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

Received: 12 February 2018 | Accepted: 18 April 2018

DOI: 10.1111/phc3.12506

ARTICLE

WILEY

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

Sven Nyholm

Eindhoven University of Technology

Correspondence
Sven Nyholm, Eindhoven University of Technology, Eindhoven, 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, we need to think about who should be held responsible when self-driving cars crash and people are injured or killed. We also need to examine what new ethical obligations might be created for car users by the safety potential of self-driving cars. The article first considers what lessons might be learned from the growing legal literature on responsibility for crashes with self-driving cars. Next, worries about responsibility gaps and retribution gaps from the philosophical literature are introduced. This leads to a discussion of whether self-driving cars are a form of agents that act independently of human agents. It is suggested that it is better to analyze their apparent agency in terms of human–robot collaborations, within which humans play the most important roles. The next topic is the idea that the safety potential of self-driving cars might create a duty to either switch to self-driving cars or seek means of making conventional cars safer. Lastly, there is a short discussion of ethical issues related to safe human–robot coordination within mixed traffic featuring both self-driving cars and conventional cars.

1 | INTRODUCTION

Some major car manufacturers have recently promised that when the fully self-driving cars that they are developing are ready to go on the market, they will take responsibility in case any crashes occur. Volvo and Audi are among the

companies that have made this promise (Atiyeh, 2015; Maric, 2017). This indicates that they expect their self-driving cars to be very safe. In contrast, car manufacturers that already have cars for sale with a certain degree of automation – like Tesla with their "autopilot" feature – are more cautious. For example, Tesla currently says that when accidents occur, the responsibility remains with the car owner (Tesla, 2016). In line with what Volvo and Audi claim about their future self-driving cars, however, Tesla claims that their cars with the autopilot feature are already safer than regular cars without any type of automation. In general, the promise from all major car companies developing self-driving cars is that future versions of self-driving cars will be much safer than any cars currently available on the market.

What should we make of all of this from an ethical point of view? When there are crashes involving self-driving cars, who should be held responsible? If a fully automated car is seemingly operating as some form of artificial autonomous agent on the road, can any human beings fairly be held responsible for any harm the car might cause? And if it really is the case that self-driving cars will prove to be much safer than regular cars, what new ethical duties might this create for people who use cars?

These are the questions discussed below. My aim is to provide an opinionated overview and assessment of the discussion about these topics so far in the literature. I start with the question of responsibility. Then I consider whether self-driving cars can be understood as a type of autonomous agents that operate with some significant degree of independence from human agents. I end by discussing what new ethical duties might be created by self-driving cars' potential to be safer than conventional cars.

2 | PART II: WHAT SHOULD HAPPEN AFTER A CRASH?

2.1 | Lessons from the legal literature

Legal scholars started discussing crashes involving autonomous vehicles a little earlier than moral philosophers did. Accordingly, there is more literature about legal issues related to holding people responsible for crashes with autonomous cars than there is about ethical issues related to this. The legal literature has mostly focused on the American context. It has discussed things such as how to extend American civil and criminal law to the issue of holding people or organizations responsible for crashes with autonomous cars[1] (e.g., Beiker, 2012; Gurney, 2013; Gurney, 2015; Gurney, 2016; Marchant & Lindor, 2012; Peterson, 2012; Ravid, 2014). So we need to be careful about generalizing the findings in the legal literature to other contexts, such as other legal systems, or to ethics more generally. Nevertheless, there are some interesting general lessons to be learned from the legal literature on this topic. I will consider two broad lessons here.

The first is that the introduction of self-driving cars onto public roads raises a sort of "existential crisis" within the domain of traffic. That is to say, when more and more driving tasks are handed over to automated systems, and we move towards full automation, this raises questions about how to understand the new role and responsibilities of the driver. It also raises questions about how to reinterpret the relation between the driver and his or her car. Self-driving cars are being introduced into society before we have clear ethical and legal frameworks for how to understand them and our relation to them. Consider, firstly, the role of the driver in a fully automated car: Is it more like that of a manager, a supervisor, a commander, or what? Consider next the relation between a user of an automated car and the car itself. Is it more like the relation between an employer and their employee, between a master and a slave, a superior and a subordinate, a principal and an agent, or what? The legal literature has usefully pointed out that we need to reconceptualize the way that we typically understand the relation between a car user and a car. This can significantly affect what car users' responsibilities should be thought to be.

