Chapter 5
Responsibility and Liability in the Case of AI Systems

This chapter discusses the question of who is responsible in the case of an accident involving a robot or an AI system that results in some form of damage. Assigning liability is challenging because of the complexity of the situation and because of the complexity of the system. We discuss examples of an autonomous vehicle accident and mistargeting by an autonomous weapon in detail to highlight the relationship of the actors and technologies involved.

People are not concerned with liability when everything is going fine and no one suffers harm. This concept comes into play only when something goes wrong. Some damage is done, people get hurt, property gets stolen. Then people want to know who is responsible, who gets blamed and who is liable for damages or compensation.

What holds for the non-digital world also holds for AI. Only when these technologies cause problems will there be a case for holding some person or entity responsible or liable. And these problems might be serious problems, i.e., problems that are encountered not just in a single case, but in many instances, in a systematic way. So what does "go wrong" mean in the case of AI technologies? Let's discuss two examples: the crash of an autonomous vehicle and mistargeting by an autonomous weapon.

5.1 Example 1: Crash of an Autonomous Vehicle

Imagine a vehicle driving in autonomous mode. Let's make this more specific: What does "autonomous" mean in this case? The European guidelines (and similarly the US NHTSA guidelines) distinguish between 6 levels of automated driving, from simple driver assistance systems which allow the car to brake or accelerate automatically to

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_5

fully autonomous vehicles which do not require a driver at all (see Sect. 10.1). The crucial leap happens from level 2 to level 3. On level 3 a driver is not required to monitor the functions of their car at any time. This implies that a car has to be able to deal with a broad range of cases, including critical situations, on its own. The vehicle needs to do this at least for a certain amount of time. Car manufacturers often set this time frame to 10 s before requiring the driver to take control again.

So if an accident happens during the time that the car was in control, who would be held responsible or liable? The driver was not in control, not even required to do so. The technology did just as it was programmed to do. The company maybe? But what if they did everything possible to prevent an accident—and still it happened?

5.2 Example 2: Mistargeting by an Autonomous Weapon

In Chap. 11 we talk about the military use of artificial intelligence in more detail. Here, we will focus our attention on only the liability aspects of autonomous weapons. In the military context the system in question is typically an autonomous weapon. International Humanitarian Law (IHL) permits combatants to kill, wound and capture enemy combatants. They are also allowed to destroy enemy war materiel, facilities and installations that support the enemy war effort. However, combatants are not allowed to kill non-combatants unless they are actively engaged in the enemy war effort. Thus it is lawful to bomb an enemy factory, phone exchange or power station. Indeed, it is lawful to kill civilians assisting the enemy war effort if they are working in factories making military equipment that contributes to the war effort. However there are many objects that are specifically protected. One cannot bomb a hospital or school or attack an ambulance. In this example, we discuss the questions of blame, responsibility and liability with respect to an autonomous weapon killing a person it is not permitted to kill under IHL.

Let us suppose there is a war between a Red state and a Blue state. The commander of the Red Air Force is Samantha Jones. It so happens, she has an identical twin sister, Jennifer Jones, who is a hairdresser. She makes no direct contribution to the war effort. She does not work in an arms factory. She does not bear arms. One day during the war, Jennifer takes her child to a kindergarten as she does every day. Blue has facial data of Samantha but knows nothing about her twin sister, Jennifer. An "autonomous" drone using facial recognition software spots Jennifer taking her daughter to kindergarten. It misidentifies Jennifer as a "high value target" namely her sister, Samantha. Just after she drops her child off at the school, the drone fires a missile at her car, killing her instantly.

In this situation, who or what is to blame for the wrongful killing of Jennifer? Who or what should be held responsible for this wrong. Who or what is liable to be prosecuted for a war crime for this killing? Who or what would be punished?

First let's clarify what would happen if the drone was remotely piloted by human pilots. In the human scenario, the pilots also use facial recognition software. They manoeuvred the drone into good position to get clear images. They ran several images

of Jennifer through their facial recognition system. They also looked at the photos of Samantha Jones they had themselves. All the images identified the image as General Samantha Jones. None identified the image as Jennifer Jones. This is because the facial recognition system only had images of high values targets, not of the entire enemy population. They were following orders from General Smith of the Blue Air Force to seek and destroy his counterpart in the Red Air Force.

5.2.1 Attribution of Responsibility and Liability

To blame an agent for a wrong, we must first establish the causal chain of events that led to the wrong. Also, we must understand who or what made the decisions that resulted in the wrong. In the human case, the humans relied on software to identify the target. They also used their own senses. The humans pressed the trigger that fired the missile. They did so because the software identified the target. The mistake was made because the software had incomplete information. There was an intelligence failure.

Is there much difference (if any) with regard to the Autonomous Weapon System (AWS) case? The AWS was following the same rules as the humans did. It fired for the same reason as the humans did. A computer database of facial images returned a match. Jennifer Jones did look like her sister and that is what led to her death.

There are one or two differences but both cases are very similar. Does it make sense to "blame" the AWS? Does it make sense to "blame" the humans? The fault lies in what was **not** in the facial recognition system. If Jennifer Jones was in the system the system might have reported that this image could be either Jennifer or Samantha. Then the humans would have to make a call as to what to do. With the facial recognition reporting "that is Samantha" and no one else, then a human would most likely assume they had their target too.

5.2.2 Moral Responsibility Versus Liability

We might well ask who is morally responsible for the wrongful death of Jennifer? We might consider the following people: the political leader who declared war, the pilot who fired the missile in the human case, the programmers who programmed the drone to fire missiles in the AWS case, or the intelligence officers who missed Samantha's sister. In reality, these are all "contributing factors" to the death of Jennifer. However, we can make some distinctions between moral responsibility and legal responsibility (i.e., liability).

