5

Moral Responsibility and Autonomous Technologies

Does AI Face a Responsibility Gap?

Lode Lauwaert and Ann-Katrien Oimann

5.1 INTRODUCTION

There are several ethical conundrums associated with the development and use of AI. Questions around the avoidance of bias, the protection of privacy, and the risks associated with opacity are three examples, which are discussed in several chapters of this book. However, society's increased reliance on autonomous systems also raises questions around responsibility, and more specifically the question whether a so-called responsibility gap exists. When autonomous systems make a mistake, is it unjustified to hold anyone responsible for it?[1] In recent years, several philosophers have answered in the affirmative – we think primarily of Andreas Matthias and Robert Sparrow. If, for example, a self-driving car hits someone, in their opinion, no one can be held responsible. The argument we put forward in this chapter is twofold. First, there does not necessarily exist a responsibility gap in the context of AI systems and second, even if there would be, this is not necessarily a problem.

We proceed as follows. First, we provide some conceptual background by discussing respectively what autonomous systems are, how the notion of responsibility can be understood, and what the responsibility gap is about. Second, we explore to which extent it could make sense to assign responsibility to artificial systems. Third, we argue that the use of autonomous system does not necessarily lead to a responsibility gap. In the fourth and last section of this chapter, we set out why the responsibility gap is not necessarily problematic and provide some concluding remarks.

5.2 CONCEPTUAL CLARIFICATIONS

In the section, we first discuss what autonomous systems are. Next, we explain the concept of responsibility and what the responsibility gap is about. Finally, we describe how the responsibility gap differs from related issues.

[1] By "AI" in this text, we mean "autonomous AI."

5.2.1 *Autonomous Systems*

Before we turn to responsibility, let us begin with a brief exploration of AI systems, which are discussed in more details in the second chapter of this book. One of the most controversial examples are autonomous weapons systems or the so-called "killer robots,"[2] designed to kill without human intervention. It is to date unclear to which extent such technology currently already exists in fully autonomous form, yet the use of AI in warfare (which is also discussed in Chapter 20 of this book) is on the rise. For instance, a report by the UN Panel of Experts on Libya of 2020 mentions the system Kargu-2, a drone which may have hunted down and attacked retreating soldiers without any data connectivity between the operator and the system.[3] Unsurprisingly, the propensity toward ever greater autonomy in weapon systems is also accompanied by much speculation, debate, and protest.

For another example of an AI system, one can think of Sony's 1999 robot dog AIBO, a type of toy that can act as a substitute for a pet, which is capable of learning. The robot dog learns to respond to specific phrases of its "owner," or learns to adapt its originally programmed walking motion to the specific shape of the owner's house. AI systems are, however, not necessarily embedded in hardware. Consider, for instance, a software-based AI system that is capable of detecting lung cancer based on a pattern analysis of radiographic images, which can be especially useful in poorer regions where there are not enough radiologists. Amazon's Mechanical Turk platform is also a good example, as the software autonomously allocates tasks to suitable workers who subscribed to the platform, and subsequently handles their payment in case it – autonomously – verified that the task was adequately carried out. The uptake of AI systems is on the rise in all societal domains, which also means that questions around responsibility arise in various contexts.

5.2.2 *Notions of Responsibility*

The term "responsibility" can be interpreted in several ways. When we say "I am responsible," we can mean more than one thing by it. In general, a distinction can be made between three meanings: causal responsibility, moral responsibility, and role responsibility.[4] We will discuss each in turn.

[2] See among others: video Slaughterbots of 2017 by the Future of Life Institute and AI expert Stuart Russell, open letters in 2015 and 2017 by renowned technology experts to raise awareness among the general public around the dangers associated with the technology, The Campaign to Stop Killer Robots calling for a new international treaty.

[3] UN Security Council, Final report of the Panel of Experts on Libya established pursuant to Security Council resolution 1973 (2011), S/2021/229, 8 March 2021, https://documents.un.org/doc/undoc/gen/n21/037/72/pdf/n2103772.pdf?token=DtEs8GLFoOLY8vGG39&fe=true

[4] These terms already make it clear that we are not concerned here with the domain of the law and are therefore not talking about liability or legal responsibility. For a good overview of the different kinds of responsibility, see Nicole A. Vincent, "A Structured Taxonomy of Responsibility Concepts" in

Suppose a scientist works in a laboratory and uses a glass tube that contains toxic substances that if released would result in the death of many colleagues. Normally the scientist is careful, but a fly in the eye causes her to stumble. The result is that the glass tube breaks and the toxins are released, causing deaths. Asked who is responsible for the havoc, some will answer that it is the scientist. They then understand "responsibility" in a well-defined sense, namely in a causal sense. They mean that the scientist is (causally) responsible because she plays a role in the course of events leading to the undesirable result.

