CHAPTER 15

Moral Dilemmas of Self-Driving Cars

Sven Nyholm

Abstract

This chapter provides an overview of some of the most important ethical issues related to autonomous vehicles, also known as self-driving cars. The chapter begins with a discussion of ethical issues related to different levels and kinds of autonomation in cars. It next considers issues having to do with safety precautions, and after that turns to issues related to risks created by self-driving cars. The chapter then discusses the trolley problem, empirical approaches to the ethics of self-driving cars, traditional moral theories, and, lastly, questions related to moral responsibility for harm caused by self-driving cars.

KEYWORDS

self-driving cars, safety, risks, the trolley problem, moral dilemmas, responsibility gaps

Introduction

This chapter provides an overview of some important ethical issues related to autonomous vehicles, also known as self-driving cars. Notably, automation in vehicles comes in different degrees and kinds (Nyholm & Smids, 2020). Cars can be partially or fully automated. It is common practice to distinguish among five different levels of automation, where level zero means no automation and level five means full automation. In cars with some automation (level one to four

cars), drivers are expected to sometimes take over some of the driving tasks. But they are also able to hand over some or perhaps all of the tasks to the car itself, at least in some traffic situations. For the purposes of this chapter, the expression "self-driving car" will refer to any type of vehicle that is either fully automated or could operate in an autonomous mode in at least some traffic situations. Fully autonomous level-five cars may not become available for a very long time. But cars that can operate in autonomous mode in some traffic situations already exist and are on the market.

There are ethical questions related to the differences in levels and types of automation in cars. For example, is it to expect too much of human drivers to require them to sometimes take over the operation of the vehicle, if the artificial intelligence in the car suddenly requests that control be handed back to them? In particular, is it reasonable to expect people to be sufficiently alert at all times, so that they can easily take over if necessary (Hevelke & Nida-Rümelin, 2015)? Some authors argue that it is not reasonable to expect this. So, we should either have fully autonomous self-driving cars or manually driven cars, according to these authors (Sparrow & Howard, 2017).

To this one might potentially respond that cars should never ask or expect human drivers to take back control, but that it should always be a human choice when one wants to drive and when one wants to hand over control to the artificial intelligence in the car. Such questions about handing control over to, and taking control back from, cars are intriguing. In what follows, however, the focus will not be on that topic. The questions below will all be about those times that the vehicle is operating in autonomous mode, whether or not it is possible to hand control back to the human occupant(s) riding in the car.

This is a fairly new topic. Moral philosophers started investigating ethical issues related to self-driving cars around 2014. Back then, the discussion exclusively involved hypothetical thought experiments. What if there was a crash involving a self-driving car, and somebody was injured? Who should then be held responsible? What if the artificial intelligence in the car had to react to a potential crash? What should the self-driving car do? Legal scholars had started thinking about issues involving crashes with self-driving cars a little earlier (e.g., Marchant & Lindor, 2012; Peterson, 2012) But those early articles in legal theory were also mostly about hypothetical scenarios.

In 2015, what had previously been thought experiments started happening in real life (Nyholm & Smids, 2020). That year, there were around twenty small accidents involving experimental self-driving cars. No human beings were seriously harmed, and there were only some minor scratches on some of the cars. What mostly happened was that people in regular cars rear-ended experimental self-driving cars. This typically happened because those self-driving cars were not behaving as the human drivers in the regular cars expected them to.

For example, the self-driving cars accelerated more slowly than most human-driven cars do.

After those early accidents, human drivers were usually blamed. In 2016, however, for the first time, a self-driving car clearly caused an accident (Nyholm, 2018a). On Valentine's day of that year, an experimental self-driving car from Google crashed into a bus. Google had to admit that their car had caused the crash. Later in the same year, something more tragic happened. The first person died while riding in a self-driving car. A Florida man was killed in an accident when his Tesla Model S crashed into a truck while operating in the "autopilot" mode. In 2018, in turn, the first pedestrian was hit and killed by a self-driving car. The artificial intelligence in an experimental car operated by Uber failed to recognize a human being crossing the road in time. The victim was Elaine Herzberg. She was first classified as a sign, then as a bike, and then reclassified as a human being. But by then it was too late. The car hit Herzberg, and she died on the way to the hospital.

In recent years, there has been an absolute explosion of philosophical articles about the ethics of self-driving cars. This chapter will not try to summarize everything that has been discussed in academic philosophy related to self-driving cars. The focus will instead be on some of the key issues that have received the most attention.