Legal responsibility has to be proven in a court. If the human pilots were to be subject to a court martial for the unlawful killing of a civilian, then the facts would have to be proven to a standard beyond reasonable doubt. Certainly, the defence would argue the human pilots had a "reasonable expectation" that the facial recognition

system was accurate. Blame might be shifted to the intelligence officers who failed to find and load images of Jennifer into the system, or perhaps to the vendors of the facial recognition system. On these facts it is hard to see the pilots being found guilty of a war crime. Their intention was clearly to kill an enemy general. This would be a legal act in the context of war.

It has been said that law floats on sea of ethics. It is easy to imagine that General Jones would be furious to learn of the death of her sister. In the human case, she would naturally hold the pilots who fired the missile responsible. Perhaps she would hold the officer who ordered the mission responsible. Perhaps she would hold the political leadership of the Red State responsible. Regardless of who General Jones thinks is responsible, to assign legal responsibility, evidence would need to be presented in a court or legal proceeding. By bringing such proceedings, those who commit wrongs can be held accountable and those who do wrong can then be punished.

This is fine for humans but in the case of an AWS what does it mean to "punish" a system that cannot feel or suffer? If the AWS only does what it is programmed to do or acts on the basis of the data humans input into it, how can it be "responsible" for killing the wrong person? A pragmatic remedy is simply to assign responsibility for the actions of an AWS to a human or the state that operates it. There is an established legal concept called "strict liability" (described in the following section) which can be used in this regard.

5.3 Strict Liability

Product liability is the legal responsibility to compensate others for damage or injuries that a certain product has caused. Product liability is a concept that companies around the globe are familiar with and have to take into account. Clearly, when a company or one of its employees causes harm to others by neglecting their duties, the company will be held liable.

In some cases, there may not even have been a concrete wrongdoing. In the case of "strict liability," a company or a person can also be held liable even if they did not do anything wrong in the strict sense. For example, if someone owns a cat, and this cat causes damage to someone else's property, the owner is held liable in this sense. Or, a technique that has many beneficial consequences might also have some negative ones. For example, while vaccination is in general beneficial for the members of a society, there might be some cases where children suffer bad consequences from vaccines. These must be compensated for within the framework of strict liability. In the US the National Vaccine Injury Compensation Program provides this compensation.

A similar situation might arise in the context of autonomous vehicles or AI technologies in general, since a concrete (or even an abstract) harm may have been neither intended nor planned (otherwise it would have been a case of deliberate injury or fraud). Still, strict liability would be the framework within which to deal with these issues. Moreover, in many cases of AI technologies, harm might not have been foreseeable. This might even be a signature characteristic of these technologies, since

they often operate in ways which are, in a sense, opaque to observers. Even to the programmers themselves it is often not clear how exactly the system arrived at this or that conclusion or result.

In the case of autonomous vehicles a several shifts have been made to move the liability from the driver or car owner to the car manufacturer or, if applicable, the company operating or developing the software. This shift was made possible through the adaptation of the Vienna Convention on Road Traffic. More specifically through its update from the 1968 version to the 2014 that took effect inf 2016 (United Nations 1968).

In the German Ethics Code for Automated and Connected Driving (Luetge 2017), this step is explicitly being taken, for those cases where the car's system was in control. This implies, among others, that monitoring (black box) devices will have to be installed in those cars which clearly record who was in control at each moment the driver or the car.

5.4 Complex Liability: The Problem of Many Hands

In the AWS case we can see that "many hands" may be involved in a wrongful death. General Blue ordered the mission which went wrong. The pilots confirmed the target based on faulty intelligence and pressed the button that launched the missile. The facial recognition system made a decision based on incomplete data. Those who designed the system will say the fault lies with those who input the data. How will a judge determine who is responsible?

In a complex liability matter it is often the case that no one is found to be at fault. Intuitively, people want to blame someone but in complex events, often there is no single person who can be blamed. In such cases, the normal legal solution is to assign blame to a collective entity. In this case, blame would be assigned to the Blue State not any particular individual. Strict liability can be assigned to States operating an AWS. Even if no person deserves blame, the collective entity is held responsible and sanctioned accordingly.

5.5 Consequences of Liability: Sanctions

What sanctions would eventually be imposed must be left open at this point. In many cases, there might be a focus on sanctions against a company rather than against individuals, since programmers will have to work with a certain company strategy or policy. Corporate liability will be a key solution here, as in the case of fines against Volkswagen (around EUR 25 billion in 2016) or Bank of America (around USD 16 billion in 2014). This shows that the sanctions can actually be very substantial.

Discussion Questions:

- Describe the differences between moral and legal responsibility in your own words.
- If an AI does the wrong thing, who would you blame? The software? The programmer(s)? The company making it? Explain.
- How would you compensate a person that is severely injured by an autonomous car? Give reasons.

Further Reading:

- Robin Antony Duff. Answering for crime: *Responsibility and liability in the criminal law*. Hart Publishing, 2007. ISBN 978-1849460330. URL http://www.worldcat.org/oclc/1073389374
- Martha Klein. Responsibility. In Ted Honderich, editor, *The Oxford companion to philosophy*. OUP Oxford, 2005. ISBN 978-0199264797. URL http://www.worldcat.org/oclc/180031201
- Christoph Luetge. Responsibilities of online service providers from a business ethics point of view. In Mariarosaria Taddeo and Luciano Floridi, editors, *The Responsibilities of Online Service Providers*, pages 119–133. Springer, 2017. Doi: 10.1007/978-3-319-47852-4 7. URL https://doi.org/10.1007/978-3-319-47852-47.