Let us make a slight modification. Say the same scientist works in exactly the same context with exactly the same toxic substances, but now also belongs to a terrorist group and wants the colleagues to die, and therefore deliberately drops the glass tube, resulting in several people dying. We will again hold the scientist responsible, but the content of this responsibility is clearly different from the first kind of responsibility. Without the scientist's morally wrong act, the colleagues would still be alive, and so the scientist is the cause of the colleagues' deaths. So, while the scientist is certainly causally responsible, in this case she will also be morally responsible.

Moral responsibility usually refers to one person, although it can also be about a group or organization. That person is then held responsible for something. Often this "something" is undesirable, such as death, but you can also be held responsible for good things, such as saving people. If a person is morally responsible, it means that others can respond to that person in a certain way: praise or reward when it comes to desirable things; disapproval or punishment when it comes to bad things. In addition, if one were to decide to punish or to reward, it would also mean that it is morally right to punish or reward that person. In other words, there would be good reasons to punish or reward that particular person, and not someone else. Note that moral responsibility does not necessarily involve punishment or reward. It only means that someone is the rightful *candidate* for such a response, that punishment or reward *may* follow. So, I may be responsible for something undesirable, but what happened was not so bad that I should be punished.

The third form, role responsibility, refers to the duties that come with a role or position. Parents are responsible in this sense because they must ensure their children grow up in a safe environment, just as it is the role responsibility of a teacher to ensure a safe learning environment for students. When revisiting the earlier example of the scientist, we can also discuss her responsibility without referring to her role in a chain of events (causal responsibility) or to the practice of punishment and reward (moral responsibility). Those who believe that the scientist is responsible may in fact refer to her duty to watch over the safety of the building, to ensure that the room is properly sealed, or to verify that the glass tubes she uses do not have any cracks.

Nicole Vincent, Ibo van de Poel, and Jeroen van den Hoven (eds), *Moral Responsibility. Library of Ethics and Applied Philosophy* (Dordrecht Springer, 2011) who develops a taxonomy of responsibility concepts inspired by H. L. A Hart's illustration of the drunken ship captain.

These three types of responsibilities are related. The preceding paragraphs make it clear that a person can be causally responsible without being responsible in a moral sense. We typically do not condemn the scientist who trips over a shoelace. Conversely, though, moral responsibility always rests on causal responsibility. We do not hold someone morally responsible if they are in no way part of the process that led to the (un)desired result. That causal involvement, by the way, should be interpreted in a broad sense. Suppose the scientist is following an order. The person who gave the order is then not only causally but also morally responsible, despite not having committed the murder itself. Finally, role responsibility is always accompanied by moral responsibility. If, for example, as a scientist it is your duty to ensure that the laboratory is safe, it follows at least that you are a candidate for moral disapproval or punishment if it turns out that you have not done your duty adequately, or that you can be praised or rewarded if you have met the expectations that come with your role.

5.2.3 *Responsibility Gap*

Autonomous systems lead to a responsibility gap, some claim.[5] But what does one understand by "responsibility" here? Clearly, one is not talking about causal responsibility in this context. AI systems are normally created by humans (we say "normally" because there already exist AI systems that design other AI systems). Therefore, if one would claim that no humans are involved in the creation of AI systems, this would come down to a problematic view of technology.

The responsibility gap is also not about the third form of responsibility namely role responsibility. That argument refers to the duty of engineers, not so much to create more sustainability or well-being, but to make things that have as little undesirable effect on moral values as possible, and thus to think about such possible effects in advance. Since there is no reason why this should not apply to the developers of autonomous systems, the responsibility gap does not mean that developers and users of AI systems have no special duties attached to them. On the contrary, such technology precisely affirms the importance of moral duties. Because the decision-making power is being transferred to that technology, and because it is often impossible to predict exactly what decision will be made, the developers of AI systems must think even more carefully than other tech designers about the possible

[5] See among others: Roos De Jong, "The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to Nyholm" (2020) *Science and Engineering Ethics*, 26; Robert Sparrow, "Killer robots" (2007) *Journal of Applied Philosophy*, 24; Andreas Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata" (2004) *Ethics and Information Technology*, 6. The term was first used by Andreas Matthias with respect to autonomous machines (2004) and was later applied to autonomous weapon systems by Robert Sparrow (2007). For a recent overview of the discussion on autonomous weapons, see: Ann-Katrien Oimann, "The responsibility gap and LAWS: A critical mapping of the debate" (2023) *Philosophy & Technology*, 36.

undesirable effects that may result from the algorithms' decisions in, for example, the legal or medical world.[6]

The thesis of the so-called responsibility gap is thus concerned with moral responsibility. It can be clarified as follows: in the case of mistakes made by autonomous AI, despite a possible spontaneous tendency to punish someone, that tendency has no suitable purpose as there is no candidate for punishment.

