would be permissible or right to do, taking only these features of the cases into consideration.

This is a characteristic of the trolley cases that has been criticized. Consider Allen Wood's criticism. Wood notes that, when we set aside everything else than the above-described considerations, there is "an important range of considerations that are, or should be, and in real life would be absolutely decisive in our moral thinking about these cases in the real world is systematically abstracted out." (Wood 2011, 70) Explaining what he finds problematic about this, Wood writes:

> even if some choices [in real life] 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. (Wood 2011, 73, emphasis added)

What is Wood getting at?

What Wood is missing is, among other things, due concern with moral and legal responsibility, viz. the question of who we can justifiably hold morally and legally responsible for

[9] Jan Gogoll and Julian Mueller identify three further differences between these two decision-making situations worth noting: (i) the much more static nature of standard trolley-situations as compared to the non-static situations that self-driving cars will typically face; (ii) the possibility of updating and revising accident-algorithms over time in self-driving cars, which contrasts with how trolley-situations are typically represented as isolated events; (iii) the one-sided nature of the "threat" in trolley-situations (the decision-maker is not represented as being at risk) as opposed to how in typical traffic-situations, all parties are usually subject to certain risks. (Gogoll and Müller 2016)

what is going on. Commenting specifically on how trains and trolley cars are regulated in real life, Wood writes:

> Trains and trolley cars are either the responsibility of public agencies or private companies that ought to be, and usually are, carefully regulated by the state with a view to ensuring public safety and avoiding loss of life. (Wood 2011, 74)

Developing the legal side of the issue further, Wood continues:

> ...mere bystanders ought to be, and usually are, physically prevented from getting at the switching points of a train or trolleys. They would be strictly forbidden by law from meddling with such equipment for any reason, and be held criminally responsible for any death or injury they cause through such meddling. (Wood 2011, 75)

Thus Wood thinks that trolley cases are too far removed from real life to be useful for moral philosophy. One key reason is that in real life, we hold each other responsible for what we do or fail to do. When it comes to things involving substantial risks – such as traffic – we cannot discuss the ethical issues involved without taking issues of moral and legal responsibilities into account. Since trolley cases ignore all such matters, Wood finds them irrelevant to the ethics of the real world.

We think that Wood might be going too far in making this criticism of the philosophical and psychological literature on the trolley problem. It is surely the case that sometimes examples that might not be very true to real life can serve useful purposes in moral philosophy and in various other fields of academic inquiry. But we think that the issue Wood brings up helps to highlight a stark difference between the discussion of the trolley problem and the issue of how we ought to program self-driving cars and other autonomous vehicles to respond to high-risk situations.

The point here is that when it comes to the real world issue of the introduction of self-driving cars into real world traffic, we cannot do what those who discuss the trolley problem do. We cannot stipulate away all considerations having to do with moral and legal responsibility. We must instead treat the question of how self-driving cars ought to be pre-programmed as partly being a matter of what people can be held morally and legally responsible for. (Cf. Hevelke and Nida-Rümelin 2014) Specifically, we must treat it as a question of what those who sell or use self-driving cars can be held responsible for and what the society that permits them on its roads must assume responsibility for.

With the occurrence of serious crashes and collisions – especially if they involve fatalities or serious injuries – people are disposed to want to find some person or persons who can be held responsible, both morally and legally. This is an unavoidable aspect of human interaction. It applies both to standard traffic, and traffic introducing self-driving cars. Suppose, for example, there is a collision between an autonomous car and a conventional car, and though nobody dies, people in both cars are seriously injured. This will surely not only be followed by legal proceedings. It will also naturally – and sensibly – lead to a debate about who is morally responsible for what occurred. If the parties involved are good and reasonable people, they themselves will wonder if what happened was "their fault". And so we need to reflect carefully on, and try to reach agreement about, *what* people can and cannot be held morally and legally responsible for when it comes to accidents involving self-driving cars. We also need to reflect on, and try to reach agreement about, *who* can be held responsible for the things that might happen and the harms and deaths that might occur in traffic involving these kinds of vehicles.

