The ethics of crashes with self-driving cars: a roadmap I

Received: 12 February 2018 | Accepted: 18 April 2018

DOI: 10.1111/phc3.12507

ARTICLE

WILEY

The ethics of crashes with self-driving cars: A roadmap, I

Sven Nyholm

Eindhoven University of Technology

Correspondence
Sven Nyholm, Eindhoven University of Technology, The Netherlands
Email: s.r.nyholm@tue.nl

Abstract

Self-driving cars hold out the promise of being much safer than regular cars. Yet they cannot be 100% safe. Accordingly, they need to be programmed for how to deal with crash scenarios. Should cars be programmed to always prioritize their owners, to minimize harm, or to respond to crashes on the basis of some other type of principle? The article first discusses whether everyone should have the same "ethics settings." Next, the oft-made analogy with the trolley problem is examined. Then follows an assessment of recent empirical work on lay-people's attitudes about crash algorithms relevant to the ethical issue of crash optimization. Finally, the article discusses what traditional ethical theories such as utilitarianism, Kantianism, virtue ethics, and contractualism imply about how cars should handle crash scenarios. The aim of the article is to provide an overview of the existing literature on these topics and to assess how far the discussion has gotten so far.

1 | INTRODUCTION

The year 2016 marked a turning point for crashes involving self-driving cars. Up until that point, there had been various minor crashes—about 20 of them in 2015. But no one had been injured. And the fault had always been attributed to human drivers, who had accidentally bumped in to self-driving cars (Schoettle & Sivak, 2015). In February 2016, however, it for the first time happened that a crash was obviously caused by a self-driving car. One of Google's self-driving cars crashed into a bus in Mountain View, California. Google assumed "partial responsibility," and promised to update the software of their cars. (Urmson, 2016) In May, something more tragic occurred.

The first person was killed in a self-driving car. A man riding in a Tesla Model S operating in "autopilot mode" crashed into a white truck that the car's sensors had not detected. Unlike Google, Tesla did not assume responsibility for what happened. Yet, like Google, they promised that they would update their cars—in this case their sensors—so that they would be better able to handle the given kind of accident scenarios (Tesla, 2016).

In March 2018, the first pedestrian was killed by a self-driving car. A test vehicle operated by Uber hit and killed a woman in Tempe, Arizona (Levin & Wong, 2018). This was another important turning point. These incidents in 2016 and 2018 illustrate that crashes involving self-driving cars are not merely material for hypothetical thought experiments. This is a real-world issue. It requires a serious response from both society and the developers of self-driving cars. Human lives are at risk. Accordingly, the new and developing topic of the ethics of crashes with self-driving cars is a very important one.

This is indeed a very new topic. The first papers on the topic in philosophy and ethics journals started appearing in 2014. Legal scholars had gotten a head start. They were already discussing ethically relevant legal issues related to self-driving cars in law review journals a few years earlier. But they too only really got started on this general topic very recently. Accordingly, it is still possible to have a fairly good overview of all that is going on within the field of the ethics of crashes with self-driving cars.

This is an opinionated review of the main things that have been discussed within this field so far. I will not be able to comment on everything that has come out on this topic within the space available to me here. But I will comment on as many things as possible. My aim is two-fold. On the one hand, I aim to introduce readers to some of the main contributions made so far. On the other hand, I aim to critically assess how far we have gotten in this discussion, so as to indicate where we might want to go next.

This and the accompanying article divide into three main parts. In the first part, I discuss questions related to how to program self-driving cars to respond to unavoidable crashes. In the second, I discuss who should be held responsible after a crash occurs. In the third, I discuss ethical issues related to ways in which we might try to minimize the total number of crashes involving self-driving cars. It is not only the occurrence of crashes with self-driving cars that raises ethical issues. Strategies for avoiding such crashes raise important ethical issues as well.[1]

2 | PART I: WHAT TO DO ABOUT UNAVOIDABLE CRASHES?