The second key lesson from the legal literature is that when we reflect on responsibility for crashes involving autonomous cars, we should not forget that in ordinary life, what people actually do is not the only commonly recognized ground for attributions of responsibility to them. We often also attribute responsibility for outcomes to people based on roles they perform (which could be either professional roles or private roles) or rights they enjoy (which

involve things such as ownership rights). For example, if you are a dog owner and your dog bites somebody, you might be held both legally and morally responsible for the harm caused even though it was not you, but the dog, who sank their teeth into the victim (Ravid, 2014). In this same way, when we trace the moral responsibility for outcomes caused by automated technologies back to certain people, the right way to do this may sometimes not be via the most immediate actions of these people. It may rather be via the roles they play or the rights they enjoy in relation to the automated technologies in question.

2.2 | Potential responsibility gaps and retribution gaps

Let us now consider the philosophical literature. In an influential discussion about military robots, Robert Sparrow argued in a 2007 article that military robots may create worrying responsibility gaps because people may be unable to predict and directly control what autonomous military robots will do. A "responsibility gap" here refers to a situation where it is unclear who we can justifiably hold responsible for an outcome, for example, if somebody is killed by a military robot. Programmers, Sparrow worried, will not be able to fully predict what their creations will do. And commanders may not be able to fully control the robots they deploy. It is not fair, according to this argument, to hold people responsible for what they cannot fully predict or directly control (Sparrow, 2007). The most important contribution to the philosophical discussion about responsibility for crashes with automated cars so far – a 2015 paper by Alexander Hevelke and Julian Nida-Rümelin – bears some interesting similarities to Sparrow's discussion of potential responsibility gaps related to military robots.

Before getting to the arguments similar to those in Sparrow's paper, however, Hevelke and Nida-Rümelin present an argument that they borrow from a 2012 paper by the legal scholars Gary Marchant and Rachel Lindor. This is an argument that the opening anecdote about Volvo and Audi might seem to disprove. The argument is that if we decide to hold car manufacturers responsible for harms caused by the self-driving cars they develop, this might discourage car-manufacturers from developing autonomous cars (Marchant & Lindor, 2012). That would be bad, Hevelke and Nida-Rümelin think, because the safety potential of self-driving cars is an important reason why we should encourage their development (Hevelke & Nida-Rümelin, 2015). But like I just said, recent developments with companies like Volvo and Audi seem to clash with that prediction in Hevelke and Nida-Rümelin's paper.

Moreover, we should note that whereas a pragmatic argument like the one Hevelke and Nida-Rümelin present against holding car manufacturers responsible is certainly important to consider, it does not settle the question of whether it is just or fair to hold car manufacturers responsible for harms or deaths their cars might cause. That is a separate issue from the issue of whether holding them responsible might have bad effects, such as halting the development of self-driving cars (cf. Darwall, 2006).

When it comes to other possible contenders for being held responsible, Hevelke and Nida-Rümelin think that it does not make any sense to hold the car itself responsible for any harm it might cause. But what about holding drivers accountable as people who have a special duty of care while using their cars? Here, Hevelke and Nida-Rümelin argue that people cannot be expected to pay sufficient attention to what their cars are doing. Nor will people have quick enough reactions in accident scenarios for it to be fair to hold drivers responsible. It is simply too hard to discharge this duty of care. What about the driver under the guise of somebody who introduces a risk into society? Here, too, Hevelke and Nida-Rümelin argue that it would be unfair to hold drivers whose cars crash responsible. They will not be doing anything different from others whose cars luckily do not crash. This would be an unfair type of moral luck (Hevelke & Nida-Rümelin, 2015).

