4 Responsibility and Liability

One of the greatest sources of angst around the advent of increasingly intelligent computer systems is the alleged erosion of human responsibility they'll entail. International discussions on the use of lethal autonomous weapons systems provide a forceful illustration of what's at stake. If weapons systems powered by complex and potentially opaque machine-learning technology will soon be able to decide, unassisted, when to engage a target, will we still meaningfully be able to hold human beings responsible for war crimes involving these systems? In war we're obviously talking about life and death decisions, but similar worries play out elsewhere—areas in which lives may not be directly on the line but in which advanced systems can nevertheless have significant effects on people and society, such as in law enforcement, transportation, healthcare, financial services, and social media. If AI technologies can decide things without direct human intervention in ways that are too complex or too fast for human beings to properly understand, anticipate, or change, will we still be able to hold human beings responsible for these systems? And would we *want* to hold them responsible? What would be the alternative? Can *machines* be responsible?

What exactly *is* responsibility, and what qualities of AI challenge existing ways of distributing it? Does responsibility simply vanish? Does it fade away gradually, or is something else happening entirely? In this chapter we'll look at some of these questions. We'll begin by unpacking some of the different senses of responsibility, including moral and legal responsibility. We'll then highlight some of the ways in which technology, in general, affects these senses. In the remainder of the chapter we'll home in on some of the discussions about various aspects of responsibility that developments in AI—as a special kind of technology—have generated.

Picking Apart Responsibility

Responsibility is one of those concepts most of us *think* we understand. Within moral philosophy, legal scholarship, and even daily conversation, the word is often used interchangeably with words like "accountability," "liability," "blameworthiness," "obligation," and "duty."[1] Yet there are nuanced differences between these terms. Consider this little snippet from the philosopher H. L. A. Hart, describing a (fictional) sea captain:

> As captain of the ship, [he] was responsible for the safety of his passengers and crew. But on his last voyage he got drunk every night and was responsible for the loss of the ship with all aboard. It was rumored that he was insane, but the doctors considered that he was responsible for his actions. Throughout the voyage he behaved quite irresponsibly, and various incidents in his career showed that he was not a responsible person. He always maintained that the exceptional winter storms were responsible for the loss of the ship, but in the legal proceedings brought against him he was found criminally responsible for his negligent conduct, and in separate civil proceedings he was held legally responsible for the loss of life and property. He is still alive and he is morally responsible for the deaths of many women and children.[2]

Hart's story isn't supposed to be riveting—which is just as well!—but it does make its point vividly: responsibility is a complex, multifaceted notion. It can refer to causal contribution, as when winter storms cause the loss of a ship or to the formal mechanism used to attribute legal responsibility. It can be used to describe a character trait and also the obligations and duties that come with the professional role of a sea captain. It can mean any or all of these things and other things besides. Responsibilities are also by nature diffuse. Events can rarely be pinned down to one person or factor alone. Why was the drunken captain allowed to be captain if he had a history of alcoholism? Why wasn't this picked up? Who cleared the captain as safe? Who designed the screening protocol?

In this chapter, we're not going to consider all the forms of responsibility Hart explored in his story. Our focus instead will be on moral and legal notions of responsibility, since it's at these points where the rubber hits the road, so to speak, as far as artificial intelligence is concerned. Let's start by considering how they differ.

First, moral responsibility is very often (but not always) *forward*-looking, and legal responsibility is very often (but not always) *backward*-looking.[3] This is a crucial point, in fact, because many of the differences between

these two forms of responsibility can be traced back to this one pivotal distinction. What does it *mean* to be forward-looking or backward-looking?