The Ethics of Safety and Experiments with Self-Driving Cars

One important reason why many people are excited about the prospect of self-driving cars relates to traffic safety. Eventually, self-driving cars are hoped to become much safer than regular cars (Gurney 2016). So far, though, this has not been proven in practice. This already raises interesting ethical questions.

To improve the safety of self-driving cars in a wide range of real-life traffic scenarios, engineers need to experiment with self-driving cars in actual traffic. This involves imposing risks on people who live in the communities where experimental self-driving cars are being tested. A key question here is how great the risks are that we can justifiably impose on people in this experimental stage to make sure that we later save many lives because very safe self-driving cars have by then been developed. Recall that Elaine Herzberg was killed by an experimental self-driving car. A grim question arises: Should such deaths be tolerated now because of the greater number of lives that might later be saved if the tech industry is permitted to experiment with self-driving cars on public roads?

Robert Sparrow and Mark Howard (2017) present an interesting claim about the ethics of risk and safety in the development of self-driving cars.

They argue that so long as self-driving cars have not been proven to be safer than regular cars, it should be illegal to sell self-driving cars. However, once self-driving cars have been proven to be safer than regular cars, regular cars should be forbidden. That argument seems to implicitly rest on the following general moral principle: if a safer alternative is introduced into some risky domain of life, it is immoral to use older, less safe alternatives. Only the safest alternative should be permitted in a dangerous domain like traffic. Is this right?

What if people who drive older, otherwise less safe cars are willing to use special safety precautions (Nyholm, 2018b; Nyholm & Smids, 2020)? For example, the requirements for getting a driver's license could be made much more stringent. Moreover, manually driven cars could perhaps be equipped with alcohol locks and speed-limiting technologies, which would make it impossible to drive while drunk or to do any dangerous speeding. Could such added safety precautions perhaps offset the greater risks otherwise involved with manually driven cars? Whatever we think about these issues, the following seems to hold true: when or if self-driving cars become safer than regular cars, this will put pressure on those who still wish to drive regular cars to justify why they should still be permitted to do so.

Ethics Programming and the Trolley Problem

As noted above, self-driving cars hold the promise of eventually becoming much safer than regular cars. Yet they cannot be 100% safe. Even the safest self-driving cars will sometimes crash. Anything that is heavy and moving fast, and that could malfunction, like any technology can, will sometimes cause accidents. So, we need to think about accident scenarios involving self-driving cars (Goodall, 2014).

It is sometimes suggested that humans should always take over control in crash scenarios or that cars should simply brake in risky scenarios. However, these responses are problematic. Average human reaction times are slow. It won't always be possible for people to react in time. Moreover, in some situations, it is not possible to simply apply the brakes. And no option open to the car may be safe for everyone involved. So, it seems that automated cars need to be programmed for how to respond to accident scenarios. Coming up with such programming requires thinking about potential choices that impose serious risks on different people. The issue of what self-driving cars should do in situations in which crashes are unavoidable is an inherently ethical issue. Therefore, some philosophers talk about the need to equip self-driving cars with "ethics settings." For example, should the car always try to protect the people riding in the car?

Or should self-driving cars simply try to minimize overall harm when crashes are unavoidable (Goodall, 2014; Nyholm, 2018a).[1]

Imagine the following scenario. A self-driving car with five passengers suddenly detects a large obstacle on the road. Unless the car turns, the five passengers are likely to die. The only way to turn is onto a sidewalk. But a pedestrian walking there. So, the only way the car can save the five is if it sacrifices the one. What should the self-driving car be programmed to do in such a situation? Now consider an alternative scenario: in this case, there is only one person in the automated car. Again, some large obstacle falls onto the road. And again, the car can only save the passenger if it turns onto the sidewalk. This time, however, there are five pedestrians on the sidewalk. What should the car do now? Should it sacrifice the one for the sake of the five?

These examples are designed to sound similar to the so-called trolley problem (Kamm, 2015; Nyholm & Smids, 2016). The trolley problem is a well-known philosophical thought-experiment in which an out-of-control trolley is about to hit five people on train tracks. You are standing next to a switch. If you pull the switch, the trolley will be redirected to a side track, where there is only one person. So, to save the five, one person would have to be sacrificed. In another variation, the only way to save the five on the tracks is to push a large and heavy person off a bridge down onto the train tracks in front of the trolley. The large person's hefty weight will then set off the automatic breaks of the trolley before it hits the five. This would kill the one but save the five. What should be done in these cases? The challenge of explaining and justifying differences in people's intuitions about such cases is what is usually referred to by the phrase "the trolley problem" (Kamm, 2015).