5.2.4 *Related but Different Issues*

Before we examine whether the thesis of the responsibility gap holds water, it is useful to briefly touch upon the difference between the alleged problem of the responsibility gap and two other problems. The first problem is reminiscent of a particular view of God and the second is the so-called problem of many hands.

Imagine strolling through the city on a sunny afternoon and stepping in chewing gum. You feel it immediately: with every step, your shoe sticks a little to the ground and your mood changes, the sunny afternoon is gone (at least for a while) and you are looking for a culprit. However, the person who left the gum on the ground is long gone. There is definitely someone causally responsible here: someone dropped the gum at some point. And the causally responsible person is also morally responsible. You're not supposed to leave gum, and if you do it anyway, then you're ignoring your civic duty and you're justified in being reprimanded. However, the annoying thing about the situation is that it is not possible to detect the morally responsible person.

The problem in this example is that you do not know who the morally responsible person is, even though there is a responsible person. This is reminiscent of the relationship between man and God as described in the Old Testament. God created the world, but has subsequently distanced Himself so far from His creation that it is impossible for man to perceive Him. In the case of the responsibility gap the problem is of a different nature. Here it is not an epistemic problem, but an ontological problem. The difficulty is not that I do not know who is responsible; the problem is that there is no one morally responsible for the errors caused by an autonomous system, so the lack of knowledge cannot be the problem here.

The second problem that deviates from the responsibility gap is the problem of many hands.[7] This term is used to describe situations where many actors have contributed to an action that has caused harm and it is unclear how responsibility

[6] See in this regard also Hans Jonas' study *Das Prinzip Vernatwortung* (1979), which is one of the first major works on the ethics of technology. Jonas suggested that in a modern world, the effects of technology are not so uncertain that designers need to think about the consequences even more so than before.

[7] The expression "many hands" was reportedly first used by Dennis Thompson, "Moral responsibility and public officials: The problem of many hands" (1980) *American Political Science Review*, 74 and later applied to computer technology by Helen Nissenbaum, "Accountability in a computerized society" (1996) *Science and Engineering Ethics*, 2.

should be allocated. It is often used with respect to new technologies such as AI systems because a large number of actors are involved in their development and use, but the problem also occurs in nontechnical areas such as climate change.

To illustrate this problem, we turn to the disaster of the Herald of Free Enterprise, the boat that capsized on March 6, 1987, resulting in the deaths of nearly 200 people. An investigation revealed that water had flowed into the boat. As a result, the already unstable cargo began to shift to one side. This displacement eventually caused the ferry to disappear under the waves just outside the port of Zeebrugge in Belgium. This fatal outcome was not the result of just one cause. Several things led to the boat capsizing. Doors had been left open, the ship was not stable in the first place, the bulkheads that had been placed on the car deck were not watertight, there were no lights in the captain's cabin, and so on. Needless to say, this implies that several people were involved: the assistant boatman who had gone to sleep and left the doors open; the person who had not checked whether the doors were closed; and finally, the designers of the boat who had not fitted it with lights.

There are so many people involved in this case that not only one person can be held responsible. But this differs from saying that no one is responsible. The case is not an example of an ontological problem; there is no lack of moral responsibility in the case of the capsized ferry. Indeed, there are multiple individuals who are morally responsible. There is, however, an epistemic problem. The problem is that there are so many hands involved that it is very difficult (if not impossible) to know exactly who is responsible for what and to what extent each person involved is responsible. In the case of the Herald of Free Enterprise, many knots had to be untangled in terms of moral responsibility, but that is different from claiming that the use of a technology is associated with a responsibility gap.

5.3 CAN AI BE MORALLY RESPONSIBLE?

Is it true that there is no moral responsibility for mistakes made by an AI system? There is an answer to that question that is often either not taken seriously or overlooked, namely the possibility of AI systems being responsible themselves.[8] To be clear, we refer here to moral responsibility and not the causal type of responsibility. After all, autonomous technologies very often play a causal role in a chain of events with an (un)desirable outcome. Our question is: is it utter nonsense to see an AI system as the object of punishment and reward, praise, and indignation?