2.1 | The need for "ethics settings"

Self-driving cars hold out the promise of being much safer than regular cars. Whether or not this ultimately turns out being true, it is their main selling point. Yet even the safest self-driving cars will not be a 100% safe. Noah Goodall and Patrick Lin convincingly explain why in some of the earliest papers on this general topic (Goodall, 2014a, Goodall, 2014b; Lin, 2015).

Self-driving cars will drive in the midst of unpredictable pedestrians, bicyclists, human drivers, animals, and whatever else might appear in their paths. They will have to do so in different weather conditions, on different kinds of roads, with different qualities of upkeep (Nyholm & Smids, 2016). Even if there were only self-driving cars on our roads, the facts that they will be moving at high speeds and that they will be rather heavy will be sufficient to guarantee that they won't always be able to simply stop in order to avoid accidents (Goodall, 2014a, Goodall, 2014b). To think that we will not need to program self-driving cars for how to handle accident scenarios would be like thinking that the Titanic did not need any life-boats since it was an "unsinkable" ship.

It might seem like a good idea to always hand over control to a human driver in any accident scenario. However, typical human reaction-times are too slow for this to always be a good idea (Hevelke & Nida-Rümelin, 2015). So the cars themselves need to be pre-programmed for how to respond to different types of accident scenarios. Some accident scenarios will be such that (a) there are different options open to the self-driving cars; and (b) depending on what option is selected, different people will be put at risk (Goodall, 2014a, Goodall, 2014b; Lin, 2015). This is the basic reason why the choice between different possible accident-programs is an inherently ethical choice.[2]

One question here is whether all cars should have the same "ethics settings" (Gogoll & Müller, 2017). Should people who buy a self-driving car have a choice about whether, say, their cars should try to save them in an accident scenario, whether they should try to minimize overall harm, or whether they should perhaps be programmed according to some other principle? Jason Millar argues that a person's car should function as a "proxy" for their ethical outlook. People should therefore be able to choose their own ethics settings (Millar, 2014; see also Sandberg & Bradshaw-Martin, 2013). Similarly, Giuseppe Contissa and colleagues argue that self-driving cars should be equipped with an "ethical knob," so that whoever is currently using the car can set it to their preferred settings. (Contissa, Lagioia, & Sartor, 2017)

Jan Gogoll and Julian Müller, in contrast, argue that we all have self-interested reasons to want everyone's cars to be programmed according to the same settings. (Gogoll & Müller, 2017) Cars should have whatever ethics settings best promotes our chances of surviving any accidents we might be involved in. This would best be achieved, Gogoll and Müller argue, if cars were coordinated with each other in the ways they crash. Not mincing his words, Lin has called the idea of people getting to choose their own ethics settings a "terrible idea" (Lin, 2014). It is a terrible idea because some people might then potentially choose their ethics settings based on racist ideologies or other types of wholly unacceptable outlooks.

One question here is whether there might be a middle ground between the view that there should be an open choice of ethics settings and the view that everyone should have the same ethics settings. Perhaps there should be certain general boundaries within which everyone's cars need to make their choices about how to crash. But within these boundaries, perhaps some people should be permitted, say, to be more altruistic than others are expected to be, if that truly is what they sincerely wish. One advantage to giving people a certain degree of choice here is that this might make it easier to hold them responsible for any bad outcomes that crashes involving their vehicles might give rise to (Sandberg & Bradshaw-Martin, 2013; cf. Lin, 2014). As we will see in the next article, some authors worry about potential responsibility gaps opened up by self-driving cars.

2.2 | An applied trolley problem?

Consider now what types of dilemmas self-driving cars might conceivably face in accident scenarios. It is tempting to model these on the so-called trolley cases that have been widely discussed both in philosophy and psychology (Greene, 2013; Kamm, 2015). In the most classic trolley dilemmas, you can either save five people from being hit by a train by redirecting the train to a side-track where there is one person or by pushing somebody in front of the train. Figuring out why people typically respond very differently to these different cases is what is known as the "trolley problem" (Greene, 2013; Kamm, 2015).