Many articles—both in the mass media and the academic literature—have likened the ethics of self-driving cars to the trolley problem. However, we should be careful not to draw too close of an analogy between the philosophy of the trolley problem and the real-world ethics of crashes involving self-driving cars (Hevelke & Nida-Rümelin, 2015; Nyholm & Smids, 2016). There are at least three reasons why.

Firstly, in academic discussions of the trolley problem, we are asked to concentrate only on a small set of stylized situational considerations. In the real-world ethics of automated driving, in contrast, we should take as many considerations as possible into account. Secondly, in philosophical trolley-problem discussions, we are typically asked to completely set aside questions about legal and moral responsibility. In the real-world ethics of automated driving, we cannot simply set aside questions about responsibility. Thirdly, in trolley-problem discussions, we assume that we know with certainty what the outcomes of different possible actions would be. In the real-world ethics of automated driving, in contrast, we are dealing with risks and uncertainty.

For these reasons, the literature on the so-called trolley problem may be less helpful than many people might think when it comes to the ethics of automated driving. That is not to say that the literature about the trolley problem and the comparison between trolley problem-inspired cases and the ethics of self-driving cars is altogether irrelevant. If nothing else, it can be useful to compare the ethics of crashing self-driving cars with the trolley problem because identifying key differences between the two can be a good way of clarifying what matters most in the real-world ethics of self-driving cars.

Empirical Ethics

There has been some fascinating work about self-driving cars within the field of empirical ethics. Empirical ethics is an attempt to incorporate empirical investigation of ordinary people's intuitive attitudes and judgments into academic ethical analysis. For example, we can systematically study people's attitudes and moral intuitions by letting them make judgements about many different real or simulated scenarios involving crashing self-driving cars. We can then discern patterns in their judgments and intuitions. And we can try to incorporate our findings into ethical arguments.

Several psychologists and behavioral economists have been surveying ordinary people's intuitive opinions about how automated cars should handle crash scenarios. One interesting finding comes from interdisciplinary researchers at MIT (Bonnefon et al., 2016). The finding is that when people are asked about what kinds of accident algorithms they would like others to have, many people say that they want others to have cars programmed to minimize overall harm. However, when asked what kind of accident settings they would like to have in their own self-driving cars, people's responses typically change. They do not want to be required to use or buy cars that are "altruistic" by being programmed to minimize overall harm. Instead, they prefer cars that would be programmed to try to always save the people in the car, even if this does not minimize overall harm.

On the "moral machine" website also created by researchers at MIT, one can explore numerous different dilemmas and cases and make intuitive judgments about them.[2] For example, if a car would hit and kill three senior citizens if it turns left, or hit and kill three children and two cats if it turns right, what should the car do? Or what if a car with five passengers in it can either go straight and crash into a wall, or turn and crash into a pedestrian who is jaywalking when there is a red light? What should the car then do? Those are the kinds of dilemmas people are asked to have intuitions about.

Millions of people around the world have participated in this experiment. The researchers have analyzed widely shared attitudes about these moral

dilemmas involving self-driving cars (Awad et al., 2018). An interesting finding is that depending on where people live in the world, they have slightly different attitudes about whose safety should be prioritized in these imagined crash scenarios. In some parts of the world, participants were more likely to favor saving children at risk than saving older people at risk. In other parts of the world, it was the other way around. Moreover, in some parts of the world, someone breaking the traffic rules (e.g., crossing the street at a red light) was seen as weakening their right to not be hit by a self-driving car. In other parts of the world, that factor did not play any significant role in people's intuitions. There were several other fascinating cultural differences in people's attitudes around the world in these surveys.

Is this survey-based empirical methodology a good basis for ethical theorizing about crash scenarios? These findings are certainly very interesting. But there are some reasons for skepticism (Nyholm, 2018a). Here are three.

First, people do not yet have much real-world experience with traffic involving self-driving cars. It is likely that people's attitudes will change once they acquire more experience of what it is like to have lots of self-driving cars in society. This gives us reason to not put too much weight on people's current attitudes. Second, people's spontaneous gut reactions to hypothetical cases do not necessarily tell us what arguments and reasons they would present to defend their intuitive judgments. In ethical reasoning we evaluate arguments, and not only intuitive responses without any arguments or reasons to back them up. Third, people seem to have inconsistent attitudes. As was noted above, most people want others to have harm-minimizing cars. But they themselves want to have cars programmed to save them. Cars that minimize overall harm will sometimes save the car owner. But sometimes cars programmed to minimize overall harm will have to sacrifice the people in the car.