One of the sub-domains of philosophy is philosophical anthropology. A central question in that domain is whether there are properties that separate humans from, say, plants and nonhuman animals, as well as from artificial entities. In that context,

[8] An author like Joanna Bryson explicitly rejects this option, emphasizing that autonomous systems are essentially nonexistent and should be viewed as nothing more than tools: Joanna Bryson, "Robots should be slaves" in Yorick Wilks (ed), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues* (John Benjamins Publishing Company, 2010).

one can think, for instance, of the ability to play and communicate, to suffer psychologically or to get gray hair. However, it is almost impossible not to consider responsibility here. After all, today, we only attribute that moral responsibility to human beings. Sure, there are some exceptions to this rule. For instance, we do not hold people with mental disabilities or disorders responsible for a number of things. And we also punish and reward animals that are not humans. But moral responsibility is something we currently reserve exclusively for humans, and thus do not attribute to artifacts.

Part of the reason we do not hold artificial entities responsible has to do with what responsibility entails. We recall that a morally responsible person is the justifiable target of moral reactions such as punishment and reward, anger and indignation. Those reactions do not necessarily follow, but if they follow then the responsible person is the one who is justifiably the subject of such a reaction. But that presupposes the possibility of some form of sensation, the ability to be affected in the broad sense, whether on a mental or physical level. There is no point in designating someone as responsible if that person cannot be affected by the moral reactions of others. But where do we draw the line? Of course, the ability to experience pain or pleasure in the broad sense of the word is not sufficient to be morally responsible. This is evident from our dealings with nonhuman animals: dogs can experience pain and pleasure, but we do not hold them responsible when they tip a vase with their tail. However, the ability to be physically or mentally affected by another's reaction is a necessary condition. And since artifacts such as autonomous technologies do not currently have that ability, it would be downright absurd to hold them responsible for what they do.

On the other hand, moral practices are not necessarily fixed forever. They can change over the course of history. Think about the allocation of legal rights. At the end of the eighteenth century, people were still arguing against women's rights based on the following argument: if we grant rights to women, then we must also grant rights to animals. The concealed assumption was that animal rights are unthinkable. Meanwhile, it is completely immoral to deny women rights that are equal to those of men. One can also think of the example of the robot Sophia who was granted citizenship of Saudi Arabia in 2017. If throughout history more and more people have been granted rights, and if other moral practices have changed over time, why couldn't there be a change when it comes to moral responsibility? At the time of writing, we cannot hold artifacts responsible; but might it be possible in the future?

That question only makes sense if it is not excluded that robots in the future may be affected on a physical or mental level, and they may later experience pain or pleasure in some way. If that ability can never exist, then it is out of the question that our moral attitudes will change, then we will never hold AI systems morally responsible. And exactly that, some say, is the most realistic scenario: we will never praise technology because it will never be capable of sensation on a physical or mental level.

Much can be said about that assertion. We will limit ourselves to a brief response to the following thought experiment that is sometimes given to support that claim.

Suppose a robot that looks like a human falls down the stairs and reacts as humans normally do by providing the output that usually follows the feeling of pain: yelling, crying, and so on. Is the robot in pain? Someone may react to the robot's fall, for example, because it is a human reflex to react to signs of pain. However, one is unlikely to respond because the robot is in pain. Although the robot does show signs of pain, there is no pain, just as computer programs such as Google Translate and DeepL do not really understand the sentence that they can nevertheless translate perfectly.

AI can produce things that indicate pain in humans, but those signals, in the case of the software, are not in themselves a sufficient reason to conclude that the technology is in pain. However, we cannot conclude at this point that AI systems will never be able to experience pain nor exclude that machines will never be able to be affected mentally. Indeed, next to software, technologies usually consist of hardware as well and the latter might be a reason not to immediately cast aside the possibility of pain.[9] Why?

Like all physiological systems of a human body, the nervous system is made up of cells, mainly neurons, which constantly interact. This causal link ensures that incoming signals lead to the sensation of pain. Now, suppose that you are beaten up, and that for 60 minutes, you are actually in pain, but that science has advanced to the point where the neurons can be replaced by a prosthesis, microchips, for example, without it making any difference otherwise. The chips are made on a slice of silicon – but other than that, those artificial entities do exactly the same thing as the neurons: they send signals to other cells and provide sensation. Well, imagine that, during one month and step by step, a scientist replaces every cell with a microchip so that your body is no longer only made up of cells but also of chips. Is it still utter nonsense to claim that robots might one day be able to feel pain?