It is easy to come up with similar types of dilemma-scenarios for self-driving cars. Accordingly, many academic articles, opinion pieces, and articles in other venues have likened the ethics of crashes with self-driving cars to the trolley problem (e.g., Achenbach, 2015; Bonnefon, Shariff, & Rahwan, 2015; Doctorow, 2015; Lin, 2014; Lin, 2015; Wallach & Allen, 2009; Windsor, 2015; Worstall, 2014).

For example, imagine that a self-driving car carrying five passengers suddenly detects that a crash with a heavy obstacle on the road is unavoidable. The five will die unless the car swerves onto the sidewalk, where there is one pedestrian, who would be killed as a result. Imagine next that there is only one passenger in the self-driving car. Once again there suddenly appears a large obstacle on the road. The only way to avoid a deadly crash is to once more swerve onto the sidewalk. In this variation of the example, though, there are five pedestrians on the sidewalk. Should the self-driving car let its passenger die, by crashing into the large obstacle rather than driving up on the sidewalk and potentially killing the five? In the first variation, should the self-driving car save the five in the car, by crashing into the one pedestrian on the sidewalk?

These examples are designed to be similar to the examples typically discussed in the literature on the trolley problem. Lots of other examples modeled on that literature have also been suggested. One such example involves the choice between crashing into a motorcyclist wearing a helmet and crashing into another not wearing a helmet

(Goodall, 2014b). Another example—constructed by a critic poking fun at these dilemmas—involves choosing between crashing into a criminal and crashing into a nun (McFarland, 2015). Just like trolley problems have gotten rather fanciful in the hands of authors like Frances Kamm, these dilemmas involving self-driving cars have gotten very fanciful in the hands of some writers as well.

One of the questions this raises is whether the vast literature on the trolley problem might be a useful source of ideas about how to deal with the ethics of crashing self-driving cars. Together with Jilles Smids, I have put forward three reasons for being skeptical about relying very heavily on the trolley problem literature here (Nyholm & Smids, 2016).[3]

Firstly, in the trolley literature, we are typically asked to imagine that the only morally relevant factors are a very small set of factors. Any bigger and more complex sets of considerations are imagined away. In the real-world ethics of self-driving cars, however, we do not have the luxury of simply abstracting away the real decision-making scenario we are facing, with all its complexity and messiness that we need to take into account. Secondly, in most trolley discussions, we are asked to set all questions of moral and legal responsibility aside, and only focus on the choice between the one and the five. In actual traffic ethics, we cannot ignore questions about responsibility. They are among the most central questions here. Thirdly, in trolley discussions, a fully deterministic scenario is imagined. It is assumed that we know with certainty what the outcomes of our available choices would be. In contrast, when we are prospectively programming self-driving cars for how to deal with accident scenarios, we do not know what scenarios they will face. We must make risk-assessments. We must make most of our decisions in the face of uncertainty (Nyholm & Smids, 2016).

The just-mentioned last point has also been stressed by Goodall in some of his recent papers. In one of those, he urges us to move away from deterministic trolley problems, and towards an ethics of risk-management instead (Goodall, 2016; cf. JafariNaimi, 2018). At the same time, Goodall also argues that considering thought experiments similar to those associated with the trolley problem has an undeniable value in this discussion. It can help us to stress-test our initial ideas about how self-driving cars should be programmed. It can also open up other—and perhaps more fruitful—avenues of inquiry. Moreover, it can help to awaken people's curiosity and interest in the ethics of crashes with self-driving cars. So the literature on the trolley problem may not always have been discussing issues directly relevant to the ethics of self-driving cars. But we should not take that to mean that there is no point in considering similar thought experiments within the ethics of crashes with self-driving cars.