Nevertheless, Hevelke and Nida-Rümelin think that there is a solution that is both fair and good. Namely, we can hold all users of self-driving cars collectively responsible for the risks that they as a group introduce into society. They can be held responsible even before the fact of an actual accident by being forced to pay a mandatory tax or insurance motivated by the risks that they impose on others simply by using these cars (Hevelke & Nida-Rümelin, 2015).

One worry here is that this collective tax solution might make it seem like no one in particular is responsible for any bad outcomes that might occur. This might give rise to what John Danaher calls "retribution gaps." As Danaher

points out, most people have a tendency to want to see particular individuals punished when somebody is harmed or killed as a result of a risk that somebody brings about (Danaher, 2016). A general tax on all users of self-driving cars may not satisfy people's sense of justice when it takes the form of a desire to find somebody to punish.

Danaher himself worries that technologies like self-driving cars and military robots will give rise to many retribution gaps. Similar to how Sparrow reasons in his paper, Danaher argues that different kinds of autonomous robots will soon operate in ways that most people will not fully understand. And our control over these technologies will be limited. These robots themselves will not be fit to be held responsible for what they do. This will open up many responsibility and retribution gaps in the domain of automation and robotic technologies (Danaher, 2016).[2]

Relating this back to the lessons from the legal literature discussed above, the first thing to note about these arguments is that they seem to assume that the main sources of responsibility there are all have to do with the direct agency we exercise as individuals. That is, these arguments seem to overlook that we can have also responsibilities based on the roles that we play or the rights that we enjoy in relation to people or things that might bring about bad outcomes.

The second thing to notice here is that these arguments appear to portray self-driving cars as operating in a very independent way, outside of the control and supervision of human beings. They are treated as independent agents of their own – which is the reason why retribution and responsibility gaps are supposed to open up here. Are these machines agents that can act independently of human beings?

2.3 | Agency and human–robot collaborations

It is worth highlighting that in the ethics literature about self-driving cars, many writers happily attribute agency to these cars. They often do so without giving any evidence of intending to speak in a loose and metaphorical way. Here are some sample quotes:

> *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, p. 69)*

> *Driverless cars, like [automated weapons systems], would likely be required to make life and death decisions in the course of operation. (Purves et al., 2015, p. 855)*

> *Driverless systems put machines in the position of making split-second decisions that could have life or death implications. (Wallach & Allen, 2009, p. 14)*

If a robot were indeed able to "replicate the human decision-making process," "make life and death decisions," or make "split-second decisions," the robot would be an agent. Hence, these just-cited commentators all attribute agency, or decision-making capacities, to the automated systems themselves. Mark Coeckelbergh is even more explicit about this. He writes that when a human being uses an automated car, "all agency is entirely transferred to the machine" (Coeckelbergh, 2016, p. 754). If all agency is indeed entirely transferred to the machine, this might justify the worries about responsibility and retribution gaps we considered above. But is all agency entirely transferred to these machines?

Let us now consider an alternative way of thinking about agency in automated systems. In doing so, we can start with the following quote about military robots from a 2012 report published by the US Defense Science Board:

> *… there are no fully autonomous systems just as there are no fully autonomous soldiers, sailors, airmen, or Marines …. Perhaps the most important message for commanders is that all machines are supervised by humans to some degree, and the best capabilities result from the coordination and collaboration of humans and machines. (US Defense Science Board, 2012, p. 24)*

Similarly, the engineer and technology historian David Mindell argues that the perspective in the quote should be taken on most types of automated systems, from space ships to deep sea exploration robots to airplanes, and so on (Mindell, 2015). Taking inspiration from this point of view, I have recently argued that we should also think of

self-driving cars in these terms. They too are always supervised by humans to some degree. Therefore, if we do attribute agency to them, we should think of this as a form of collaborative agency, where the key partners in these human–robot collaborations are certain humans. After all, humans set self-driving cars' goals (e.g., going to the grocery store). Humans update them when they are not happy with their performance or simply stop using them if they do not live up to our expectations (Nyholm, forthcoming). In these ways, we can understand self-driving cars as working for us. And we can be understood as having power over them (Nyholm, forthcoming; cf. Pettit, 1996).