To avoid confusion, we would like to stress the following: we are not claiming that intelligent systems will one day be able to feel pain, that robots will one day resemble us – us, humans – in terms of sensation. At most, the last thought experiment was meant to indicate that it is perhaps a bit short-sighted to simply brush this option aside as nonsense. Furthermore, if it does turn out that AI systems can experience pain, we will not automatically hold them morally responsible for the things they do. The reason is that the ability to feel pain is not enough to be held responsible. Our relationships with nonhuman animals, for example, demonstrate

[9] Debates about embodied information are discussed for many years in philosophy of mind. In this regard, see, among others: Daniel Dennett, *Consciousness Explained* (Little Brown, 1992); John Rogers Searle, "Minds, brains, and programs" (1980) *The Behavioral and Brain Sciences*, 3: 417–57; Hubert Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence* (Harper & Row, 1972); Alan Turing, "Computer Machinery and Intelligence" in Edward A. Feigenbaum and Julian Feldman (eds), *Computers and Thought* (McGraw-Hill, 1963).

this, as we pointed out earlier. Suppose, however, that all conditions are met (we will explain the conditions in the next section), would that immediately imply that we will see AI systems as candidates for punishment and reward? Attributing responsibility exclusively to humans is an age-old moral practice, which is why this may not change any time soon. At the same time, history shows that moral practices are not necessarily eternal, and that the time-honored practice of attributing rights only to humans is only gradually changing in favor of animals that are not humans. That alone is a reason to suspect that ascribing moral responsibility to robots may not become a reality in the near future, even if robots could be affected physically or mentally by reward or punishment.

5.4 THERE IS NO RESPONSIBILITY GAP

So we must return to the central question: do AI systems create a responsibility gap? Technologies themselves cannot be held morally responsible today, but does the same apply to the people behind the technology?

There is reason to suspect that you can hold people morally responsible for mistakes made by an AI system. Consider an army officer who engages a child soldier. The child is given a weapon to fight the enemy. But in the end, the child kills innocent civilians, thus committing a war crime. Perhaps not many people would say that the child is responsible for the civilian casualties, but in all likelihood we would believe that at least someone is responsible, that is, the officer. However, is there a difference between this case and the use of, for example, autonomous weapons? If so, is that difference relevant? Of course, child soldiers are human beings, robots are not. In both cases, however, a person undertakes an action knowing that undesirable situations may follow and that one can no longer control them. If the officer is morally responsible, why shouldn't the same apply to those who decide to use autonomous AI systems? Are autonomous weapons and other autonomous AI systems something exceptional in that regard?

At the same time, there is also reason to be skeptical about the possibility of assigning moral responsibility. Suppose you are a soldier and kill a terror suspect. If you used a classic weapon that functions as it should, a 9-mm pistol for example, then without a doubt you are entirely – or at least to a large extent – responsible for the death of the suspect. Suppose, however, that you want to kill the same person, and you only have a semiautomatic drone. You are in a room far away from the war zone where the suspect is, and you give the drone all the information about the person you are looking for. The drone is able to scout the area itself, and when the technology indicates that the search process is over, you can assess the result of the search and then decide whether or not the drone should fire. Based on the information you gathered, you give the order to fire. But what actually happens? The person killed is not the terror suspect and was therefore killed by mistake. That mistake has everything to do with a manufacturing error, which led to a defect in

the drone's operating. Of course, that does not imply that you are in no way morally responsible for the death of the suspect. So, there is no responsibility gap, but probably most people would feel justified in saying that you are less responsible than if you used a 9-mm pistol. This has to do with the fact that the decision to fire is based on information that comes not from yourself but from the drone, information that incidentally happens to be incorrect.

For many, the decrease in the soldier's causal role through technology is accompanied by a decrease in responsibility. The siphoning off of an activity – the acquisition of information –implies, not that humans are not responsible, but that they are responsible to a lesser degree. This fuels the suspicion that devolving all decisions onto AI systems leads to the so-called responsibility gap. But is that suspicion correct? If not, why? These questions bring us to the heart of the analysis of the issue of moral responsibility and AI.

5.4.1 *Conditions for Responsibility*

Our thesis is that reliance on autonomous technologies does not imply that we can never hold anyone responsible for their mistakes. To argue this, we must consider whether the classical conditions for responsibility are also met. We already referred to the capacity for sensation in the broad sense of the word, but what other conditions must be fulfilled for someone to be held responsible? Classically, these are three sufficient conditions for moral responsibility: causal responsibility, autonomy, and knowledge.

It goes without saying that moral responsibility presupposes causal responsibility. Someone who is not involved at all in the creation of the (undesirable) result of an action cannot be held morally responsible for that result. In the context of AI systems, several people meet this condition: the programmer, the manufacturer, and the user. However, this does not mean that we have undermined the responsibility gap theorem. Not every (causal) involvement is associated with moral responsibility. Recall the example of the scientist in the laboratory we discussed earlier: we do hold the scientist responsible, but only in a causal sense.

Thus, more is needed. Moral responsibility also requires autonomy. This concept can be understood in at least two ways. First, "autonomy" can be interpreted in a negative way. In that case, it means that the one who is autonomous in that respect can function completely independently, without human intervention. For our reasoning, only the second, positive form is relevant. This variant means that you can weigh things against each other, and that you can make your own decision based on that. However, the fact that you are able to deliberate and decide is not sufficient to be held morally responsible. For example, you may make the justifiable decision to kill the king, but when the king is killed, you are not necessarily responsible for it, for example, because someone else does it just before you pull the trigger and independently of your decision. You are only responsible if your deliberate decision

is at the root of the murder, that is, if there is a causal link between the autonomy and the act.