2.3 | Empirical ethics

Staying on the topic of thought experiments a little longer, we can next consider another interesting way in which this topic has been approached. We can consider the "empirical ethics" approach taken by Jean-François Bonnefon and colleagues—a group of psychologists and behavioral economists (Bonnefon, Sharrif, & Rahwan, 2016). They were inspired by Joshua Greene's earlier work. Greene had combined psychological studies about people's reactions to moral dilemmas with philosophical premises, so as to generate empirically informed ethical arguments (Greene, 2013). That is what is meant by empirical ethics here.

The most widely publicized finding from Bonnefon et al. is the following. When people are asked about how other people's cars should be programmed to handle accident scenarios, they are inclined to want them to simply minimize overall harm. However, when surveyed about what kinds of cars they themselves would want to use, people tend to favor cars that would save them in an accident scenario. At the very least, they do not want to be forced to buy cars that would be "altruistic" by seeking to minimize overall harm (Bonnefon et al., 2016).[4]

On a website set up by these same researchers—the "moral machine" website—one can explore various other versions of these ethical dilemmas. One can also create one's own dilemma-scenarios.[5] The researchers can then use this data in their further work on interesting patterns in the general public's attitudes. But how good are these findings as normative premises in ethical arguments about how self-driving cars should be programmed to crash?

Again, I want to voice some skepticism, and again I will offer three reasons for being skeptical. Firstly, most people still have very little real experience with self-driving cars. It is likely that their attitudes will change once they have more actual experience with them. This suggests that we should not put too much weight on people's current attitudes about this technology.

Secondly, when people are asked to offer up an intuitive response to an imagined moral dilemma, they are typically not asked to justify their responses. Instead, researchers are simply testing what preferences subjects have among the choices presented to them. But in ethical arguments, it is important to articulate and assess arguments in favor of or against the different options that are being considered (cf. Kahane, 2015).

Thirdly, people appear to have inconsistent or paradoxical attitudes. In the finding mentioned above, many people want others to have harm-minimizing cars, while themselves wanting to have cars that would favor them. It is of course very interesting to know that people's attitudes exhibit these asymmetries. But we could not plausibly put forward an ethical argument according to which some people should have harm-minimizing cars (everyone else!) while others should have cars saving their owners (us!).[6]

It is also interesting to note here that, in a way, this scenario recently played out in real life recently as well. A representative of Mercedes, Christoph von Hugo, was interviewed during an auto-show in Paris in 2016. 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). Given people's above-discussed attitudes suggesting that they would prefer buying such a car, 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. Von Hugo ended up claiming that his previous statements—which included 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).

What should we make of this? Perhaps the people who were outraged about what von Hugo had said were thinking of themselves as not being among those who would buy self-driving Mercedes cars. This would be consistent with people's apparently widely shared attitudes of wanting others (in this case Mercedes-owners) to have harm-minimizing cars. I take this to be a further indication that what we need are neither intuitive reactions (as in the research by Bonnefon et al.) nor off-the-cuff arguments quickly formulated when we are put on the spot (as in von Hugo's case). We need carefully thought-out arguments that can be fully articulated and critically assessed by all who participate in this debate.

2.4 | Traditional ethical theories

One way of formulating more fully spelled out arguments about how self-driving cars should crash is to draw on the traditional ethical theories as they are sometimes used in applied ethics taking a "top-down" approach.[7] That is, we can consider what utilitarians (or consequentialists more broadly), Kantians (or deontologists more broadly), virtue ethicists, or contractualists would recommend regarding this topic.

Utilitarian ethics is about maximizing overall happiness, while minimizing overall suffering. Kantian ethics is about adopting a set of basic principles ("maxims") fit to serve as universal laws, in accordance with which all are treated as ends-in-themselves and never as mere means. Virtue ethics is about cultivating and then fully realizing a set of basic virtues and excellences. Contractualist ethics is about formulating guidelines people would be willing to adopt as a shared set of rules, based on nonmoral or self-interested reasons, in a hypothetical scenario where they would be making an unforced agreement about how to live together[8] (see, e.g., Suikkanen, 2014). What arguments could we formulate about crashes with self-driving cars using these theories?