This means that when we allocate responsibility for any harms or deaths caused by these technologies in terms of the agency involved in these cases, we should not only focus on theories of individual agency and responsibility for individual agency. We should rather draw on philosophical analyses of collaborative agency and responsibility for such agency. In particular, my suggestion is that we should draw on hierarchical models of collaborative agency, where some agents within the collaborations are under other agents' supervision and authority. People can be responsible for the apparent agency of automated cars precisely for the reason that they are supervisors, managers, or some other type of authority figures in relation to the cars (Nyholm, forthcoming).[3]

A similar view is developed in work by Catrin Misselhorn and colleagues while discussing robots more generally. They analyze robotic agency in terms of human–robot collaborations as well (Misselhorn, 2015). This leads them to argue that we should think in terms of what they call "responsibility networks" (Loh & Loh, 2017; Neuhäuser, 2015). Humans can be responsible for the apparent agency on the part of robotic agents such as automated cars based on their roles within these responsibility networks.

Others are even more skeptical about the whole idea of attributing agency to any robots – whether it is self-driving cars, autonomous weapons systems, or whatever. Philip Brey, for example, argues that agency involves acting on beliefs and desires and that robots lack beliefs and desires (Brey, 2013). Duncan Purves and colleagues argue that agency involves acting on the basis of reasons and that robots lack the capacity to act on the basis of reasons (Purves et al., 2015). My own view is that there are some aspects of agency that we can certainly associate with robots such as self-driving cars. For example, there is the capacity to pursue goals in a way that is sensitive to representations of the environment. That is something a self-driving car is seemingly able to do (Nyholm, forthcoming). Nevertheless, I agree with writers like Brey and Purves et al. that there are other dimensions of human agency that no current types of robots are able to replicate.

It is worth noting here, however, that there are also papers that go very much in the opposite direction. Jaap Hage, for example, argues that the attribution of agency and the attribution of responsibility are both wholly conventional matters. If there are important goods that we can realize by attributing agency and responsibility to automated systems – such as self-driving cars – there is nothing in principle to stop us from doing so (Hage, 2017, but see also Brożek & Jakubiec, 2017). From this point of view, the only real question here is whether there would be any significant gain in having a practice of sometimes attributing responsibility and blame to the machines themselves.

Whatever we might want to say about the different aspects of agency that self-driving cars and other current robots may or may not be able to exemplify, the most important things to investigate, I suggest, are as follows: (a) the roles people have in relation to these machines; (b) the rights they enjoy in relation to them; and (c) the ways in which they exercise not only direct but also indirect control over them. Investigating these things will help us to close some of the responsibility gaps and retribution gaps that many authors worry about.

3 | PART III: CRASH-AVOIDANCE STRATEGIES

3.1 | A duty to switch over to self-driving cars?

We have been discussing how to deal with potential crashes involving self-driving cars. But the promise of self-driving cars is supposed to be that they will help us, as much as possible, to avoid crashes. Let us therefore look at the bigger picture and consider arguments in favor of developing and introducing self-driving cars related to their safety potential.

Let us first return briefly to Hevelke and Nida-Rümelin's discussion. When they argue that we should not hold car manufacturers responsible for crashes because of the importance of developing self-driving cars, their reasoning suggests the following argument. There is a moral imperative to seek ways of making traffic with automobiles safer; self-driving cars promise to be much safer than conventional cars; and so therefore there is a moral imperative to develop and introduce self-driving cars into society (Hevelke & Nida-Rümelin, 2015). Similarly, Elon Musk of Tesla – in an interview about Tesla's "autopilot" feature – predicted and seemed to approve of the idea that, in the future, conventional cars will be forbidden, and only automated cars will be permitted (Bloomberg, 2014). The implied argument in that interview was that the safety potential of self-driving cars creates an imperative for people to switch over to such cars and to stop using conventional cars altogether. Robert Sparrow and Mark Howard make virtually the same argument. In their version, self-driving cars should be forbidden so long as they are less safe than regular cars. But once self-driving cars become safer than regular cars, it is regular cars that should be forbidden (Sparrow & Howard, 2017). Moreover, the same general type of argument has also been made within the domain of public health research, by Janet Fleetwood. She writes that automated driving is an "incredible invention" likely to "transform transportation" while "saving lives." For this reason, public health leaders "should welcome" this development (Fleetwood, 2017, p. 536).