Knowledge is the final condition. You can only be held morally responsible if you have the necessary relevant knowledge.[10] One who does not know that an action is wrong cannot be responsible for it. Furthermore, if the consequences of an act are unforeseeable, then you cannot be punished either. Note that, the absence of knowledge does not necessarily exonerate you. If you may not know certain things while you should have known them, and the lack of knowledge leads to an undesirable result, then you are still morally responsible for that result. For example, if a driver runs a red light and causes an accident as a result, then the driver is still responsible for the accident, even if it turns out that she was unaware of the prohibition against running a red light. After all, it is your duty as a citizen and car driver – read: your role responsibility – to be aware of that rule.[11]

5.4.2 *Control as Requirement*

So, whoever is involved in the use of a technology, whoever makes the well-considered decision to use that technology, and whoever is aware of the necessary relevant consequences of that technology, they can all be held morally responsible for everything that goes wrong with the technology. At least that is what the classical analysis of responsibility implies. So why do authors such as Matthias and Sparrow nevertheless conclude that there are responsibility gaps?

They point to an additional condition that must be met. Once an action or certain course of events has been set in motion, they believe you must have control over it. So even if you are causally involved, for example, because you have made the decision that the action or course of events should take place, while you can do nothing else about it at the time it was initiated, it would be unfair to punish you when it all results in an undesirable outcome. They argue that, since AI systems can function completely independently, in such a way that you cannot influence their decisions due to their high degree of autonomy and capacity for self-learning, you cannot hold anyone responsible for the consequences.

[10] According to an ordinary conception of responsibility attribution, it is only fitting to hold someone responsible if the agent can foresee that the device will or is likely to create a certain kind of outcome. This is usually termed the epistemic condition and many philosophers agree that such a requirement is a necessary condition for moral responsibility. See among others: John Martin Fischer and Neal A. Tognazzini, "The truth about tracing" (2009) *Noûs*, 43; John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge University Press, 2000); Michael J. Zimmerman, "Moral responsibility and ignorance" (1997) *Ethics*, 107.

[11] The epistemic condition often relies on a *tracing* strategy and plays an important role in many theories of responsibility. It is used in cases where an agent is blameworthy for the harm caused based on the ground that her responsibility can be traced back to previous acts of the agent when she did meet the conditions to fulfill on moral responsibility. See for example: John Martin Fischer and Neal A. Tognazzini, "The truth about tracing" (2009) *Noûs*, 43.

If you are held responsible for an action, it usually means that you have control. As CEO, I am responsible for my company's poor numbers because I could have made different decisions that benefited the company more. Conversely, I have no control over a large number of factors and thus bear no responsibility for them. For example, I have no control over the weather conditions, nor do I bear any responsibility for the consequences of good or bad weather. Thus, responsibility is often accompanied by control, just as the absence of control is usually accompanied by the absence of responsibility. Yet we argue that it is false to say that you must have control over an initiated action or course of events to be held responsible, and that not having control takes away your responsibility. This is demonstrated by the following.

Imagine you are driving and after a few minutes, you have an epileptic seizure that causes you to lose control of the wheel and to seriously injure a cyclist. It is not certain that you will be punished, let alone receive a severe sentence, but perhaps few, if any, will hold you responsible for the cyclist's injury, in spite of your lack of control of the car's steering wheel. This is mainly the case because you possess all the relevant knowledge. You do not know that a seizure will occur within a few minutes, but as someone with epilepsy you do know that there is a risk of a seizure and that it may be accompanied by an accident. Furthermore, you are autonomous (in a positive sense). You are able to weigh up the desire to drive somewhere by yourself against the risk of an attack, and to decide on that basis. Finally, you purposefully get in the car. As a result, you are causally connected to the undesirable consequence in a way that sufficiently grounds moral responsibility. After all, if you decide knowing that it may lead to undesirable consequences, then you are justified in considering yourself a candidate for punishment at the time the undesirable consequence actually occurs. Again, it is not certain that punishment will follow, but those who take a risk are responsible for that risk, and thus can be punished when it turns out that the undesirable consequence actually occurs.

We can conclude from the above that not having control does not absolve moral responsibility. Therefore, we do not believe that AI systems are associated with a responsibility gap due to a lack of control over the technology. However, we cannot conclude from the foregoing that the idea of a responsibility gap in the case of autonomous AI is incorrect and that in all cases someone is responsible for the errors caused by that technology. After all, perhaps situations might occur in which the other conditions for moral responsibility are not met, thus still leading us to conclude that the use of autonomous AI goes hand in hand with a responsibility gap.