I present this idea in this slightly hypothetical-sounding way, because at the time of writing, very few articles have been published that (a) explicitly endorse one of these types of ethical reasoning and then (b) applies that to this issue in a top-down sort of way. Two key exceptions are papers by Gogoll and Müller and by Derek Leben. Those papers can be classified as employing forms of contractualist moral reasoning in defense of specific conclusions

regarding accident-programming. But most papers that discuss the above-mentioned moral theories in relation to these ethical issues do so in an exploratory way. They investigate what these theories might imply here, while often raising worries about using these different theories in such arguments. Or they investigate whether it would at all be possible to program self-driving cars using algorithms based on these moral theories (e.g., Gerdes & Thornton, 2015; Kumfer & Burgess, 2015).

For example, in a thorough investigation of the ethics and law of automated driving, the legal scholar Jeffrey Gurney first imagines a series of dilemma-scenarios of the sorts discussed above. He then investigates what utilitarians and Kantians, as he understands them, would say about those dilemmas (Gurney, 2016). Lin and Goodall do similar things in their early papers (Goodall, 2014a, Goodall, 2014b; Lin, 2015). But I will focus on what Gurney says about utilitarianism to illustrate that what the moral theories imply about the choice of accident-programming is not always a straightforward matter. Rather, it is something that is up for debate.

Gurney suggests that a self-driving car could be equipped with a powerful computer enabling it to make utilitarian calculations about the expected utility of different options in a much quicker and more reliable way than a human could ever do. Accordingly, Gurney argues that utilitarians would recommend that self-driving cars be equipped with these capabilities, and that they would be programmed to always crash in ways that maximize expected utility (Gurney, 2016).

However, it is not altogether clear that this is what a utilitarian would necessarily recommend. A utilitarian would be mindful of the fact that people might be scared of taking rides in "utilitarian" cars, instead preferring cars programmed to prioritize their passengers. After all, this is what was indicated by the surveys from Bonnefon et al. described above. A clever utilitarian might take this into account, and recommend that self-driving cars be programmed to save their owners. The utilitarian might recommend this if it were the case that having a maximum of people willingly using self-driving cars rather than regular cars would be likely to bring down the overall number of deaths and injuries in traffic. Utilitarians would recommend whatever solution would best promote overall happiness. This might mean that people must be lured into self-driving cars by the promise that their cars would be programmed to behave in "non-utilitarian" ways in crashes.

This also highlights that we need to consider who exactly is the moral agent who needs to make a choice. Is it the car itself? Perhaps its best way of maximizing utility would be to "tell" people that it is programmed in whatever way they prefer, but then actually crash in whatever ways would maximize utility. Or is the moral agent the person designing the car? Or perhaps the regulatory body permitting certain types of cars on the road? When we try to apply traditional moral theories to this problem, it is useful to ask ourselves who exactly the moral agent is who is making the decisions about how self-driving cars should crash.

Another issue here is whether we need to choose among these different moral theories, such that we can either use utilitarian or Kantian or virtue ethical or contractualist reasoning. One view would be that we have to make a choice and that we cannot draw on more than one type of reasoning here. My own opinion is that there are lessons to learn from all of these different perspectives. For example, the lesson to be learned from utilitarian ethics might be that when we think about how self-driving cars should crash, we should do this partly with an eye to what would be best for the overall good, and everyone's happiness. The lesson from Kantian ethics might be that we should choose rules we would be willing to have as universal laws applying equally to all—so as to make everything fair, and not give some people an unjustified advantage in crash-scenarios.