I myself – together with Jilles Smids – have put forward a related argument about the situation in which there are both self-driving cars and conventional cars available. This type of "mixed traffic" is perhaps the most realistic situation to anticipate for the coming years ahead. On the supposition that self-driving cars would actually be safer than conventional cars, our suggestion was as follows: People have a duty to either switch over to the safer alternative, namely, autonomous cars, or to use or accept added safety precautions when using the less safe alternative, namely, conventional cars (Nyholm & Smids, forthcoming).

This type of reasoning could help to reopen otherwise mostly dormant debates about safety technologies such as speed limiters and alcohol locks in regular cars. There are contributions to the applied ethics literature that present moral arguments in favor of such technologies, such as papers by Jilles Smids (Smids, 2018) and by Kalle Grill and Jessica Nihlén-Fahlquist (Grill & Nihlén Fahlquist, 2012). But on the whole, such technologies – or ways of mitigating traffic-risks more generally – have not received a great deal of attention within practical ethics. However, the introduction of self-driving cars might put some pressure on people to either try to make their conventional cars safer or switch over to self-driving cars instead. Technologies that could help to make their conventional cars safer would precisely be things such as speed limiters and alcohol locks.

This choice between automated cars and conventional cars is another thing that can be related back to the comparison between self-driving cars and military robots. Military robots have also sometimes been lauded for their supposed life-saving potential. It is thought that they can save the lives of soldiers whose lives are at risk on the battlefield and potentially also the lives of some innocent non-combatants. Military robots have also been said to have a potential to better conform to ethics and the law than human soldiers sometimes do. In particular, Ronald Arkin has argued that if military robots can be programmed to follow the laws of war and rules of engagement in a much stricter way than human soldiers do, armies will have a better chance of preventing war crimes. That, Arkin argues, is a good reason to switch over from deploying human soldiers to instead deploying more reliably law-abiding military robots (Arkin, 2010).

How good is this particular analogy between self-driving cars and military robots? What further ethical issues might reflecting more on this apparent analogy help to bring up? In considering this, Purves and colleagues discuss an objection to military robots that they worry would potentially apply to self-driving cars as well. But, ultimately, they go on to argue that it does not apply to self-driving cars in the same way as it does to military robots. So what is the objection?

Purves et al. argue that it is an inherently problematic feature of military robots that they are programmed to perform targeted killings of human beings. In general, the idea of a machine preprogrammed to kill human beings is a disturbing and unacceptable idea to many people. The worry here about a possible analogy back to self-driving cars is that they too might need to be programmed to kill humans under certain circumstances – namely, in the sense that was discussed in the first of these two articles (Purves et al., 2015).

That is, self-driving cars might be programmed to crash in unavoidable accident scenarios in ways that would target certain people (e.g., the smaller number of people in a "trolley" scenario where the car can either crash into a bigger or smaller number of people). Would the fact that a self-driving car would be pre-programmed to kill people under certain circumstances undermine the argument in their favor based on the prediction that overall they will kill far fewer people than regular cars do?[4]

Purves and colleagues agree that self-driving cars would indeed need to be programmed to kill under certain circumstances. But they think that this is not bad and immoral in the same way that it is bad and immoral, in their view, that military robots are programmed to kill. The argument of Purves et al. for this standpoint is reminiscent of the so-called doctrine of double effect. They argue that whereas military robots are problematic because their primary goal is to kill people, self-driving cars are not problematic in the same way. Self-driving cars' main goal is not to crash into people. It is rather to allow people to drive in a safer way than they are able to do using regular cars. That self-driving cars may need to be pre-programmed to crash into people in rare circumstances is a foreseen and acknowledged bad side effect of self-driving cars. But it is not the primary function they are meant to perform. Therefore, one can coherently oppose military robots even if they on the whole would make wars safer even as one is in favor of self-driving cars even though they too would be pre-programmed to kill human beings under certain circumstances (Purves et al., 2015).

