5 Morality and the Machine

5.1 The Uber Autonomous Car Accident in Arizona

In March, 2018 a Volvo XC90 SUV – one of Uber's self-driving cars[1] – hit and killed a forty-nine-year-old pedestrian who was walking a bicycle across the street in Tempe, Arizona (US). This was not the first fatal accident with a self-driving car: there had already been several cases of fatal accidents with Tesla cars driving on autopilot.[2] The Uber accident caused a lot of commotion nonetheless, as it was the first time a self-driving car had caused the death of a pedestrian.

A key question in such accidents is: Who is responsible? Assigning responsibility in complex engineering designs is generally a complicated matter: to put it in terms of the "problem of many hands" discussed in Chapter 1, there are so many "hands" involved in the design of complex technologies that it is difficult to say precisely whose hands caused an accident. This problem is even more complicated when the "pair of hands" involved is that of a machine. This is a typical situation arising from a joint decision between human and machine, raising the questions: Can we ascribe any form of responsibility to the car,[3] or does the responsibility lie solely with the car designer or manufacturer? Who else might have contributed to the accident?

Let us discuss the question of responsibility by reviewing the Uber accident while focusing on the various actors involved in the development of the Uber car and the implementation of the experiment that led to this fatal

[1] In this chapter, I sometimes refer to self-driving cars as "autonomous cars" because this is the term used most often in the literature and the media, but almost all cars that are described as autonomous are in fact semiautonomous.

[2] Stewart, "Tesla's Self-Driving Autopilot Involved in Another Deadly Crash."

[3] Matthias, "The Responsibility Gap."

accident.[4] In so doing, we need to also consider the findings of the National Transportation Safety Board (NTSB), both the preliminary findings of May, 2018 and those of the extensive final report released in November, 2019.[5] The NTSB is formally in charge of the accident investigation.

In the literature on responsibility, various notions have been used to pinpoint particular categories of responsibility, including passive and active responsibility, accountability, blameworthiness, and liability. Ibo van de Poel and Lambèrt Royakkers have suggested clear demarcations of these concepts that are helpful for the purposes of this chapter.[6] Passive responsibility is a backward-looking responsibility, relevant for consideration after an accident. Within passive responsibility, there is a distinction between accountability – concerning who can be held to account for the harm to the victims or to society at large – and blameworthiness, which concerns the question of who is to blame.[7] Liability is a legal passive responsibility, according to the law, often "related to the obligation to pay a fine or repair or repay damage."[8] Active responsibility involves proactive thinking about responsibilities and situations when technologies are to be used in the future. In reviewing the accident, we will be focusing on passive responsibility for this specific accident, but the purpose of this chapter (and also of this book) is to use the Uber case study to illustrate how to designing of *morality into the machine*. So our primary interest is in what Van de Poel and Royakkers call proactive and active responsibility.

5.1.1 The Human Driver

In earlier accidents with self-driving cars, it was often claimed that the bulk of responsibility lies with the human driver, especially in discussions of legal responsibility or liability.[9] That is, the so-called autonomous systems are only *semiautonomous*, which means they are meant to assist but not replace

[4] Jack Stilgoe's recent book discusses the case study of the Uber accident in great detail; see Stilgoe, *Who's Driving Innovation?*

[5] Hawkins, "Serious Safety Lapses Led to Uber's Fatal Self-Driving Crash, New Documents Suggest."

[6] Van de Poel and Royakkers, *Ethics, Technology and Engineering*.

[7] Ibid., 10–13.

[8] Ibid., 258.

