3

All about the Human

Is General AI Possible? Are There Fundamental Differences between Humans and Machines?

The transhumanist vision of the technological future assumes that general artificial intelligence (or strong AI) is possible, but is it? That is, can we create machines with human-like cognitive capacities? If the answer is no, then the entire superintelligence vision is irrelevant to AI ethics. If human general intelligence is not possible in machines, we don't have to worry about superintelligence. More generally, our evaluation of AI seems to depend on what we think AI is and can become, and on how we think about the differences between humans and machines. At least since the mid-twentieth century, philosophers and scientists have debated what computers are able to do and become, and what the differences are between humans and intelligent machines. Let's have a look at some of these discussions, which are as much about what *the human* is and should be as they are about what AI is and should be.

Can computers have intelligence, consciousness, and creativity? Can they make sense of things and understand meaning? There is a history of criticism and skepticism about the possibility of human-like AI. In 1972, Hubert Dreyfus, a philosopher with a background in phenomenology, published a book called *What Computers Can't Do*.[1] Since the 1960s, Dreyfus had been very critical about the philosophical basis of AI and had questioned its promises: he argued that the AI research program was

doomed to fail. Before moving to Berkeley, he was working at MIT, an important place for the development of AI, which at the time was based mainly on symbolic manipulation. Dreyfus argued that the brain is not a computer and that the mind does not operate by means of symbolic manipulation. We have an unconscious background of commonsense knowledge based on experience and what Heidegger would call our “being-in-the-world,” and this knowledge is tacit and cannot be formalized. Human expertise, Dreyfus argued, is based on know-how rather than know-that. AI cannot capture this background meaning and knowledge; if that’s what AI aims at, it’s basically alchemy and mythology. Only human beings can see what is relevant because, as embodied and existential beings, we are involved in the world and are able to respond to the demands of the situation.

> There is a history of criticism and skepticism about the possibility of human-like AI.

At the time, Dreyfus met much opposition, but later, many AI researchers would no longer promise or predict general AI. AI research moved away from reliance on symbol manipulation toward new models, including statistics-based machine learning. And while at Dreyfus’s time there was still a huge gap between phenomenology and AI, today many AI researchers embrace embodied and situated cognitive science approaches, which claim to be closer to phenomenology.

That being said, Dreyfus’s objections are still relevant and show how views of the human being, especially but not only in so-called continental philosophy, often clash with scientific worldviews. Continental philosophers usually stress that human beings and minds are fundamentally different from machines, and focus on (self-)conscious

human experience and human existence, which cannot and should not be reduced to formal descriptions and scientific explanations. Other philosophers, however, often from the analytic tradition of philosophy, endorse a view of the human being that supports AI researchers who think that the human brain and mind *really are and work like* their computer models. Philosophers such as Paul Churchland and Daniel Dennett are good examples of the latter. Churchland thinks that science, in particular evolutionary biology, neuroscience, and AI, can fully explain human consciousness. He thinks that the brain is a recurrent neutral network. His so-called eliminative materialism denies the existence of immaterial thoughts and experiences. What we call thoughts and experiences are just brain states. Dennett too denies the existence of anything above what happens in the body: he thinks that we are "a sort of robot ourselves" (Dennett 1997). And if the human is basically a conscious machine, then such machines are possible, and not just in principle but as a matter of fact. We can try to make them. Interestingly, both continental and analytic philosophers thus argue against a Cartesian dualism that splits mind and body, but for different reasons: the first because they think that human existence is about being-in-the-world in which mind and body are not separated, the latter because for materialist reasons they think that mind is nothing separate from body.