How good is this argument? As I noted, the argument seems to have a lot in common with the doctrine of double effect. It is a general moral principle according to which causing harm as an intended side effect is more permissible than causing harm as a primary goal. This so-called doctrine of double effect does to a large extent seem to fit with common sense, as David Edmonds argues in his book about the trolley problem (Edmonds, 2013). That being said, it is also worth noticing that many moral philosophers have presented general objections to the doctrine of double effect (e.g., Scanlon, 2008). Whether we should accept this doctrine falls outside of the scope of this discussion.

3.2 | Responsible human–robot coordination within mixed traffic

The arguments discussed at the beginning of the section above are looking ahead to a potential future (20, 30 years from now?) when self-driving cars have proven themselves to be much more safe than human-driven cars. As the last thing we will do, let us return to the present once again and the ethical issues we are already facing when it comes to automation in cars.

As was mentioned in the introduction of the first of these two articles, there have already been a number of crashes involving self-driving cars and conventional cars within mixed traffic. Let us briefly reflect on what we might do to avoid these crashes and what ethical choices might be involved in choosing safety strategies.

Traffic psychologists such as Roald van Loon and Marieke Martens and traffic researchers such as Brandon Schoettle and Michael Sivak have analyzed these crashes in mixed traffic (Schoettle & Sivak, 2015; van Loon & Martens, 2015). Their main finding is that there is a noteworthy clash between the ways in which self-driving cars function and the ways in which humans typically drive. Self-driving cars have optimizing driving styles and are very strict rule followers. Human drivers, in contrast, are typically satisficers: They drive just well enough to satisfy their driving goals. And humans are more flexible in their attitudes to traffic rules. An additional complicating factor is that in the coming years, there will be cars with different levels and different types of automation sharing our roads. This may lead to a further variety in driving styles that need to be coordinated with each other (Yang, Gao, & Li, 2016).

When it started coming out that self-driving cars and regular cars crash into each other on account of their different driving styles, some people suggested that the fault lies with self-driving cars. Their "fatal flaw," one commentator claimed, was that self-driving cars are too strict in their rule-following. This causes humans to crash into self-driving cars (Naughton, 2015). To make self-driving cars better coordinated with human-driven cars, self-driving cars should be made to drive more like human drivers do. They should be programmed to sometimes speed, drive "aggressively," and, in other ways, behave in the non-optimal ways that humans behave on our roads (Naughton, 2015; see also Gerdes & Thornton, 2016)

That, however, is a slightly paradoxical or self-defeating suggestion. It might undo some of the main motivations for introducing self-driving cars in the first place. After all, we are looking for a more optimal and more law-abiding type of driving that can help to make our roads safer and automotive traffic more optimal in various other ways as well.

The question is as follows: Should we try to adjust the optimized robotic driving of self-driving cars to human driving and thereby make automated driving less optimal? Or should we rather seek means – including technological means, stricter traffic rules, or whatever – for adjusting human driving to the more optimal driving of automated cars? Either option can be seen as involving costs from an ethical point of view. The former might seem to give up on some of the extra safety benefits we are hoping to realize with self-driving cars. It can also be seen as letting the wrong type of driving (i.e., unsafe and rule-bending human driving) set the standard. The latter – namely, seeking means for making people drive more like robots – might seem to constrain people's freedom and their personal autonomy on the road. There is a need for more ethical debate about how to handle these sorts of problems with human–robot coordination in this domain, as well as in other domains into which robots are increasingly being introduced (Nyholm & Smids, forthcoming).