Again, people's attitudes and intuitions are certainly important and interesting to consider when we think about the ethics of automated driving and accident scenarios. But it is not clear that we can easily move from premises about people's intuitive attitudes to any solid conclusions about how best to argue about self-driving cars and accident scenarios.

Traditional Ethical Theories

We can next briefly explore the option of using traditional ethical theories from moral philosophy when thinking about how self-driving cars should behave. Specifically, let us consider utilitarianism, Kantian ethics, and virtue ethics. Utilitarianism is the theory that we should always promote the overall good, by promoting everyone's well-being. Kantian ethics says that we should adopt a set

of basic principles we would be willing to have as universal laws, so that we treat everyone with equal respect. Virtue ethics tells us that we should live our lives in ways that help us to exercise various virtues and excellences. We can use these theories to explore the question of how self-driving cars should handle accident scenarios as well as how they should behave more generally (Gurney, 2016).

Importantly, these ethical theories were originally developed to be about what humans should do, not about what technologies equipped with artificial intelligence should do. So, it is not obvious that we can simply carry over the moral principles that are supposed to guide human choices to the ethics of how self-driving cars should behave. It might be unclear what principles of translation should be used when we export traditional theories about human–human interaction into the new domain of human–machine interaction. This is a new form of ethics, where different rules and principles might potentially be taken to apply.

Let us nevertheless consider how these theories might be used in this context. Some philosophers will say that we need to make a choice here. We can only use one moral theory. But it is also possible to suggest that in thinking about the ethics of how cars should behave around human beings, we could make use of all three moral theories. There is clearly something to learn from each traditional moral theory.

The lesson from utilitarianism (or consequentialism more generally) might be that however cars are programmed to handle crash scenarios or behave more generally, we should think about this issue with an eye to the greater good of society. We should reason carefully about what promotes overall well-being and other important human values.

The lesson from Kantian ethics could be that whatever rules we decide on regarding the behavior of self-driving cars, these rules should be "universal laws" that are respectful of everyone. For the sake of fairness and equal treatment, people's cars should behave and handle crash scenarios according to a shared set of rules, applying equally to everyone.

Consider lastly virtue ethics. Currently, there are important virtues that many people tend to exhibit in traffic. For example, people tend to conduct themselves in fairly responsible ways. Of course, there are many exceptions. But most people feel responsible and mostly also act responsibly when they use very risky technologies like cars. The philosopher Mark Coeckelbergh (2016) has argued that people's tendencies to feel a sense of responsibility when they use cars is influenced by the design and technology of the car. This is relevant from a virtue-ethical perspective. After all, behaving responsibly is an important virtue. So, it can be argued that self-driving cars should be designed to make people who use such cars still feel responsible for what happens when they are riding in these cars. This is an important virtue.

In general, then, we could use the traditional ethical theories of utilitarianism, Kantian ethics, and virtue ethics to argue for the following general moral principle regarding how self-driving cars should behave in society: self-driving cars should be made to behave in a way that promotes everyone's well-being, according to principles that apply equally to all and that are respectful toward all, and that help to promote human virtue. That is a very general moral principle, and there may be lots of disagreement about what this would mean in practice. But it provides general guidelines for ethical thinking about the behavior of self-driving cars that seems highly plausible.

Moral Responsibility

In the introduction above, it was mentioned that the ethics of self-driving cars has gone from using hypothetical thought experiments to being about real-world events. When this development was mentioned above, there was also a brief mention of the issue of who should be held responsible when there are accidents involving self-driving cars. This has been a key issue in the real-world ethics of automated driving. In this last section, let us therefore briefly consider some issues related to responsibility and self-driving cars.

Starting with the real-world cases mentioned earlier, Google has usually denied responsibility whenever their experimental self-driving cars have been involved in crashes. However, as was mentioned above, there was one case—the case on Valentine's Day of 2016—in which Google admitted that a crash had been caused by their car. The Google car had crashed into a bus. Google admitted "partial responsibility." They also promised to update the software of their car, so that it would become better at predicting the behavior of buses (Nyholm, 2018b).