Responsibility is forward-looking when it relates to the prospective duties and obligations a person may have. Expecting a certain degree of care and civility in human relationships is normal. Every time we step into a car, we take on various obligations—some to pedestrians, some to motorists, others to public and private property-holders. In our workplaces we expect that a certain level of professional courtesy will be extended to us. In the domain of goods and services, we naturally expect that manufacturers will have an eye to safe construction and assembly. And when it comes to autonomous systems, we naturally expect that software engineers will take the risks of certain design decisions into account when configuring their programs. If they know a credit-risk assessment tool has a good chance of unfairly disadvantaging a particular group of people, they have a responsibility—if not a legal one than certainly a moral one—to adjust the system in order to reduce the risk of unfair bias. Such forward-looking responsibilities are about events that will happen in the future and are the bread and butter of instruction manuals, professional codes, regulations, training programs, and company policies.

Backward-looking responsibility is different altogether; it's about accountability for events that have *already* happened. If an accident occurs, chains of responsibility will be traced back in order to determine who was at fault and should be called on to answer for their actions. Backward-looking responsibilities most often result in legal liability (see below), or some other kind of formal or informal reckoning.

So in what other ways do moral and legal responsibility differ? Well, let's take a step back for a moment.

A useful way of thinking about *any* kind of responsibility is to carefully consider the relationships in which it actually arises and how burdens of responsibility are distributed within them. After all, responsibility is fundamentally a relational concept—it's about what we owe each other and the demands it's reasonable to place on our fellows.

Relationships of responsibility have several components. In a simple breakdown,

- the *agent* is the person (or group) who performs the action;
- the *patient* is the person (or group) who receives or is somehow affected by the action;

- the *forum* is the person (or group) who determines who should be responsible (this can be an agent reflecting on their own conduct, but it can also be a judge, the general public, or some other entity);
- the *enforcement mechanism* is the means by which censure or approbation is registered (the forum assigns blame or praise and maybe even punishment based on commonly accepted standards of adjudication).

Various configurations of these components lead to differences in the kinds of responsibility we can expect in particular relationships. We'll see, for example, that the burden of legal responsibility often mischievously shifts back and forth between agent and patient, depending on the context. But let's consider moral responsibility first.

In traditional moral philosophy, moral responsibility is firmly human-centered.[4] Particularly in liberal democracies, human beings are commonly thought to possess autonomy and free will, which is the basis for being held morally responsible. This capacity is what sets us apart from machines and animals. And although human autonomy should definitely be celebrated, there *are* strings attached. Precisely because we get to decide for ourselves—assuming there is no one forcing or coercing us to do otherwise—we *must* live with the consequences of our choices. The flip side of being free to choose as we please is that in the end the buck stops with us.

Autonomy is a necessary condition for the attribution of moral responsibility (we'll unpack it a little more in chapter 7), but there are two other conditions frequently discussed in moral philosophy.[5] One is that there should be a causal connection between the conduct of an agent and the outcome of events; the other is that the agent must be able to foresee the predictable consequences of their actions. To be held morally responsible, a person should knowingly be able to change the outcome of events in the world and thus foresee (to some extent) the likely consequences of their actions. It simply wouldn't be reasonable blaming someone for harm they couldn't have known would result from their actions.

What exactly these three conditions mean and to what extent they properly ground responsibility is subject to continuing debate within moral philosophy. There are plenty of open questions here. Are we *really* free to act or are our actions determined more by nature, nurture, and culture? Do we *really* have control over the outcome of events? To what extent can we know the outcome of events? Yet a common feature of many of these debates is that the agent takes center stage—on this there is little disagreement.

Debates focus on the conditions for an agent to be reasonably held responsible, such as possession of a sound mind. The patient tends to disappear from view or is too easily dismissed as a passive, extraneous factor with no independent bearing. Similarly, the forum is an abstract and inconsequential entity—the moral philosopher, professional opinion, or perhaps the wider moral community.

Certain kinds of legal responsibility have a much broader view of the responsibility relationship. Legal responsibility, generally called "liability," is related to and may overlap with moral responsibility, but it's not quite the same thing. Someone can be legally responsible without being considered morally responsible. In many countries, the driver of a vehicle can be held liable for an accident without clear evidence of moral wrongdoing.[6] Indeed the very word "accident" itself denotes the arbitrary and happenstance character of the event—the fact that no one is truly to blame. For instance, if you're driving along a road and happen to get stung by a bee, swerve, and crash into an oncoming vehicle, it's unlikely anyone would be morally blameworthy if the sting was, let's say, to your eyelid, and unavoidable by taking all reasonable precautions. Conversely, we might consider that social media platforms have a moral responsibility to tackle the problem of fake news, but in most jurisdictions they have no legal responsibility to do so and can't be held liable for harms caused by this form of misinformation.