Questions concerning both "forward-looking" and "backward-looking" responsibility arise here. Forward-looking responsibility is the responsibility that people can have to try to shape

what happens in the near or distant future in certain ways. Backward-looking responsibility is the responsibility that people can have for what has happened in the past, either because of what they have done or what they have allowed to happen. (Van de Poel 2011) Applied to risk-management and the choice of accident-algorithms for self-driving cars, both kinds of responsibility are highly relevant. One set of important questions here concerns moral and legal responsibility for how cars that will be introduced into traffic are to be programmed to deal with the various different kinds of risky situations they might encounter in traffic. Another set of questions concerns who exactly should be held responsible, and for what exactly they should be held responsible, if and when accidents occur. The former set of questions are about forward-looking responsibility, the second about backward-looking responsibility. Both sets of questions are crucial elements of the ethics of self-driving cars.[10]

We will not delve into how to answer these difficult questions about moral and legal responsibility here. Our point in the present context is rather that these are pressing questions we cannot ignore, but must instead necessarily grapple with when it comes to the ethics of accident-algorithms for self-driving cars. Such questions concerning moral and legal responsibility are typically simply set aside in discussions of the trolley problem. For some of the theoretical purposes the trolley cases are meant to serve, it might be perfectly justifiable to do so. In contrast, it is not justifiable to set aside basic questions of moral and legal responsibility when we are dealing with accident-algorithms for self-driving cars. So we here have a second very important disanalogy between these two topics.

5 Stipulated Facts and Certainties vs. Risks, Probabilities, and Uncertainties

We now turn to the third and last major disanalogy we wish to highlight. Here, too, we will approach this disanalogy via a criticism that has been raised against the trolley problem and its relevance to the ethical issues we face in the real world. What we have in mind is Sven Ove Hansson's criticism of standard moral theory and what he regards as its inability to properly deal with the risks and uncertainties involved in many real world ethical issues. In one of his recent papers on this general topic, Hansson specifically brings up the trolley problem as one clear case in point of what he has in mind. Hansson writes:

> The exclusion of risk-taking from consideration in most of moral theory can be clearly seen from the deterministic assumptions commonly made in the standard type of life-or-death examples that are used to explore the implications of moral theories. In the famous trolley problem, you are assumed to know that if you flip the switch, then one person will be killed, whereas if you don't flip it, then five other persons will be killed. (Hansson 2012, 44).

What is Hansson's worry here? What is wrong with being asked to stipulate that we know the facts and that there is no uncertainty as regards what will happen in the different sequences of events we could initiate? Hansson comments on this aspect of the trolley cases in the following way:

[10] Current practice typically assigns backward-looking responsibility for accidents to drivers. But the introduction of self-driving cars is likely to shift backward-looking responsibility-attributions towards car-manufacturers. If justified, this would make backward- and forward-looking responsibility for accidents more closely related and coordinated. We owe these observations to an anonymous reviewer.

> This is in stark contrast to ethical quandaries in real life, where action problems with human lives at stake seldom come with certain knowledge of the consequences of the alternative courses of action. Instead, uncertainty about the consequences of one's actions is a major complicating factor in most real-life dilemmas. (Ibid.)

Hansson is not alone in making this criticism. Others have also worried that it is absurd to suppose, in any realistic situation, that doing something such as to push a large person in front of a trolley car would be sure to stop the trolley and save any people who might be on the tracks. As before, however, this may not be a fatal objection to the trolley cases if we conceive of them as a set of stylized thought experiments we use for certain circumscribed purely theoretical and abstract purposes. But again, we also see here that the trolley cases are far removed from the reality that we face when we turn to the ethical problem of how to program self-driving cars to respond to risky situations when we introduce these cars into actual traffic and thereby bring them into the real world with all its messiness and uncertainty.

We will illustrate this point by taking a closer look at our scenario from section I above. This was the scenario in which a heavy truck suddenly appears in the path of a self-driving car carrying five passengers, and in which the only way for the self-driving car to save the five appeared to be to swerve to the right, where it would kill an elderly pedestrian on the sidewalk. Under this brief description, the scenario might appear to involve an ethical dilemma in which we need to choose between outcomes whose features are known with certainty. But once we add more details, it becomes clear that there is bound to be a lot of uncertainty involved in a more fully described and maximally realistic version of the case.