In contrast, when a man died in a Tesla Model S car operating in "autopilot" mode later that same year, Tesla denied all responsibility. They published a blog post expressing sympathy with the family of the deceased. But the company noted that it was part of their user agreement that users of their "autopilot" function must take responsibility for whatever problems might arise. At the same time, however, Tesla also said that they would update their hardware, so that their cars would be better able to detect dangerous obstacles.

To some people, Google's abovementioned response made more sense than Tesla's. Google assumed partial responsibility. So, it made sense that they would also say that they were going to update their software. This contrasts in an interesting way with Tesla's response. Tesla admitted that it would be a good idea to update their hardware. Was that not an admission of responsibility for the crash? If Tesla was not to blame for the crash, then what need was there to update the technology in their car?

Some commentators find it unfair to blame human users of self-driving cars for accidents that their cars cause (Hevelke & Nida-Rümelin, 2015). Users of self-driving cars who are lucky because their cars do not crash may not do anything differently than users of self-driving cars who are unlucky because their cars do cause accidents.

Why not always simply blame the companies who create the self-driving cars? Some scholars who discuss this issue have worried that this might make car companies less motivated to create these cars (Marchant & Lindor 2012; Hevelke & Nida-Rümelin 2015). That would be a bad development, it has been suggested, since self-driving cars are thought to potentially have many benefits, particularly related to traffic safety.

Another reason that is sometimes suggested for why car designers should not be responsible for crashes involving self-driving cars is that they will not be able to reliably predict what these cars will do once they are on the road. Once the cars are out in traffic and operating autonomously, the people who built the cars will no longer directly control what the cars are doing. After all, they are self-driving cars. They are supposed to be operating autonomously. And no human might be able to fully predict what the artificial intelligence in the car will decide is the best course of action in certain traffic situations (Hevelke & Nida-Rümelin, 2015).

Some philosophers who discuss issues like these worry that self-driving cars might give rise to "responsibility gaps" (Nyholm, 2018b). This would mean that there is nobody who can be sensibly blamed when self-driving cars crash and people are harmed, even though it might seem as if somebody should be held responsible.

Do these worries about possible responsibility gaps make sense? Perhaps some traditional ways of thinking about responsibility for crashes involving human-driven cars cannot be directly carried over to the new case of crashes involving self-driving cars. However, there are ways in which one can understand moral and legal responsibility that can be brought to bear on the issue of responsibility for crashes involving self-driving cars. For example, we can think in terms of what roles and rights people have. This can ground responsibilities. There are also other possible arguments. Among other things, one can think about who benefits most from the presence of self-driving cars on the road. If a car company rents out self-driving cars, and that company makes a lot of money, then it might only be fair that they should be held responsible for any accidents that may occur involving their lucrative self-driving cars.

Moreover, even if people may lack direct control over how self-driving cars behave on the road, they will still have indirect control over the behavior of self-driving cars. Self-driving cars will be updated and maintained. And updates and maintenance will be based on people's opinions about how self-driving cars

should function (Nyholm, 2018b). This will help to make anybody in charge of updating and maintaining these cars at least partly responsible for how the cars behave on the road. This will give people at least indirect control over what self-driving cars do. That might be enough to make them responsible for the behavior of the cars.

References

Awad, E., Dsouza, S., Kim, R., Shulz, J., Henrich, J., Shariff, A., Bonnefon, J-F., & Rahwan, I. (2018): The moral machine experiment. *Nature*, *563*, 59–64.

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

Coeckelberg, M. (2016). Responsibility and the moral phenomenology of using self-driving cars. *Applied Artificial Intelligence*, *30*(8), 748–757 https://www.tandfonline.com/doi/full/10.1080/08839514.2016.1229759

Goodall, N. J. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board*, *2424*(1), 58–65.

Gurney, J. (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.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics*, *21*(3), 619–630.

Kamm, F. (2015). *The trolley problem mysteries*. 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.

Nyholm, S. (2018a). The ethics of crashes with self-driving cars, a roadmap I. *Philosophy Compass*, *13*(7): e12507. https://onlinelibrary.wiley.com/doi/full/10.1111/phc3.12507

Nyholm, S. (2018b). The ethics of crashes with self-driving cars, a Roadmap II. *Philosophy Compass*, *13*(7): e12506. https://onlinelibrary.wiley.com/doi/full/10.1111/phc3.12506

Nyholm, S., & Smids, J. (2020). Automated cars meet human drivers: Responsible human-robot coordination and the ethics of mixed traffic. *Ethics and Information Technology*, *22*(4): 335–344.

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.

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

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.