Liability is to some extent a contrivance designed to regulate the behavior of citizens.[7] As such, it comes in different shapes and sizes depending on the legal system and which behavior it's meant to regulate. Criminal liability, for instance, focuses primarily on the conduct and mental state of the agent (and is thus a little more like moral responsibility), whereas civil liability places relatively more emphasis on the consequences for the patient and the need to distribute the burdens of harm fairly.[8]

Even *within* the area of civil liability the emphasis can shift again—there's no way to predict where the burden will fall just by knowing that a case is a "civil" rather than "criminal" matter. Take *fault-based liability* and *strict liability*. Fault-based liability requires proof that someone did something wrong or neglected to act in a certain way. In order for a person to be held liable, there needs to be evidence of a causal link between the outcome of events and the actions or inactions of the person. We expect manufacturers to ensure that the products they sell are safe and won't cause harm. If a microwave is defective, the manufacturer can be held liable for any harm

caused by the microwave if it can be shown that the manufacturer failed to take reasonable precautions to minimize the risk of harm. However, it can be difficult to establish causal connections between conduct (or omissions) and outcomes, making it hard for victims to be compensated for harm. The burden of the incident will then squarely fall on them, which many people may consider unfair.

Strict liability provides a way to even out the unfairness that fault-based liability may entail. Although strict liability comes in many forms, the general idea is that it doesn't require evidence of fault. Where fault-based liability looks at the conduct and intentions of agents and how much control over the outcome of events they have, strict liability shifts most of its attention onto the patient. For defective products that cause harm, this means that the victim only has to show that the product was defective and that the defect caused harm. The victim doesn't have to prove that the manufacturer was negligent, that is, failed to take reasonable precautions against harm—it's enough to show that the product was defective and caused the victim harm. This kind of liability makes it easier for victims to be compensated and places some of the burden on the actors in the relationship that are better placed to absorb the costs than the victims.

Some forms of strict liability don't even require the identification of individual victims—the patient can be a group, society in general, or even the environment! Think of human rights. As the legal scholar Karen Yeung points out, any interference with human rights, such as freedom of speech, attracts responsibility without proof of fault even if there are no clear victims.[9] By taking the agents, patients, and potentially the whole of society into account, as well as the fair distribution of the burdens of harm, legal responsibility has a much wider gaze than moral responsibility.

Another difference between moral and legal forms of responsibility is that moral responsibility most commonly attaches to individuals, whereas it's not at all unusual to hear of companies or conglomerates being held legally responsible for their misdeeds and liable to pay sometimes multimillion dollar damages or fines for breaches of consumer safety or environmental protection laws.

We should note one final, important difference here too. The sanction itself (e.g., fines, compensation, etc.), as well as the forum that determines the sanction (e.g., a tribunal or local council authority), are both crucially important considerations when establishing legal responsibility. It very

much matters *who* is deciding the outcome and *what* the outcome is. If a company errs, the denunciations of commercial competitors won't carry much weight with the offender. The rulings of a supreme court, on the other hand, surely will. The sanction, too, is paramount in law. Remember that liability tends to be a backward-looking form of responsibility in which agents have to make amends. Although these sanctions are often imposed for breaches of forward-looking duties, the emphasis is on paying up (retribution), *giving* up (disgorgement), or giving *back* (restitution). These features of legal responsibility aren't always present in the moral sphere—the sanction in particular has little relevance, although admittedly, *who* expresses moral disapproval can matter a great deal to the moral wrongdoer. Children, for instance, may be more affected by their teachers' rebukes than by their parents'. Of course, moral disapproval can result in social consequences—ostracism, injury to reputation, and the like—but these "sanctions" aren't usually formal, planned, or structured (the way a fifty-hour community service order is). More often than not, they arise from the instinctive resentments people feel when they've been mistreated.