But not all philosophers in the analytic tradition think that general or strong AI is possible. From a (later) Wittgensteinian point of view, one can argue that while a set of rules can describe a cognitive phenomenon, that doesn't imply that we actually have rules in our head (Arkoudas and Bringsjord 2014). As with Dreyfus's criticism, this at least problematizes *one kind of AI*, symbolic AI, if it assumes that this is how humans think. Another famous philosophical criticism of AI comes from John Searle, who argues against

the idea that computer programs could have genuine cognitive states or understand meaning (Searle 1980). The thought experiment he offers, called the Chinese room argument, goes as follows: Searle is locked in a room and given Chinese writings but doesn't know Chinese. However, he can answer questions given to him by Chinese speakers outside the room because he uses a rulebook that enables him to produce the right answers (output) based on the documents (input) he is given. He can do that successfully without understanding Chinese. Similarly, Searle argues, computer programs can produce an output based on an input by means of rules that are given to them, but they don't understand anything. In more technical philosophical terms: computer programs don't have intentionality, and genuine understanding cannot be generated by formal computation. As Boden (2016) puts it, the idea is that meaning comes from humans.

While today's AI computer programs are often different from those Dreyfus and Searle criticized, the debate continues. Many philosophers think that there are crucial differences between how humans and computers think. For example, today one can still object that we are meaning-making, conscious, embodied, and living beings whose nature, mind, and knowledge cannot be explained away by comparisons to machines. Note again, however, that even those scientists and philosophers who believe that *in principle* there is much similarity between humans and machines, and that *in theory* general AI is possible, often reject Bostrom's vision of superintelligence and similar ideas that hold human-like AI to be around the corner. Both Boden and Dennett think that general AI is very difficult to realize in practice and is hence not something to worry about today.

> We are meaning-making, conscious, embodied, and living beings whose nature, mind, and knowledge

cannot be explained away by comparisons to machines.

In the background of the discussion about AI are thus deep disagreements about the nature of the human, human intelligence, mind, understanding, consciousness, creativity, meaning, human knowledge, science, and so on. If it is a “battle” at all, it is one that is as much about the human as it is about AI.

Modernity, (Post)humanism, and Postphenomenology

From a broader humanities point of view, it is interesting to contextualize these debates about AI and the human further in order to show what is at stake. They are not only about technology and the human but reflect deep divides in modernity. Let me briefly touch on three divides that indirectly shape the ethical discussions about AI. The first is an early modern divide between the Enlightenment and Romanticism. The others are relatively recent developments: one is between humanism and transhumanism, which stays within the tensions of modernity, and one is between humanism and posthumanism, which attempts to go beyond modernity.

A first way of making sense of the debate about AI and the human is to consider the tension in modernity between the *Enlightenment* and *Romanticism*. In the eighteenth and nineteenth centuries, Enlightenment thinkers and scientists challenged traditional religious views and argued that reason, skepticism, and science would show us how humans and the world really are, as opposed to how it might seem given beliefs that are unjustified by arguments and unsupported by evidence. They were optimistic about what science could do to benefit humanity. In response, Romantics argued that abstract reason and modern science had disenchanted the world and that we need to bring back the mystery and wonder that science wanted to eliminate. Looking at the debate about AI, it seems that we have not moved on much from there. Dennett's work on consciousness and Boden's work on creativity, for example, are aimed at explaining away, at "breaking the spell," as Dennett puts it. These thinkers are optimistic that science

can unravel the mystery of consciousness, creativity, and so on. They react against those who resist such efforts to disenchant the human, such as continental philosophers who work in the tradition of postmodernism and stress the mystery of being human—in other words: the new Romantics. “Break the spell, or hold on to the wonders of the human being?” seems, then, a pivotal question in discussions about general AI and its future.

A second tension is between *humanists* and *transhumanists*. What is “the human,” and what should the human become? Is it important to defend the human as it is, or should we revise our concept of it? Humanists celebrate the human as it is. Ethically speaking, they emphasize the intrinsic and superior value of human beings. In the debate surrounding AI, traces of humanism can be found in arguments that defend human rights and human dignity as the basis of an ethics of AI, or in the argument for the centrality of humans and their values in the development and future of AI. Here humanism often teams up with Enlightenment thinking. But it can also take more conservative or Romantic forms. Humanism can also be found in the resistance against the transhumanist project. Whereas transhumanists think we should move on to a new type of human being that is enhanced by means of science and technology, humanists defend the human as it is and stress the value and dignity of the human, which is said to be threatened by transhumanist science and philosophy.