5.4.3 *Is Someone Responsible?*

To prove that it is not true that no one can ever be held responsible, we invoke some previously cited examples: a civilian is killed by an autonomous weapon and a self-driving car hits a cyclist.

To begin with, it is important to note that both dramatic accidents are the result of a long chain of events that stretch from the demand for production, through the search for funding, and finally to the programming and use. If we are looking for a culprit, we might be able to identify several people – one could think of the designer or producer, for example – but the most obvious culprit is the user: the commander who decides to deploy an autonomous weapon during a conflict, or the occupant of the autonomous car. It is justified to put them forward as candidates for punishment for the following reasons, just as the epilepsy patient is responsible for the cyclist's injury.

First of all, both are aware of the context of the use and of the possible undesirable consequences. They do not know whether or not an accident will happen, let alone where and when exactly. After all, autonomous cars and weapons are (mainly) based on machine learning, which means that it is not (always) possible to predict what decision will be made. But the kinds of accidents that can happen are not unlimited. Killing civilians and destroying their homes (autonomous weapons) and hitting a cyclist or crashing into a group of people (self-driving car) are dramatic but foreseeable; as a user, you know such things can happen. And if you don't know, that is a failure from your part: you should know. It is your duty, your role responsibility, to consider the possible negative consequences of the things you use.

Second, both commander and owner are sufficiently autonomous. They are able to weigh up the advantages and disadvantages: the chance of fewer deaths in their own ranks and war crimes (autonomous weapons), the chance of being able to work while on the move and traffic casualties (self-driving car).

Third, if, based on these considerations, the decision is made to effectively pursue the use of autonomous cars and weapons, while knowing that it may bring undesirable consequences, then it is justifiable to hold both the commander and owner responsible for deliberately allowing the undesirable anticipated consequences to occur. Those who take risks accept responsibility for that risk; they accept that they may be penalized in the event that the unwanted, unforeseen consequence actually occurs.

Thus, in terms of responsibility, the use of AI systems is consistent with an existing moral practice. Just as you can hold people responsible for using nonautonomous technologies, people are also responsible for things over which they have no control but with which they are connected in a relevant way. So not only does the autonomy of technology not erase the role responsibility of the user; it does not absolve moral responsibility either. The path the system takes to decide may be completely opaque to the user, but the system does not create a responsibility gap.

Those who disagree must either demonstrate what is wrong with the existing moral practice in which we ascribe responsibility to people or demonstrate the relevant difference between moral responsibility in the case of autonomous systems and everyday moral practice. Of course, there are differences between using an autonomous system on the one hand and driving a car as a patient on the other. The question, however, is whether those differences matter when it comes to moral responsibility.

To be clear, our claim here is only that the absence of control does not necessarily lead to a gap. The thesis we put forward is not that there can never be a gap in the case of AI. The reason is that the third, epistemic condition must be met. There is no gap if the consequences are and should be foreseen (and if there is autonomy and a causal link). In contrast, there may be a gap in case the consequences are unforeseeable (or in case one of the other conditions is not met).

5.5 IS A RESPONSIBILITY GAP PROBLEMATIC?

We think there are good reasons to believe that at least someone is responsible when autonomous AI makes mistakes – maybe there is even collective responsibility[12] – since it is sufficient to identify one responsible person to undermine the thesis of a responsibility gap (assuming the other conditions are met). However, suppose that our analysis goes wrong in several places, and that you really cannot hold anyone responsible for the damage caused by the toy robot AIBO, Google's self-driving car, Amazon's recruitment system, or the autonomous weapon system. In that case, would that make an argument for the conclusion that ethics is being disrupted by autonomous systems? In other words, would this gap also be morally problematic? To answer that question, we look at two explanations for the existence of the practice of responsibility. The first has to do with prevention; the second points to the symbolic meaning of punishment.

Someone robs a bank, a soldier kills a civilian, and a car driver ignores a red light: these are all examples of undesirable situations that we, as a society, do not want to happen. To prevent this, to ensure that the norm is not infringed again later, something like the imputation of responsibility was created, a moral practice based on the psychological mechanism of classical conditioning. After a violation, a person is held responsible and is a candidate for unpleasant treatment, with the goal of preventing the violation from happening again in the future.

That goal, prevention, must obviously be there, and it is clear that the means – punishing the responsible party – is often sufficient to achieve the goal. Yet prevention is not necessarily related to punishment; punishing the person responsible is not necessary for the purpose of prevention. There are ways other than punishment to ensure that the same mistake is not made again. You can teach people to follow the rules, for example, by giving them extra explanations and setting a good example. It is possible that undesirable situations will not occur in the future without moral responsibility. This appears to be exactly the case in the context of AI.