That we face these sorts of choices illustrates that the ethics of automated driving is not only about how self-driving cars should crash or about who should be held responsible for crashes. The more general issue of risk management in choosing the driving styles of self-driving cars and otherwise optimizing their driving also raises significant ethical issues.[5]

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] This applies to the legal literature published in English. There has also been articles by legal scholars publishing in other languages – for example, German – which have focused on other legal systems.

[2] Danaher himself is ambivalent about retributive blame. He suggests that the creation of retribution gaps might afford an opportunity to move away from a retributive culture (Danaher, 2016).

[3] In a response to my work, Roos de Jong argues that analyzing responsibility for automated technologies in terms of collaborative agency may in some cases give rise to a new version of the so-called "problem of many hands" (De Jong, forthcoming).

[4] A related worry, from Hin-Yan Liu, is that if all cars are programmed to crash in the same ways, this might have aggregated effects whereby certain groups of people become targeted more often than others (Liu, 2017).

[5] Noah Goodall makes this same point when he discusses why "routine driving requires ethics." Goodall argues that the following are all examples of choices made in routine driving of regular cars that are ethically significant: choosing the right following distance; braking strategies; lateral positioning within a lane; and the issue of to what extent we should permit violations of the law (Goodall, under review). One reason why these choices have an ethical component is that they involve balancing different ethically relevant objectives (such as safety, mobility, and legality). Another reason is that these choices also involve distributing risks among different individuals. Risk-management is an inherently ethical matter, involving aspects such as protecting people's safety and promoting equality and fairness. If ordinary routine driving involves making these types of ethically loaded choices, then so does programming self-driving cars for how they should deal with these aspects of routine driving (Goodall, under review).

ORCID

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

WORKS CITED

Arkin, R. (2010). The case for ethical autonomy in unmanned systems. *Journal of Military Ethics*, *9*(4), 332–341.

Atiyeh, C. (2015). Volvo will take responsibility if its self-driving cars crash, *Car and Driver* October 8, 2015. Retrieved from https://blog.caranddriver.com/volvo-will-take-responsibility-if-its-self-driving-cars-crash/ (Accessed January 31, 2018)

Beiker, S. (2012). Legal aspects of autonomous driving. *Santa Clara Law Review*, *52*(4), 1145–1156.

Bloomberg. (2014). Elon Musk on Tesla's auto pilot and legal liability. Retrieved from https://www.youtube.com/watch?v=60-b09XsyqU (Accessed January 31, 2018)

Brey, P. (2013). From moral agents to moral factors: The structural ethics approach. In P. Kroes, & P.-P. Verbeer (Eds.), *The moral status of artifacts* (pp. 125–142). Springer.

Brożek, B., & Jakubiec, M. (2017). On the legal responsibility of autonomous machines. *Artificial Intelligence and Law*, *25*(3), 293–304.

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

Danaher, J. (2016). Robots, law and the retribution-gap. *Ethics and Information Technology*, *18*(4), 299–309.

Darwall, S. (2006). *The second-person standpoint*. Cambridge, MA: Harvard University Press.

De Jong, R. (forthcoming). The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to Nyholm. *Science and Engineering Ethics*.

Edmonds, D. (2013). *Would you kill the fat man? The trolley problem and what your answer tells us about right and wrong*. Princeton, New Jersey: Princeton University Press.

Fleetwood, J. (2017). Public health, ethics, and autonomous vehicles. *American Journal of Public Health*, *107*(4), 532–537.

Gerdes, J. C., & Thornton, S. M. (2016). Implementable ethics for autonomous vehicles. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous driving* (pp. 87–102). Berlin: Springer.

Goodall, N. J. (under review). More than Trolleys: Plausible, ethically ambiguous scenarios likely to be encountered by automated vehicles.

Grill, K., & Nihlén Fahlquist, J. (2012). Responsibility, paternalism and alcohol interlocks. *Public Health Ethics*, *5*(2), 116–127.

Gurney, J. K. (2013). Sue my car not me: Products liability and accidents involving autonomous vehicles. *Journal of Law, Technology & Policy*, *2*, 247–277.