Technology and Responsibility

The philosopher Carl Mitcham noted that technology and responsibility seem to have co-evolved since the industrial revolution and the rise of liberal democracy.[10] Responsibility filled a gap that was created by the introduction of industrial technologies. These technologies extended human capacities, enabling human beings to do things that they couldn't do before and giving them tremendous power over nature and each other. As the levers of control got longer and the distance between action and consequence greater, discussions grew about how that increasing power could be held in check. Responsibility was a solution: with great power must come great responsibility.

Mitcham's observation underscores the special relationship between technology and responsibility. The introduction of a technology changes human activity in ways that affect the conditions for the attribution of responsibility. If we take moral responsibility as an example here, technologies affect the extent to which human beings are in control or free to act, how their actions contribute to outcomes, and their ability to anticipate the consequences of their actions. Let's take these in turn.

Freedom to Act

Technologies can affect the freedom we have to make informed choices. On the one hand, they can augment our capabilities and broaden the set of options available to us. On the other hand, they frequently constrain these very capabilities and options. The internet affords us a virtually limitless space of opinions and information we can freely indulge. At the same time—and as we'll see in chapters 6 and 7—it's the data collected about us while online that can be parlayed into the targeting algorithms that restrict the kinds of opportunities, opinions, and advertising we're exposed to.

Think also of the automated administrative systems that make it easier to process large numbers of cases efficiently (see chapter 8). These systems, *by design*, reduce the discretionary scope of lower-level bureaucrats to make decisions on a case-by-case basis.[11] Technologies can empower us, to be sure, but they can also rein us in. And the better they are at what they do, the more difficult it becomes to manage without them. They have a way of inducing dependency. Does anyone under the age of twenty-five even *know* what a street map is? Even those of us over twenty-five must admit we'd find life just a little harder without Google Maps on our phones.

Causal Contribution

Technology can obscure the causal connections between the actions a person takes and the eventual consequences. It wouldn't make sense to blame someone for something they had limited control over. Complex technological systems are problematic in this respect, because they often require the collective effort of many different people in order to work. The difficulty of finding responsible individuals among the many who contributed to the development and use of a technological system is known as *the problem of many hands*.[12] It can be a real challenge ascribing responsibility in highly technological environments when there isn't a single individual with complete control or knowledge of how events will turn out. Pilots aren't the only ones needed to keep planes aloft. An aircraft is a staggeringly complex piece of kit that incorporates many different subsystems and personnel. None of the people involved has direct control of what's happening. None has a complete understanding of all the components in the operation. Air traffic controllers, maintenance personnel, engineers, managers, and regulators all have a role to play in ensuring the safe flight of an aircraft. When an accident occurs, it's often the result of an accumulation of minor errors

that on their own might not have turned out disastrously. This isn't to say that no one's responsible at all. Each actor contributed in a small way to the outcome and thus has at least *some* responsibility for what took place, but the cooperative nature of the endeavor often makes it extremely difficult to isolate individual contributions. It's not quite as difficult as isolating the bananas, berries, and kiwis in a blended concoction, but it can be *nearly* as difficult!

Adding to the problem of many hands is the temporal and physical distance that technologies can create between a person and the consequences of their actions. This distance can blur the causal connection between actions and events. Technologies extend the reach of human activity through time and space. With the help of communication technologies people can, for example, interact with others on the other side of the world. Such distances can limit or change how agents experience the consequences of actions and as a result limit the extent to which they feel responsible. Sending a mean tweet may be easier than saying it directly to someone's face because you don't directly see the consequences of your actions. Similarly, the designers of an automated decision-making system will determine ahead of time how decisions should be made, but they'll rarely see how these decisions will impact the individuals they affect—impacts that may only be felt years later.