[9] For instance, a legal investigation of a 2016 Tesla accident found the driver of the car responsible.

the driver. An assisting system is therefore considered to be correctly used only when there is human supervision. Tesla has manufactured fairly advanced “auto pilot systems,” and its website explicitly states that “while using Autopilot, it is your responsibility to stay alert, keep your hands on the steering wheel at all times and maintain control of your car.”[10]

While there have been test drives with driverless autonomous cars, these have been in restricted areas using predetermined routes. The consensus seems to be that autonomous or semiautonomous cars cannot yet operate safely without a human supervisory driver. That is, their supporting software needs to get used to actual roads, with all their real-world difficulties. The cars, therefore, need to be “trained” in real-life situations, and particularly in responding to unanticipated events; for this, a supervisory driver is indispensable and takes over from the autonomous system when necessary.

When instructing and testing the car, it is thus the driver’s responsibility to remain vigilant at all times, continuously paying attention to the road while the car drives itself. This is a fundamental difficulty with self-driving and driver-assisted systems. On the one hand, autonomous systems are often introduced to replace the *fallible human driver*, assuming that a well-programmed system will make fewer mistakes. On the other hand, training these autonomous systems depends on those same fallible humans.[11] What further complicates this is that humans are good drivers when they pay continuous attention, but remaining vigilant is exactly what we do not excel at, given that our minds are easily distracted and exhausted, especially during repetitive tasks. Experiments show that we are least vigilant when we have to continually repeat a task that our minds perceive as boring. For instance, if we have to watch an automatic car driving along the road without needing to regularly intervene, our minds will potentially become numbed if we do not add some form of stimulus.[12] “Your brain doesn’t like to sit idle. It is painful,” according to Missy Cummings, director of the Humans and Autonomy Laboratory and Duke Robotics at Duke

[10] Quoted from the website of the manufacturer: www.tesla.com/support/autopilot (consulted August 22, 2019).

[11] Davies, “The Unavoidable Folly of Making Humans Train Self-Driving Cars.”

[12] Heikoop et al., “Human Behaviour with Automated Driving Systems”; Calvert et al., “Gaps in the Control of Automated Vehicles on Roads.”

University.[13] There have been many examples of Tesla drivers sleeping at the wheel while the car drove itself; fortunately, they didn't lead to any accidents.

So we have become complacent in trusting an automatic system to run smoothly with no need for us to intervene. This complacency – to which I will return in the coming sections – was a factor in the Uber accident. In a video from the car's dashboard camera, it became apparent that the safety driver, who was supposed to oversee the driving at all times, was looking down at their lap. While the safety driver at first claimed that they were looking at the autonomous system's interface, which is built into the car's center console, a police investigation revealed that an episode of the television show *The Voice* was being streamed on their cell phone at the time of the accident.[14]

Uber denied sole responsibility for the accident, implying that it was at least partly the safety driver's fault for not paying attention. If the driver is paid to sit at the wheel and pay attention to the road, this is what they are expected to do. However, this does not answer the question of whether the driver, in full mental capacity and paying full attention, would have been able to identify the pedestrian and stop, or at least mitigate the consequences – for instance, by swerving or braking. It seems indisputable that the driver was at least partly responsible for the accident, but was perhaps not the only responsible party. Uber also acknowledged that the software hadfaults.

5.1.2 The Software Designer

When software is designed for automatic vehicles, an important question is how to deal with "false positives" – that is, situations in which an object is detected but, because it is neither a danger to anyone outside the car nor likely to damage the car, does not necessitate stopping the car. Examples of such objects are a plastic bag or a piece of newspaper floating in the air, or an empty can in the path of the car. The software needs to be fine-tuned to decide when to respond to false positives. Uber's car was tuned in such a way

[13] Quoted from Davies, "The Unavoidable Folly of Making Humans Train Self-Driving Cars."

[14] Ibid.

that it would not brake overcautiously or stop abruptly, which can be a source of discomfort: “emergency braking manoeuvres are not enabled while the vehicle is under computer control, to reduce the potential for erratic vehicle behaviour.”[15] However, as Uber executives admitted, the fine-tuning of the software to ignore false positives had probably gone too far, because in the case of this accident the car treated a pedestrian pushing a bicycle as a false positive.