First, the self-driving car cannot acquire certain knowledge about the truck's trajectory, its speed at the time of collision, and its actual weight. This creates uncertainty because each of these factors has a strong causal influence on the fatality risk for the passengers of the self-driving car. (Berends 2009; Evans 2001) Moreover, the truck-driver might try to prevent the accident by steering back to her lane. Or if it's already too late, she might start braking just half a second before the crash (thereby significantly reducing the truck's speed and impact). The self-driving car's software can only work with estimates of these alternative courses of events.[11]

Second, focusing on the self-driving car itself, in order to calculate the optimal trajectory, the self-driving car needs (among other things) to have perfect knowledge of the state of the road, since any slipperiness of the road limits its maximal deceleration. But even very good data from advanced sensors can only yield estimates of the road's exact condition. Moreover, regarding each of the five passengers: their chances of surviving the head-on collision with the truck depends on many factors, for example their age, whether they are wearing seat belts, whether they are drunk or not, and their overall state of health. (Evans 2008) The car's technology might enable it to gather partial, but by no means full, information about these issues.[12]

Finally, if we turn to the elderly pedestrian, again we can easily identify a number of sources of uncertainty. Using facial recognition software, the self-driving car can perhaps estimate his age with some degree of precision and confidence. (Goodall 2014a) But it may merely guess his actual state of health and overall physical robustness.[13] And whereas

[11] Furthermore, while the self-driving car may recognize the truck-type and know its empty mass, the truck may carry a load whose weight is unknown to the self-driving car.

[12] There is, of course, also the question of whether these kinds of facts about the passengers should count ethically here and if so, how exactly? Cf. Lin 2015

[13] It should be noted here that it is controversial whether we should assign any ethical weight to the fact that an elderly person might have a lower chance of surviving an accident than a younger, less fragile person might have. We are not taking a stand on that issue here.

statistical fatality rates for car-pedestrian collisions apply to a whole population, these might ultimately have fairly low predictive value for the elderly pedestrian's more precise chances of survival.[14] Of course, in real life, the scenario also involves the possibility that the pedestrian might avoid being hit by quickly stepping out of the self-driving car's path. The self-driving car necessarily has to work with an estimate of what the pedestrian is likely to do. And this estimation may need to be based on simulation-experiments rather than actual statistics.

As we start filling in these various further details, it quickly becomes clear that what we are dealing with here are not outcomes whose features are known with certainty. We are rather dealing with plenty of uncertainty and numerous more or less confident risk-assessments. (Cf. Goodall 2014b, 96) This means that we need to approach the ethics of self-driving cars using a type of moral reasoning we don't have occasion or reason to use in thinking about the standard cases discussed in the trolley problem literature.

In the former case, we need to engage in moral reasoning about risks and risk-management. We also need to engage in moral reasoning about decisions under uncertainty. In contrast, the moral reasoning that somebody facing a trolley case uses is not about risks and how to respond to different risks. Nor is it about how to make decisions in the face of uncertainty. This is a categorical difference between trolley-ethics and the ethics of accident-algorithms for self-driving cars. Reasoning about risks and uncertainty is categorically different from reasoning about known facts and certain outcomes. The key concepts used differ drastically in what inferences they warrant. And what we pick out using these concepts are things within different metaphysical categories, with differing modal status (e.g. risks of harm, on one side, versus actual harms, on the other).[15]

Thus the distinctive and difficult ethical questions that risks and uncertainty give rise to are not in play in the trolley cases. But they certainly are in the ethics of self-driving cars. Let us give just one illustration. A significant number of people may find hitting the pedestrian morally unacceptable if this was certain to kill him. (Cf. Thomson 2008) But what if the estimated chance of a fatal collision were 10 %? Or just 1 %? To many people, imposing a 1 % chance of death on an innocent pedestrian in order to save five car-passengers might appear to be the morally right choice. The trolley cases don't require any such judgments. In the scenarios involved in the trolley cases, all outcomes are assumed to be a 100 % certain, and hence there is no need to reflect on how to weigh different uncertain and/or risky outcomes against each other.[16]

Yet again, in other words, we find that the two different issues differ in striking and non-trivial ways. In one case, difficult questions concerning risks and uncertainty immediately arise, whereas in the other, no such issues are involved. This is another important disanalogy between the ethics of accident-algorithms for self-driving cars and the trolley problem. It is a disanalogy that exposes a categorical difference between these two different subjects.

[14] Research on pedestrian fatality rates is still in progress. (Rosén et al. 2011)

[15] Reasoning about risks and uncertainty is about what could happen even if it never does, whereas reasoning about known facts is about what is actually the case.