The connection between the designers' choices and the consequences of their choices is further blurred because the design often doesn't determine exactly how the technology will be used. We, as users, often still get to decide how and when to use these technologies. We can even put them to uses that designers may not have expected. Students with a smart phone no longer need to spend hours in the library hunched over a photocopier. Why bother, when you can take snaps of the relevant pages with your phone? Did the engineers who made smart phone camera functionality a reality ever imagine it would replace a photocopier? Possibly, but it surely wouldn't have occurred to everyone or been uppermost in their minds.

Foreseeing the Consequences of One's Actions

The distancing effect that technologies can have not only enables and constrains human activities, it also mediates our relationship to the future. The various sensors and measurement instruments in a cockpit translate a range of observations such as the altitude or pitch of a plane into numbers and signs that the pilot can use to inform their awareness of the situation. In so doing, these instruments help the pilot better anticipate the consequences of

particular decisions. At the same time, however, the pilot may only have a partial understanding of the mechanisms, assumptions, models, and theories behind the technology being used, and this very opacity can itself compromise a pilot's ability to assess the consequences of particular decisions.

The novelty of a technology can also affect our foresight. It requires knowledge, skill, and experience to operate a technology responsibly and appreciate how it behaves under different conditions. The learning curve is steeper for some technologies than for others. In general, inexperience definitely complicates things.

Again it's important to stress that none of this makes responsibility a mirage. We've managed to create relatively successful practices in which different, complementary forms of responsibility, including legal, moral, and professional responsibility, are distributed across multiple actors despite the problem of many hands. Some of these forms of responsibility don't require direct or full control nor even the ability to foresee likely consequences, and when technologies constrain the autonomy of one individual in a chain, responsibility is often redistributed to other actors higher up in the chain. After all, the effects of technology on human conduct are still a function of the activities of the human agents that created them and made them available. People create and deploy technologies with the objective of producing some effect in the world. The intentions of developers to influence users in particular ways are inscribed within the technology. This kind of control and power necessarily attracts responsibility—although whether this continues to be the case depends on the kind of agency technology can be expected to acquire in the future.

AI and Responsibility

As we've seen, concerns about responsibility in light of technological developments aren't new, yet AI technologies seem to present new challenges. Increasing complexity, the ability to learn from experience, and the seemingly autonomous nature of these technologies suggest that they are qualitatively different from other computer technologies. If these technologies are increasingly able to operate without the direct control or intervention of human beings, it may become hard for software developers, operators, or users to grasp and anticipate their behavior and intervene when necessary. Some suggest that, as these technologies become more complex and autonomous,

we can't reasonably hold human beings responsible when things go wrong. The philosopher Andreas Matthias calls this the "responsibility gap"—the more autonomous technologies become, the less we can hold human beings responsible.[13] Some therefore argue that there may come a point at which we should hold AI technologies responsible and attribute them some kind of legal personhood or moral agency.[14] However, before addressing this suggestion, let's take a closer look at the alleged "responsibility gap."

Underlying the idea of a responsibility gap are several assumptions about responsibility. One assumption is that human beings must have direct control over the outcome of their actions before they can be held responsible for them. However, as we've seen, this narrow notion of responsibility doesn't accurately reflect how we deal with responsibility in many cases, since various notions of responsibility make do without direct control.

Another set of assumptions underlying the idea of a "responsibility gap" has to do with AI itself. AI in these debates is often unhelpfully framed as an independent, monolithic thing possessing certain human-like abilities, but this is misleading in several ways. First, it misses the larger context in which the technology is embedded and the human beings "standing just off stage."[15] In truth, it requires considerable work from human beings for current AI technologies to operate independently. Not only do humans have to design, develop, deploy, and operate these systems, they also have to adjust themselves and their environments to make sure that the technologies work successfully. Self-driving cars aren't independent of the roads they drive on or the interests of the cyclists and pedestrians they come across. These cars operate in environments regulated by rules intended to minimize risks and balance the competing interests of the various actors present in those environments. In short, for an autonomous technology to work, many human beings have to make many decisions, and these all involve choices for which they could be held responsible.