What about virtue ethics? It is hard to come up with any virtue ethical ideas about how self-driving cars should crash (cf. Gurney, 2016). But virtue ethics might help when we think about the ethics of automated driving more generally. A paper by Mark Coeckelbergh about self-driving cars and the phenomenology of moral responsibility might be relevant here (Coeckelbergh, 2016). Coeckelbergh argues that how careful people are while using their cars and how responsible they feel about their car-use both depend on the design of the cars they are using. Why is this observation relevant to virtue ethics? Well, being careful and taking responsibility for one's actions are virtues we value in people who use risky technologies like cars. So perhaps a lesson from a virtue ethical perspective is that we should try to design and program cars in ways that help to make people act carefully and responsibly when they

use self-driving cars. Self-driving cars could then become what Mark Alfano calls "moral technologies" (Alfano, 2013). That is, they could become technologies that help to bring out virtues in people.[9]

Those are all hypothetical suggestions about how one might use traditional moral theories to formulate arguments in this discussion. But as I mentioned above, there have also been contributions that take a more firm stance in favor of certain types of contractualist reasoning about this topic. Firstly, Gogoll and Müller discuss the question of what we all prospectively have self-interested reasons to want in this domain. This is precisely the type of question a contractualist would want us to ask about this topic. (Gogoll & Müller, 2017) What Gogoll and Müller argue using this type of argument is that we all have self-interested reasons to want everyone to use cars programmed to minimize harm in crash-scenarios. This is a contractualist justification since the justification is not based on the moral value of minimizing harm, but rather on the self-interested value for each person of maximizing their own chances of surviving any accidents they might be involved in (cf. Harsanyi, 1977).

Secondly, another argument drawing on the contractualist tradition is featured in a paper by Derek Leben. Leben presents an argument in favor of a "Rawlsian" accident-algorithm. Behind a veil of ignorance, Leben argues, it would be rational to choose accident-algorithms that would protect whoever is most vulnerable in any accident scenario. That way, we would make things as good as possible for whoever is worse off in a crash (Leben, 2017). This is another type of contractualist argument. (Rawls, 1971).

What should we think about these arguments?[10] In a response to Leben's paper, Geoff Keeling argues, firstly, that Leben has not given us a correct interpretation of how John Rawls (or somebody like Rawls) would reason about this topic. Secondly, Keeling also argues that Leben's most important decision-theoretical claim—namely, that it is prospectively rational to favor a "Rawlsian" accident-algorithm—does not correctly identify what is rational to choose from a decision-theoretical perspective (Keeling, 2018b).

I myself am perhaps less interested in whether Leben's argument is a good interpretation of Rawls. I am more concerned with whether the argument itself is a good one on its own terms. My own objection to both Leben's argument and to Gogoll and Müller's argument would rather be that we should not accept the basic contractualist premise that moral arguments should be made by first setting aside moral values and then asking what it would be rational to choose for self-interested reasons. It seems better to me to formulate arguments that explicitly refer to distinctly moral values or principles. One reason for this is that after a crash, when the outcome is evaluated, we will be wanting to know if there was any clear moral value or principle that could be used to directly justify the way the crash happened.

There is clearly more than can be said about these arguments. But this concludes the first part. In the next part, we start things off by first picking up the thread about what should happen after a crash. In particular, I will discuss who should be held responsible for any bad outcomes after a crash. This will lead us to also ask what type (s) of agency, if any, we can attribute to self-driving cars. How does that artificial agency relate to the human agency of those who design, update, or use self-driving cars? Having discussed that topic, we will also consider ethical issues that arise in preemptive risk-management aimed at preventing crashes in the first place.

ACKNOWLEDGEMENT

Many thanks to John Danaher, Noah Goodall, Dieter Huebner, Geoff Keeling, Will McNeill, Lucie White, and an anonymous reviewer for very helpful comments.

ENDNOTES

[1] There are also other interesting ethical issues related to self-driving cars that have nothing directly to do with either crashes or crash-avoidance. Other ethical questions concern employment, economic impact, privacy, cyber security, possible effects on well-being, and so on. These other issues have not been explored much in academic ethics yet. Accordingly, I set them aside here.

[2] Not all crashes (or potential crashes) will involve interesting ethical dilemmas. In many scenarios, the best thing for the car to do is simply to brake. However, in other scenarios, this will not be a realistic option—for example, if the road is slippery or if heavy traffic is approaching from behind. All the realistic options will sometimes involve one or more persons getting

hurt or some other type of nontrivial sacrifice. Simply hoping that such scenarios will not arise very often and only planning for accident-scenarios where it is obvious to everyone what the best thing to do is would be deeply irresponsible.