Gurney, J. K. (2015). Driving into the unknown: Examining the crossroads of criminal law and autonomous vehicles. *Wake Forest Journal of Law and Policy*, *5*(2), 393–442.

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.

Hage, J. (2017). Theoretical foundations for the responsibility of autonomous agents. *Artificial Intelligence and Law*, *25*(3), 255–271.

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

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.

Liu, H.-Y. (2017). Irresponsibilities, inequalities and injustice for autonomous vehicles. *Ethics and Information Technology*, *19*(3), 193–207.

Loh, W., & Loh, J. (2017). Autonomy and responsibility in hybrid systems. In P. Lin, et al. (Eds.), *Robot ethics 2.0* (pp. 35–50). New York, NY: Oxford University Press.

Marchant, G., & Lindor, R. (2012). The coming collision between autonomous cars and the liability system. *Santa Clara Legal Review*, *52*(4), 1321–1340.

Maric, P. (2017). Audi to take full responsibility in event of autonomous vehicle crash, *Car Advice* September 11, 2017. Retrieved from http://www.caradvice.com.au/582380/audi-to-take-full-responsibility-in-event-of-autonomous-vehicle-crash/ (Accessed January 31, 2018)

Mindell, D. (2015). *Our robots, ourselves: Robotics and the myths of autonomy*. New York: Viking.

Misselhorn, C. (2015). Collective agency and cooperation in natural and artificial systems. In C. Misselhorn (Ed.), *Collective Agency and Cooperation in Natural and Artificial Systems* (pp. 3–25). Dordrecht: Springer.

Naughton, K. (2015). Humans are slamming into driverless cars and exposing a key flaw: Bloomberg business. Retrieved February 23, 2016, from http://www.bloomberg.com/news/articles/2015-12-18/humans-are-slamming-into-driverless-cars-and-exposing-a-key-flaw

Neuhäuser, C. (2015). Some skeptical remarks regarding robot responsibility and a way forward. In C. Misselhorn (Ed.), *Collective Agency and Cooperation in Natural and Artificial Systems* (pp. 131–146). Dordrecht: Springer.

Nyholm, S. (forthcoming). Attributing agency to automated systems: Reflections on human—robot collaboration and responsibility-loci. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-017-9943-x

Nyholm, S., & Smids, J. (forthcoming). Automated cars meet human drivers: Responsible human—robot coordination and the ethics of mixed traffic. *Ethics and Information Technology*.

Peterson, R. W. (2012). New technology—old law: Autonomous vehicles and California 's insurance framework. *Santa Clara Law Review*, *52*, 101–153.

Pettit, P. (1996). Freedom as antipower. *Ethics*, *106*(3), 576–604.

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice*, *18*(4), 851–872.

Ravid, O. (2014). Don't sue me, I was just lawfully texting and drunk when my autonomous car crashed into you. *Southwest Law Review*, *44*(1), 175–207.

Scanlon, T. M. (2008). *Moral dimensions*. Cambridge, MA: 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.

Smids, J. (2018). The moral case for intelligent speed adaptation. *Journal of Applied Philosophy*, *35*, 205–221. https://doi.org/10.1111/japp.12168

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Sparrow, R., & Howard, M. (2017). When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. *Transportation Research Part C*, *80*, 206–215.

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

US Department of Defense Science Board. (2012). The role of autonomy in DoD systems. Retrieved from https://fas.org/irp/agency/dod/dsb/autonomy.pdf. (Accessed January 31, 2018)

van Loon, R. J., & Martens, M. H. (2015). Automated driving and its effect on the safety ecosystem: How do compatibility issues affect the transition period? *Procedia Manufacturing*, *3*, 3280–3285.

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

Yang, Q., Gao, Y., & Li, Y. (2016). Suppose future traffic accidents based on development of self-driving vehicles. In S. Long, & B. S. Dhillon (Eds.), *Man-machine-environment system engineering, lecture notes in electrical engineering*. New York: Springer.

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, II. *Philosophy Compass*. 2018;13:e12506. https://doi.org/10.1111/phc3.12506