Take an algorithm that ignores all women's cover letters, or the Amazon Mechanical Turk platform that wrongfully blocks your account, preventing you

[12] It is debated whether collective entities can be qualified as group agents that can be held morally responsible. See: Neta C. Crawford "Organizational responsibility" in *Accountability for Killing: Moral Responsibility for Collateral Damage in America's Post-9/11 Wars* (Oxford University Press, 2013); Christian List, "Group agency and artificial intelligence" (2021) *Philosophy & Technology*, 34.

from accepting jobs. To prevent such a morally problematic event from occurring again in the future, it is natural that the AI system is tinkered with by someone with sufficient technical knowledge, such as the programmer. It is quite possible that the system has so many layers that the designer cannot see the problem and therefore cannot fix it. But it is also possible that the programmer can successfully intervene, to the extent that the AI system will not make that mistake in future. In that case, the technical work is sufficient for preventing the problem, and further, for the purpose of prevention, you don't need anyone to be a candidate for punishment – we raise again that this is the definition of moral responsibility. In other words, if the goal is purely preventive in nature, then the solely technical intervention of the designer can suffice and thus the alleged absence of moral responsibility is not a problem.

There is another purpose that is often cited to justify the imputation of responsibility. That purpose has a symbolic character. Namely, it is about respecting the dignity of a human being. Is that goal, too, related to the designation of a candidate for punishment? In light of that objective, would a responsibility gap be a problem?

In a liberal democracy, everyone has moral standing. Whatever your characteristics are and regardless of what you do, you have moral standing due to the mere fact of being a human, and that counts for everyone. That value is only substantial insofar as legal rights are attached to that value. The principle that every human being has moral value implies that you have rights and that others have duties toward you. Among other things, you have the right to education and employment, and others may not intentionally hurt or insult you without good reason. It is permitted for an employer to decide not to hire you on the basis of relevant criteria, but it flagrantly violates your status as a being with moral standing if they belittle or ridicule you during a job interview without good reason.

Imagine the latter happens. This is a problem, because it is a denial of the fact that you have moral standing. Well, the practice of imputing moral responsibility is at least in part a response to such a problem. Something undesirable takes place – a person's dignity is violated – and in response someone is punished, or at least that person is designated as a candidate for punishment. Punishment here means that a person is hurt and experiences an unpleasant sensation, something that you do not wish for. Now the purpose of that punishment, that unpleasant experience, is to underscore that the violation of dignity was a moral wrong, and thus to affirm the dignity of the victim. The punishment does not heal the wound or undo the error, but it has symbolic importance. It cuts through the denial of the moral status that was inherent to the crime.

The affirmation of moral value is clearly a good, and a goal that can be realized by means of punishment. However, it is questionable whether that goal can be achieved exclusively by these means. Suppose an autonomous weapon kills a soldier. Suppose, moreover, that it is true, contrary to what we have just argued, that no one can be held responsible for this death. Does that mean that the moral

value of the soldier can no longer be emphasized? It is true that assigning responsibility expresses the idea that the value of the soldier is taken seriously. Moreover, it is undoubtedly desirable that, out of respect for the value of individual, someone should be designated as a candidate for punishment. However, the claim that responsibility is necessary for the recognition of dignity is false. One can also do justice to the deceased without holding anyone responsible. Perhaps the most obvious example of this is a funeral. After all, the significance of this ritual lies primarily in the fact that it underscores that the deceased has intrinsic value.

To be clear, we are not claiming that ascribing moral responsibility is a meaningless practice. Nor do we mean to say that, if the use of AI led to a gap, the impossibility of holding someone responsible would never be a problem. Our point is that prevention and respect are not in themselves sufficient reasons to conclude that a responsibility gap in the context of AI is a moral tragedy.[13]

5.6 CONCLUSION

AI offers many opportunities, but also comes with (potential) problems – many of which are discussed in the various chapters of this handbook. In this contribution, we focused on the relationship between AI and moral responsibility, and make two arguments. First, the use of autonomous AI does not necessarily involve a responsibility gap. Second, even if this were the case, we argued why that is not necessarily morally problematic.

[13] This manuscript is based partly on: Lode Lauwaert, *Wij robots: Een filosofische blik op technologie en artificiële intelligentie* (LannooCampus, 2021); Lode Lauwaert, "Artificial intelligence and responsibility" (2021) *AI & Society*; Lode Lauwaert, "Artificiële intelligentie en normatieve ethiek: Wie is verantwoordelijk voor de misdaden van LAWS?" (2019) *Algemeen Nederlands tijdschrift voor wijsbegeerte*.