The NTSB’s preliminary report revealed that about six seconds before the crash, the car’s sensors detected the pedestrian first as an unknown object, then as a vehicle, and finally as a pedestrian.[16] By the time the car detected that it needed to stop, less than two seconds remained, requiring abrupt braking, which by default was not possible with this software, as explained above. Thus, intervention by the driver was the only option; this might not have been sufficient to prevent the fatal accident but could potentially have reduced the impact on the pedestrian by slowing down or swerving. However, this in turn could have given rise to another accident, perhaps an impact with another car coming from the opposite direction.

Let us review for a moment how other manufacturers deal with false positives. Waymo, a company that emerged from Google’s self-driving car project and is considered one of Uber’s biggest competitors for building self-driving cars, has similar software in use. However, Waymo’s software has greater sensitivity to all objects on the road, meaning that its cars will probably brake as a result of false positives more often than Uber’s. Frequent braking omeans that “the ride can be jerky, with sudden pumping of the brakes even though there’s no threat in sight.”[17] But the flip side of a car stopping more often than is strictly necessary is that safety for pedestrians and other traffic is likely to increase. After the Uber accident, the CEO of Waymo confidently announced that his company’s self-driving cars could have avoided it.[18] Of course, it is convenient to make such a statement in hindsight, but the jerky rides and constant braking of prototype Waymo cars do add to its credibility. While Waymo considers safety to mean putting up with a bumpier ride to avoid accidents, Uber considers abrupt stopping to

[15] From the report of the NTSB, quoted from ibid. [16] Ibid.

[17] Efrati, “Uber Finds Deadly Accident Likely Caused by Software Set to Ignore Objects on Road.”

[18] Ohnsman, “Waymo CEO on Uber Crash.”

also reduce safety for both driver and other car occupants.[19] This was the reason why Uber had intentionally disabled the SUV's factory-set automatic emergency braking system in order to prevent this erratic – and thus uncomfortable and unsafe – behavior of the car in autonomous mode.[20]

If we look at this issue in terms of the vocabulary of Design for Values (see Chapter 4), we might argue that the value of safety has been operationalized differently by different manufacturers of self-driving cars. The two interpretations of safety mentioned above are dramatically different. Waymo tolerates the discomfort of more frequent braking, aiming to ensure the safety of pedestrians and cyclists, while Uber puts more emphasis on the safety of the driver, which can also be at stake if the car stops too frequently or abruptly. Uber also puts a strong emphasis on driving comfort by eliminating false positives as much as possible, producing a smooth ride, though this meant increasing the risk of overlooking a pedestrian as a false positive.

5.1.3 The Car Designer: There Was No "Autopilot Nag"

As discussed above, it is commonly known that we become complacent when a task we are carrying out is repetitive and monotonous. Thus a driver can lose attention if intervention, which would keep the mind sharp, is not required for long periods of time. This is paradoxical because on the one hand, manufacturers intend cars to become as autonomous as possible (hence requiring the least intervention from the driver), yet on the other hand, the less the driver's attention is required, the less likely they will be to pay attention at crucial moments, as in the case of the Uber accident.

This complacency issue is problematic for a number of different autonomous systems. The most prominent example is probably in highly automated aircraft cockpits, where most decisions are computerized and pilots are mainly responsible for supervision and oversight. While this division of labor leads to satisfactory results, there are situations in which the crew "must make split-second decisions on how to proceed, . . . sometimes putting the aircraft and its passengers in dangerous situations," for instance when

[19] Efrati, "Uber Finds Deadly Accident Likely Caused by Software Set to Ignore Objects on Road."