[16] When one is dealing with risks and uncertainty, one needs, among other things, to grapple with how to weigh uncertainties and risks against actual benefits. One needs to confront the difficult question of why imposing a risk onto somebody might be wrong, even if things go well in the end and certain kinds of actual harms end up not being realized. These and other difficult questions don't arise if, as is rarely the case, one knows exactly what will happen in different scenarios we might instigate. (Hayenhjelm and Wolff 2012) For some discussion of the acceptability of current driving risks, see (Smids 2015).

6 Concluding Discussion

We have isolated a number of important differences between the ethics of accident-algorithms for self-driving cars and the trolley problem. These all center around three main areas of disanalogy: with respect to the overall decision-situation and its features, with respect to the role of moral and legal responsibility, and with respect to the epistemic situation of the decision-makers. The various points we have made can be summarized and shown with the help of the following table. We here number the main areas of disanalogy as 1 through 3, and sub-divide 1 (viz. the disanalogous features of the basic decision-situations) into the three sub-disanalogies 1a-1c:

	Accident-algorithms for self-driving cars:	Trolley Problem:
1a: Decision faced by:	*Groups of individuals/ multiple stakeholders*	*One single individual*
1b: Time-perspective:	*Prospective decision/ contingency planning*	*Immediate/"here and now"*
1c: Numbers of considerations/ situational features that may be taken into account:	*Unlimited; unrestricted*	*Restricted to a small number of considerations; everything else bracketed*
2: Responsibility, moral and legal:	*Both need to be taken into account*	*Both set aside; not taken into account*
3: Modality of knowledge, or epistemic situation:	*A mix of risk-estimation and decision-making under uncertainty*	*Facts are stipulated to be both certain and known*

We started by asking just how similar, or dissimilar, the trolley problem and the issue of how self-driving cars ought to be programmed are. Return now briefly to the question of whether the literature on the trolley problem is a good, or perhaps the best, place to turn to for input for the ethics of accident-algorithms for self-driving cars. We can now argue as follows.

On the one hand, the key issues we have isolated as being 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 we have flagged as being most pressing for the ethics of accident-algorithms for self-driving cars. As we noted above, this 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. Taking these considerations together, we think it is clear 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.

Return next to the positive aim of the paper, namely, to isolate and identify key issues that the ethics of accident-algorithms for self-driving cars needs to deal with. Based on what we have argued in the previous sections – as summarized in the table above – we wish to draw the following broad conclusions about the general ethical issues that are raised by the question of

how to program self-driving cars to respond to accident-scenarios. What we are facing here are complex and difficult ethical issues relating to, among other things, the following:

(i) decision-making faced by groups and/or multiple stake-holders;
(ii) morally loaded prospective decision-making and/or contingency planning;
(iii) open-ended ethical reasoning taking wide ranges of considerations into account
(iv) ethical reasoning concerned with both backward-looking and forward-looking moral and legal responsibility
(v) ethical reasoning about risks and/or decisions under uncertainty.

We add the qualifier "among other things" here in order to make it clear that we are not of the opinion that these are the only general topics that are relevant for the ethics of accident-algorithms for self-driving cars. Rather, it is our view that these are among the general topics that are most relevant for this specific issue, but that there are certainly also other general topics that are highly relevant as we approach this ethical problem in a systematic way. Most importantly, we need to identify the ethical values, considerations, and principles that are best suited to be brought to bear on this pressing ethical issue. And we need to think about how to specify and adapt those values, considerations, and principles to the particular problem of how self-driving cars ought to be programmed to react to accident-scenarios. In other words, there is a lot of work to do here.[17]

References

Achenbach J (2015) Driverless cars are colliding with the creepy Trolley Problem. The Washington Post. . Retrieved from https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/

Berends EM (2009) De invloed van automassa op het letsel risico bij botsingen tussen twee personenauto's: een kwantitatieve analyse. [The impact of car-mass on injury risk from crashes between two person vehicles: a quantitative analysis]. SWOV

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

Doctorow C (2015) The problem with self-driving cars: who controls the code? The Guardian. Retrieved from http://www.theguardian.com/technology/2015/dec/23/the-problem-with-self-driving-cars-who-controls-the-code

Evans L (2001) Causal influence of car mass and size on driver fatality risk. Am J Public Health 91(7):1076–1081