[3] Dietmar Hübner and Lucie White (Hübner & White, under review) make a convincing case that the earliest discussions of the trolley problem—specifically Philippa Foot on negative and positive rights and Judith Jarvis Thomson on claim-based aspects of dilemma situations—can helpfully inform the ethics of crashes with self-driving cars.

[4] Other noteworthy empirical investigations are those by The Open Roboethics Initiative, who polled Robohub's readers about how they would make decisions in crash scenarios, and those by psychologists Faulhaber et al. (forthcoming), who tested people's ethical judgments made while in a car simulator. The Robohub poll is reported here: http://robohub.org/if-a-death-by-an-autonomous-car-is-unavoidable-who-should-die-results-from-our-reader-poll/

[5] http://moralmachine.mit.edu/

[6] Nassim JafariNaimi presents the following argument against empirical ethics modeled on the trolley problem: "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, 5).

[7] Another approach is to draw on established legal doctrines and to try to apply them to this issue in the ethics of crashes with self-driving cars. Filippo Santoni de Sio and Geoff Keeling have done this using the Anglo-American doctrine of necessity, and Ivó Coca-Vila has done this using doctrines of justification from criminal law (Coco-Vila, 2018; Keeling, 2018a; Santoni de Sio, 2017).

[8] There are also contractualist theories—such as that of T.M. Scanlon (1998)—that do not focus on what is self-interested to accept in a hypothetical scenario. In the context of his discussion of the legal doctrine of necessity, Keeling (2018a) briefly discusses Scanlonian contractualism in relation to crashes with automated cars. Other than that, nobody has explored Scanlonian contractualism in this context yet to my knowledge.

[9] In contrast to this, John Danaher is skeptical about automated technologies' potential to make us into morally better people. He argues that autonomous technologies' taking over many human tasks might diminish our capacity for moral agency, downgrading us into mere moral patients with greatly decreased capacities for moral agency (Danaher, forthcoming).

[10] Hübner and White (under review) argue that Gogoll and Müller's contractualist argument in favor of overall harm-minimization overlooks important ethical distinctions related to ways in which people can be more or less involved in risky traffic-situations, which should affect what rights they have.

ORCID

Sven Nyholm http://orcid.org/0000-0002-3836-5932

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

Alfano, M. (2013). *Character as a moral fiction*. Cambridge University Press.

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 1510.03346

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

Coco-Vila, I. (2018). Self-driving cars in dilemmatic situations: An approach based on the theory of justification in criminal law. *Criminal Law and Philosophy*, *12*, 59–82. https://doi.org/10.1007/s11572-017-9411-3

Coeckelbergh, M. (2016). Responsibility and the moral phenomenonology of using self-driving cars. *Applied Artificial Intelligence*, *30*(8), 748–757.

Contissa, G., Lagioia, F., & Sartor, G. (2017). The Ethical Knob: Ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, *25*(3), 365–378.

Danaher, J. (forthcoming). The rise of robots and the crisis of moral patiency. *AI & Society*. https://doi.org/10.1007/s00146-017-0773-9

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

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., … König, P. (forthcoming). Human decisions in moral dilemmas are largely described by utilitarianism. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-018-0020-x

Gerdes, J. C., & Thornton, S. M. (2015). Implementable ethics for autonomous vehicles. In M. Maurer, C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous driving. Technical, legal and social aspects* (pp. 687–706). Springer.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics*, *23*(3), 681–700.

Goodall, N. J. (2014a). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board*, *2424*, 58–65.

Goodall, N. J. (2014b). Machine ethics and automated vehicles. In G. Meyer, & S. Beiker (Eds.), *Road vehicle automation* (pp. 93–102). Dordrecht: Springer.