[20] Hawkins, "Serious Safety Lapses Led to Uber's Fatal Self-Driving Crash, New Documents Suggest."

the sensors of an airplane fail because of unexpectedly bad weather conditions.[21] This happened with Air France Flight 447 in May 2009, when extreme cold caused the "pilot tubes" to be obstructed with ice crystals so that they gave an incorrect speed reading. When the crew discovered the malfunction, it was too late to avoid crashing into the Atlantic Ocean, causing the deaths of all 228 people on board.[22] Therefore, it is necessary to ensure that pilots are constantly *aware of the situation* in order to respond properly and promptly. In other words, "in situations where sensors and automation … do not function properly, it is crucial that pilots have, or quickly regain, a good awareness and understanding of the situation at hand."[23]

A similar situation awareness also needs to be designed into the autonomous cars that partly depend on the alertness of the supervisory driver. Manufacturers have created various types of warning systems to remind and sometimes force human drivers to pay attention. Some of these fall into the category of persuasive technologies, which are technologies that persuade the user to do "the right thing"; these have been discussed in Chapter 4. For instance, Cadillac's Super Cruise has an infrared camera installed on the steering column that monitors the driver's head position and reminds them to keep their eyes on the road if they happen to look away from the road for too long.[24] Tesla's automatic pilot has built-in sensors in the steering wheel to detect whether pressure is being applied on the wheel, a sign that the driver is awake. If the driver takes their hands off the wheel for too long, they first get a "visual warning" that, if ignored, will turn into an audible warning, or beep. There are thus different degrees of persuasion built into the system, which is popularly known as "autopilot nag."[25] However, Tesla has developed this technology far beyond mere persuasion. In the most recent updates, the car refuses to continue driving if the warning

[21] Mulder, Borst, and van Paassen, "Improving Operator Situation Awareness through Ecological Interfaces," 20.

[22] See here the interim report of the accident: www.bea.aero/docspa/2009/f-cp090601e2.en/pdf/f-cp090601e2.en.pdf.

[23] Mulder, Borst, and van Paassen, "Improving Operator Situation Awareness through Ecological Interfaces," 21.

[24] Davies, "The Unavoidable Folly of Making Humans Train Self-Driving Cars."

[25] Lambert, "Tesla's Latest Autopilot Update Comes with More 'Nag' to Make Sure Drivers Keep Their Hands on the Wheel."

signal “Hands on the wheel” is ignored several times in a row: It automatically slows down, turns on the emergency lights, and eventually pulls over and shuts itself off.[26] Thus, the car autonomously decides to stop when it detects a situation that is unsafe not only for the driver but also for other traffic participants on the road.

The Uber car did not have any such system to warn the driver to pay attention. Uber did, however, claim to have instructed and trained its drivers, reminding them that they should pay attention at all times. Looking at cell phones was explicitly forbidden. What does this imply for the ascription of responsibilities after the accident? Would the cell phone clause in the drivers’ contracts also remove or reduce Uber’s moral responsibility for the accident? Here, there is a potential discrepancy between what is prohibited and the driver’s mental capacity to comply with that prohibition. Was the driver offered a detailed technical and nontechnical training to prepare them for this task that human minds are not typically good at? The answer to this question is crucial to the question of responsibility.[27]

5.1.4 The Pedestrian Who Was Jaywalking

As mentioned above, Uber claims that the driver did not pay sufficient attention to the road and that the software was tuned too loosely, failing to correctly identify the pedestrian. One reason for the failurewas that the pedestrian was jaywalking; the software was programmed to pay extra attention to pedestrians only when approaching crosswalks. According to the NTSB’s report, Uber’s software for detecting and classifying other objects “did not include a consideration for jaywalking pedestrians.”[28] On the one hand, Uber was clearly at fault for not making its software more sensitive to pedestrians who might cross the road anywhere; this has also to do with the observation made in Chapter 2 that our risk assessment methods usually assume perfect human behavior. On the other hand, one might argue that the pedestrian was at fault, too. It seems cruel to hold someone responsible for their own death when it was due to a rather minor misdemeanour such

[26] Stewart, “Tesla’s Self-Driving Autopilot Involved in Another Deadly Crash.”

[27] I thank Filippo Santoni de Sio for this important addition.