It's also clear that underlying worries about a "responsibility gap" is the assumption that machine autonomy and human autonomy are essentially similar. But there are significant differences between the two. Human autonomy is a complex moral philosophical concept that is intimately tied to ideas about what it means to be human and serves as a basis for various obligations and rights. It assumes that human beings have certain capacities that allow them to make self-willed decisions and that they should be respected for this capacity (see chapter 7).

Machine autonomy, on the other hand, generally refers to the ability of machines to operate for extended periods of time without human intervention. They are delegated tasks to perform without a human operator continuously adjusting the behavior of the system. Such tasks can include navigating an aircraft, driving on the highway, buying stock, or monitoring a manufacturing process. Most of the time, this kind of machine autonomy concerns certain well-defined processes that can be fully automated. Autonomy is then the same as the high-end of a sliding scale of automation and has nothing to do with free will or similar moral philosophical concepts.

Admittedly, the distinction between machine autonomy and human autonomy starts to blur in discussions about AI technologies that learn and adapt to their environments. Systems that automate well-defined processes are relatively predictable, whereas systems that learn from their experiences and exhibit behaviors going beyond their original programming can seem like they have a "mind of their own." AlphaZero, a system designed to play games like chess and Go, is a frequently touted example of a system able to teach itself to play like a pro without explicit human guidance and in a way that is incomprehensible to its human developers. Can human beings still be meaningfully responsible for the behavior of such systems?

Well, despite the impressive feats of a system like this, its actions still take place within the constraints set by its human operators. It takes considerable expertise and effort from human developers to make the system operate properly. They have to carefully construct, adjust, and fine-tune the algorithms and select and prepare its training data.[16] The point is that various human agents are still in a position to exert some kind of control over the system, and with that comes some degree of responsibility. And, of course, it goes without saying that AlphaZero doesn't know that it's playing Go, or that it's being watched, or what a "game" is, or what it means to "play," or what it means to win, and it doesn't do anything other than play the games it was designed to play.

This isn't to say that the development of increasingly powerful AI won't pose real challenges to established ways of allocating responsibility. The increasing complexity of AI technologies and the growing distance between the decisions of developers and the outcomes of their choices will challenge existing ideas about who should be responsible for what and to what extent. And that's very much the point. The issue isn't whether human responsibility will cease to make any sense at all; it's about *which* human

beings should be responsible and *what* their responsibility will look like. With the introduction of a new technology comes a redistribution of tasks, which will cause a shift of responsibility between the different actors in the chain. There may also be many more hands involved than before, as human-machine systems become more complex and larger. Some of these actors may become less powerful and will have less discretionary space, whereas others will gain more leverage. In any case, human beings are still making decisions about how to train and test the algorithms, when to employ them, and how to embed them in existing practices. They determine what is acceptable behavior and what should happen if the system transgresses the boundaries of acceptable behavior.

Figuring out who should be responsible for what and how isn't easy, and there are definitely bad ways of doing it. The opposite worry to the responsibility gap is that the *wrong* people will be held responsible. They become what Madeleine Elish calls a "moral crumple zone." These human actors are deemed responsible even though they have very limited or no control over an autonomous system. Elish argues that, given the complexity of these systems, the media and broader public tend to blame the nearest human operator such as pilots or maintenance staff rather than the technology itself or decision makers higher up in the decision chain.[17]

The advent of AI and the challenges it poses to existing ways of allocating responsibility raise a number of important questions. How can we properly distribute responsibility around increasingly complex technologies? What does it mean to be in control of an autonomous system? If the driver of an autonomous vehicle is no longer in charge of what the car does, on whom (or *what*) should responsibility fall? The distribution of responsibility isn't, after all, something that just happens, like a tree that falls in the woods in a way determined by the forces of nature (gravity, humidity, wind, etc.). It's something that we have to actively negotiate among ourselves, scoring off competing claims, always with an eye to the wider social ramifications of any settlement reached.