Evans L (2008) Death in traffic: why are the ethical issues ignored? Stud Ethics Law Technol 2(1). doi:10.2202/1941-6008.1014

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

Gogoll J, Müller JF (2016) Autonomous cars: in favor of a mandatory ethics setting. Sci Eng Ethics. doi:10.1007/s11948-016-9806-x

Goodall NJ (2014a) Ethical decision making during automated vehicle crashes. Transp Res Rec J Transp Res Board 2424:58–65

[17] For their helpful comments on this paper, we are grateful to Hanno Sauer, Noah Goodall, John Danaher, Andreas Spahn, the participants of Eindhoven University of Technology's Philosophy and Ethics Workshop, and two anonymous reviewers for this journal. We are also grateful to Theo Hofman from our university's mechanical engineering department for his help with technical details regarding automated driving.

Goodall NJ (2014b) Machine ethics and automated vehicles. In: Meyer G, Beiker S (eds) Road vehicle automation. Springer, Dordrecht, pp. 93–102

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

Hansson SO (2012) A panorama of the philosophy of risk. In: Roeser S, Hillerbrand R, Sandin P, Peterson M (eds) Handbook of risk theory. Springer Science, Dordrecht, pp. 27–54

Hansson SO (2013) The ethics of risk. Ethical analysis in an uncertain world. Palgrave Macmillan, New York

Hayenhjelm M, Wolff J (2012). The moral problem of risk impositions: a survey of the literature. Eur J Philos, 20, E26–E51

Hevelke A. Nida-Rümelin J. (2014). Responsibility for crashes of autonomous vehicles: an ethical analysis. Sci Eng Ethics 1–12

Kamm F (2015). The trolley mysteries. Oxford: Oxford University Press

Lin P (2013, October 8). The ethics of autonomous cars. The Atlantic. Retrieved from http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/

Lin P (2015) Why ethics matters for autonomous cars. In: Maurer M, Gerdes JC, Lenz B, Winner H (eds) Autonomes fahren: Technische, rechtliche und gesellschaftliche aspekte. Springer, Berlin, Heidelberg, pp. 69–85

Mikhail J (2013). Elements of moral cognition. Cambridge: Cambridge University Press

Millar J (2014). Technology as moral proxy: autonomy and paternalism by design. IEEE Ethics in Engineering, Science and Technology Proceedings, IEEE Explore, Online resource, doi:10.1109/ETHICS.2014.6893388

Nyholm S, Smids J (in progress) Self-driving cars meet conventional cars: the ethics of mixed traffic

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 (1971). A theory of justice. Cambridge, Mass.: Harvard University Press

Rosén E, Stigson H, Sander U (2011) Literature review of pedestrian fatality risk as a function of car impact speed. Accid Anal Prev 43(1):25–33. doi:10.1016/j.aap.2010.04.003

Smids J (2015). The moral case for intelligent speed adaptation. J Appl Philos. (Early view, doi:10.1111/japp.12168)

Thomson JJ (1985) The trolley problem. Yale Law J 94(5):1395–1515

Thomson JJ (2008) Turning the trolley. Philos Public Aff 36(4):359–374

Van de Poel I (2011). The relation between forward-looking and backward-looking responsibility. In NA Vincent, I van de Poel, J van de Hoven (eds) Moral esponsibility beyond free will and determinism, Dordrecht: Springer

van Loon RJ, Martens MH (2015). Automated driving and its effect on the safety ecosystem: how do compatibility issues affect the transition period? Procedia Manuf, 3,3280–3285. doi:10.1016/j.promfg.2015.07.401

Wallach W, Allen C (2009) Moral machines: Teaching robots right from wrong, 1st edn. Oxford Scholarship Online

Windsor M. (2015). Will your self-driving car be programmed to kill you if it means saving more strangers? Retrieved February 19, 2016, from https://www.sciencedaily.com/releases/2015/06/150615124719.htm

Wood A (2011) Humanity as an end in itself. In: Parfit D, Scheffler S (eds) On What Matters (Vol. 2). Oxford University Press

Worstall T (2014, June 18). When should your driverless car from Google be allowed to kill you? Forbes Retrieved from: http://www.forbes.com/sites/timworstall/2014/06/18/when-should-your-driverless-car-from-google-be-allowed-to-kill-you/