Goodall, N. J. (2016). Away from trolleys and toward risk-management. *Applied Artificial Intelligence*, *30*(8), 810–821.

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

Gurney, J. K. (2016). Crashing into the unknown: An examination of crash-optimization algorithms through the two lanes of ethics and law. *Albany Law Review*, *79*(1), 183–267.

Harsanyi, J. (1977). *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge: Cambridge University Press.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics*, *21*, 619–630.

Hübner, D. & White, L. (under review). Crash algorithms for autonomous cars beyond harm minimisation.

JafariNaimi, N. (2018). Our bodies in the trolley's path, or why self-driving cars must *Not* be programmed to kill. *Science, Technology, & Human Values*, *43*, 302–323. https://doi.org/10.1177/0162243917718942

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian moral judgment. *Social Neuroscience*, *10*(5), 551–560.

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

Keeling, G. (2018a). Legal necessity, pareto efficiency & justified killing autonomous vehicle collisions. *Ethical Theory and Moral Practice*. https://doi.org/10.1007/s10677-018-9887-5

Keeling, G. (2018b). Against Leben's Rawlsian collision algorithm for autonomous vehicles. In V. C. Müller (Ed.), *Philosophy and theory of artificial intelligence III*. Berlin: Springer: SAPERE.

Kumfer, W., & Burgess, R. (2015). Investigation into the role of rational ethics in crashes of automated vehicles. *Transportation Research Record*, *2489*, 130–136.

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, *19*(2), 107–115.

Levin, S, & Wong, J. C. (2018). Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian, The Guardian. (Accessed April 27, 2018)

Lin, P. (2014). Here's a terrible idea: Robot cars with adjustable ethics settings. *Wired*. Accessed on January 31, 2018

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

McFarland, M. (2015). Google's chief of self-driving cars downplays “the trolley problem”. The Washington Post December 1. (Accessed January 31, 2018)

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: https://doi.org/10.1109/ETHICS.2014.6893388

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice*, *19*(5), 1275–1289.

Orlove, R. (2016). Now Mercedes says its driversless cars won't run over pedestrians, that would be illegal. *Jalopnik* October 17, 2016. Accessed January 31, 2018

Rawls, J. (1971). *A Theory of Justice*. Cambridge, Mass: Harvard University Press.

Sandberg, A., & Bradshaw-Martin, H. (2013). In J. Romportl, et al. (Eds.), *What do cars think of trolley problems: Ethics for autonomous cars?* . Beyond AI: Artificial Golem Intelligence, Conference Proceedings. https://www.beyondai.zcu.cz/files/BAI2013_proceedings.pdf:12

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

Scanlon, T. M. (1998). *What we owe to each other*. Cambridge, Mass: Harvard University Press.

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. Google Scholar

Suikkanen, J. (2014) *This is ethics*, Wiley-Blackwell

Taylor, M. (2016). Self-driving Mercedes Benzes will prioritize occupant safety over pedestrian *Car and Driver* October 7, 2016. Retrieved from https://blog.caranddriver.com/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/ (Accessed January 31, 2018)

Tesla. (2016). A tragic loss, blogpost at https://www.tesla.com/blog/tragic-loss

Urmson, C. (2016). How a self-driving car sees the world, ted-talk. Retrieved from https://www.ted.com/talks/chris_urmson_how_a_driverless_car_sees_the_road/transcript (Accessed January 31, 2018)

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong* (1st ed.). Oxford: Oxford University Press.

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

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/

Sven Nyholm is an assistant professor of philosophy and ethics at the Eindhoven University of Technology. His research interests are ethical theory and the philosophy of technology. His articles have appeared in general philosophy journals, ethics journals, and applied ethics journals. His first book, published in 2015, was about Kantian ethics. Recently, he has published on the philosophy of love and ethical issues related to self-driving cars, sex robots, and automated weapons systems.

How to cite this article: Nyholm S. The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*. 2018;13:e12507. https://doi.org/10.1111/phc3.12507