Negotiating Responsibilities

If the human driver of an autonomous vehicle is no longer actively controlling its behavior, can they be reasonably held liable, or should liability shift to the manufacturer or perhaps some other actor? If the vehicle's software

learns from the environment and other road users,* could the manufacturer still be held liable for accidents caused by the vehicle? Are the existing rules for liability still applicable or should they be amended? Manufacturers' liability has often been strict. And, interestingly, strict liability was first recognized for the actions of substances (like dangerous chemicals) or chattels (like wandering sheep) that caused harm without the knowledge of their controller—a situation that might be thought directly parallel to the "mindless" but independent actions of machine learning algorithms that learn to classify and execute procedures without being explicitly programmed to do so by their developers. Is strict liability for manufacturers the way to go then?

One answer to the questions raised by the prospect of fully self-driving cars is that we won't need to come up with new laws, as the existing ones—including fault-based and strict liability regimes—will be sufficient to cover them. Manufacturers, it's supposed, are best placed to take precautions against the risks of harm in the construction phase and to warn users of those risks at the point of sale (or wholesale). Product liability rules should therefore apply as usual. No ifs, no buts.

Several legal scholars, however, argue that existing product liability schemes are insufficient and place an undue burden on the victims of accidents. Remember, even with strict product liability, it is the victim who must prove both that the product was defective and that the defect caused their harm. With the growing complexity of computer-enabled vehicles, it will become increasingly difficult for victims to show that the product was defective.[18] Product liability is already used sparingly because it's often difficult to show where an error originated. It may be easier to hold manufacturers liable in cases where it's clear that a particular problem occurs more often in particular designs, as when it came to light that one of Toyota's models would sometimes suddenly accelerate.[19] Even though engineers couldn't pinpoint the source of the malfunction, Toyota accepted liability given the high number of incidents. But in areas where there isn't such a high incidence rate, it can be very difficult to establish the liability of the manufacturer, strict or otherwise.

*Not that any autonomous car so far developed *can* learn on its own. To date, all autonomous vehicles ship from the factory with an algorithm learned in advance. That algorithm will probably have used training data gathered from real cars on real roads, but learning will have happened under the supervision of a team of engineers.

Moreover, like other complex systems, autonomous cars exacerbate the problem of many hands. *Multiple* actors will be involved in the manufacturing and functionality of the vehicle. Although the manufacturer of the actual physical artifact may be primarily in control of the physical artifact, other actors will contribute to the operation of the vehicle, such as various software developers, the owner of the car who needs to keep the software up to date, and the maintenance agency in charge of sensors on the road. This ecosystem of technologies, companies, government departments, and human actors makes it difficult to trace where and why mishaps occurred, especially when they involve technologies that self-learn from their environments.

In light of these difficulties, other legal scholars have championed various forms of compensation schemes that don't require proof of fault or even a human actor. Maurice Schellekens argues that the question of "who is liable for accidents involving self-driving vehicles" might be redundant. He points out that several countries, including Israel, New Zealand, and Sweden, already have no-fault compensation schemes (NFCS) in place for automobile accidents. In these countries, the owner of a car takes out mandatory insurance (or is covered by a more general state scheme that includes personal injury arising from road incidents), and when an accident occurs, the insurer will compensate the victim, even if no one is at fault or even *caused* the accident. Imagine cruising on the highway between Sydney and Canberra and swerving suddenly to avoid a kangaroo that jumped straight in front of your car—not an altogether uncommon occurrence along that stretch of road. Most likely there'd be no one to blame for any ensuing crash (much as in our previous example of the bee sting). An NFCS would at least enable the victim to be compensated quickly for any injuries sustained without the rigamarole of court proceedings.

A similar approach may be appropriate, Schellekens suggests, for accidents involving autonomous vehicles. Whether manufacturers will be incentivized to design safe vehicles under such a regime will depend on the rights of victims or insurers to pursue manufacturers for any defects that may have caused these incidents. For instance, a victim may be compensated by their insurer, but if faulty design was the issue, the insurer should acquire the victim's rights to sue the manufacturer in their place, or the victim themselves may pursue the manufacturer for compensation above the statutory (NFCS) limit. Another option would be to force *manufacturers* to pay the cost of insurance, rather than private citizens. Either way, it's

important that manufacturers be held responsible for the harms caused by their designs, otherwise they'll have little incentive to improve them.