[28] Quoted here from Hawkins, “Serious Safety Lapses Led to Uber’s Fatal Self-Driving Crash, New Documents Suggest.”

as jaywalking. It is nevertheless interesting to observe that in the minds of many, it did matter that the pedestrian was also at fault. To be sure, I am not claiming that equal shares of responsibility should be assigned to all parties. I only intend to argue that in such a moral dilemma, it may matter that the victim the accident was also at fault. In an empirical study at the Massachusetts Institute of Technology simulating similar accidents with autonomous cars, it has been shown that some participants feel less responsible for the lives of those who do not respect the law than for the lives of those who do.[29]

5.2 Crash Optimization and Programming Morality in the Machine

In the early days of autonomous technologies, there was a prominent discussion about what to do when the use of technology involved moral judgments. Are we supposed to program morality into the machine, and if so how should that be done and who should decide, for instance, how a car should react in an accident? The first writings on the subject compare such situations to a thought experiment in moral philosophy known as the Trolley Problem.[30]

5.2.1 *Autonomous Vehicles and the Trolley Problem*

Since its introduction by Philippa Foot in 1967, the Trolley Problem has been widely adopted in various forms in applied ethics because it helps to pinpoint a crucial ethical dilemma.[31] This is a hypothetical situation in which a runaway trolley is rushing down a track because its brakes do not work. There are five workers working on the track, who do not hear the trolley approaching. If nothing else happens, the trolley will kill all of them. There is an escape route, though: the trolley can be diverted onto a side track, where there is only one worker. Thus, we can reduce the number of casualties from five to one by pulling a lever and diverting the trolley. Would you intervene if you had the chance?

[29] Awad et al., "The Moral Machine Experiment."

[30] Goodall, "Ethical Decision Making during Automated Vehicle Crashes"; Lin, "Why Ethics Matters for Autonomous Cars."

[31] Foot, "The Problem of Abortion and the Doctrine of Double Effect."

The Trolley Problem indicates a hypothetical, simplistic, and highly exaggerated situation, but it does help us to understand the fundamental ethical issue at stake: If you had a chance to save five people at the expense of losing one, would you do so? It helps us to distinguish between consequentialist approaches to ethics, which assess moral rightness in terms of expected consequences, and deontological, duty-based approaches, which primarily focus on the question of whether the rights of one worker can be given up to save five others. This is the Trolley Problem in its starkest form. More sophisticated versions of the Trolley Problem further complicate the moral dilemmas. An alternative Trolley Problem, for instance, has a bridge with a person standing on it in place of the side track. You are standing on the same bridge. By pushing the other person off the bridge, you can achieve the same outcome of stopping the runaway trolley, saving five lives. In terms of the outcome, the question of whether to push the individual off the bridge is the same as the question of whether to pull the lever, but it emphasizes much more clearly that your intervention leads to someone losing their life. The Trolley Problem has countless variations and has been influential in various fields of applied ethics.

The trolley in the Trolley Problem was clearly not designed by an engineer, my engineering students often object. First, engineering systems include redundancies in order to increase safety. For instance, it is likely that the trolley has secondary brakes that will kick in if the primary brakes fail, especially if it has been designed properly. Since the trolley could potentially cause serious harm, there may even be an emergency tertiary brake system in place to prevent accidents, or at least to reduce its speed and thereby the likelihood of harm. Second, in addition to redundancies, engineering systems have features that mitigate the consequences of an unavoidable accident. A heavy trolley that could potentially kill many people would be expected to have various types of warning mechanisms that would give those in the surrounding area a chanceto escape. The third objection is a rather fundamental one and perhaps the most problematic: that is, the types of certainties that the Trolley Problem presumes are irreconcilable with basic notions of technological risk. Because of the inherent uncertainties of real-life situations, we can never be sure about the outcomes, which makes moral decision-making about them even more complex. Later in this section, I will argue that properly understanding technological risks may be a better way of addressing the ethical issues associated with autonomous cars.