Morally and Legally Responsible AI?

The challenges that AI poses to existing ways of distributing responsibility have led some to suggest that we should rethink who or what we consider to be the responsible agent. Should this always be a human agent, or can we extend the concept of agency to include nonhuman agents? As we noted earlier, conventional moral philosophical notions of responsibility are roundly individualistic and anthropocentric. Only humans can be morally responsible. Legal conceptions of responsibility, on the other hand, allow for more flexibility here, as the agent doesn't necessarily have to be an individual human being, nor do they have to be in direct control of events.

In view of the advent of AI, some philosophers have argued that the human-centered conception of moral responsibility is outdated.[20] The complexity of certain kinds of software and hardware demands a different approach, one where artificial agents can be dealt with directly when they "behave badly." Other philosophers have also argued that if these technologies become complex and intelligent enough, they could well be attributed moral agency.[21]

Critics of this suggestion have countered that such an approach diminishes the responsibility of the people that develop and deploy autonomous systems.[22] They are human-made artifacts and their design and use reflect the goals and ambitions of their designers and users. It's through human agency that computer technology is designed, developed, tested, installed, initiated, and provided with instructions to perform specified tasks. Without this human input, computers could do nothing. Attributing moral agency to computers diverts our attention away from the very forces that shape technology to behave as it does.

One response to this would be to argue that, although technology on its *own* might not have moral agency, moral agency itself is also hardly ever "purely" human anyway.[23] According to Peter-Paul Verbeek, for example, human action is a composite of different forms of agency at work: the agency of the human performing the action; the agency of the designer who helped shape the mediating role of the artifact; and the artifact itself mediating between human actions and their consequences.[24] Whenever

technology is used, moral agency is rarely concentrated in a single person; it is diffused in a complex mash-up of humans and artifacts.

In legal scholarship, a similar issue has been discussed in terms of personhood. Given the complexity of AI and digital ecosystems, some legal scholars as well as policy makers have suggested that artificial agents should be attributed some kind of personhood and be held liable the way corporations may be. In 2017, the European Parliament even invited the European Commission to consider the creation of a specific legal status that would establish electronic personhood for sufficiently sophisticated and autonomous robots and AI systems.

In law, however, the idea of nonhuman personhood is hardly exotic, as many legal systems already recognize different kinds of personhood.[25] Legal personhood is a legal fiction used to confer rights and impose obligations on certain entities like corporations, animals, and even rivers like the Ganges and Yamuna in India, the Whanganui in New Zealand, or whole ecosystems in Ecuador. Needless to say, the reasons for granting such status to these kinds of entities are entirely different from those justifying our view of each other as persons. Human beings are assigned rights and obligations based on moral considerations about, for example, their dignity, intrinsic worth, consciousness, autonomy, and ability to suffer. Where nonhuman entities are afforded personhood, it is for a great range of reasons, such as sharing certain human characteristics like the ability to suffer in the case of animals, or for economic reasons in the case of corporations. Many legal systems recognize corporations as persons to reduce the risk associated with individual liability and thereby encourage innovation and investment. AI technologies could perhaps be considered a kind of corporation. Indeed, an NFCS, extended (if necessary) to cover harms arising from the use of autonomous systems, is arguably already halfway there.

Looking Ahead

We've seen how being morally responsible for your actions goes along with *having* control over them, in some sense. Much the same is true when we use technology to achieve our ends. Being responsible for the outcomes of our actions when they have been facilitated by technology requires that we have some degree of control over that technology. This is because when we work with technology, the technology becomes a part of us—an extension

of our arms and legs and minds. Although we've said a little about control in this chapter, our next chapter will elaborate on some of the ways in which technology inherently *resists* being controlled by us. This is important because the more control an individual has over a system, the more we can hold them responsible for the results of its deployment. On the other hand, the less meaningful and effective that control, the weaker our authority to blame them when things go wrong.