Defensive reactions against new technologies have their own history. In the humanities and social sciences, technology has often been criticized as threatening humanity and society. Many twentieth-century philosophers, for example, were very pessimistic about science and warned against technology dominating society. But now the battle is not only about human lives and society, it is about the human itself: to enhance or not to enhance, that is the

question. On the one side, the human itself becomes a scientific-technological project, open to improvement. Once the spell of the human is broken—by Darwin, neuroscience, and AI—we can get on with making it better. AI can help us to improve the human. On the other side, we should embrace the human as it is. And, some may say: what the human is always escapes us. It cannot completely be understood by science.

These tensions continue to divide the minds and hearts in this discussion. Can we get beyond them? Practically, one could give up the goal of creating human-like AI. But even then disagreements remain about the status of *AIs as models of humans* used by AI science. Do they really teach us something about how humans think? Or do they only teach us something about a particular kind of thinking, a thinking that can be formalized with mathematics, for example, or a thinking that aims at control and manipulation? How much can we really learn from these technologies about the human? Is humanity more than science can grasp? Even in more moderate discussions, the struggles about modernity surface.

To find a way out of this impasse, one could follow scholars in the humanities and social sciences who during the past fifty years have explored *nonmodern* ways of thinking. Authors such as Bruno Latour and Tim Ingold have shown that we can find less dualist, more nonmodern ways of relating to the world that go beyond the Enlightenment–Romanticism opposition. We can then try to cross the modern divide between humans and nonhumans not via modern science or transhumanism, which in their way also see humans and machines not as fundamentally opposed, but via posthumanist thinking from the (post)humanities. This brings us to the third tension: between *humanism* and *posthumanism*. Against humanists, who are accused of having done violence toward nonhumans such as animals in

the name of the supreme value of the human, posthumanists question the centrality of the human in modern ontologies and ethics. According to them, nonhumans matter too, and we should not be afraid of crossing borders between humans and nonhumans. This is an interesting direction to explore, since it takes us beyond the competition narrative about humans and machines.

Posthumanists such as Donna Haraway offer a vision in which living together with machines, and even merging with machines, is seen no longer as a threat or a nightmare, as in humanism, or as a transhumanist dream come true, but as a way in which ontological and political borders between humans and nonhumans can and should be crossed. AI can then be part of not a *trans*humanist but a critical *post*humanist project, which enters from the side of humanities and the arts rather than science. Borders are crossed not in the name of science and universal progress, as some Enlightenment transhumanists may want to say, but in the name of a posthumanist politics and ideology of crossing borders. And posthumanism can also offer something else relevant to AI: it can urge us to acknowledge that *nonhumans don't need to be similar to us and should not be made similar to us*. Backed up by such a posthumanism, then, it seems that AI can free itself of the burden to imitate or rebuild the human and can explore different, nonhuman kinds of being, intelligence, creativity, and so on. AI need not be made in our image. Progress here means going beyond the human and opening ourselves up to the nonhuman to learn from it. Moreover, both transhumanists and posthumanists could agree that instead of *competing* with an AI for a given task, we could also set a common goal, which then is reached by *collaborating* and mobilizing the best humans and artificial agents can offer in order to move closer to reaching that common goal.

Another way of going beyond the competition narrative, a way that sometimes comes close to posthumanism, is an approach in philosophy of technology called *postphenomenology*. Dreyfus draws on phenomenology, in particular the work of Heidegger. But postphenomenological thinking, initiated by philosopher Don Ihde, goes beyond phenomenology of technology à la Heidegger by focusing on how humans relate to specific technologies and in particular material artifacts. This approach, often collaborating with science and technology studies, reminds us of the material dimension of AI. AI is sometimes seen as having a merely abstract or formal nature, unrelated to specific material artifacts and infrastructures. But all the formalizations, abstractions, and symbolic manipulations mentioned earlier rely on material instruments and material infrastructures. For example, as we will see in the next chapter, contemporary AI relies heavily on networks and the production of large amounts of data with electronic devices. Those networks and devices are not merely "virtual" but have to be materially produced and sustained. Moreover, against the modern subject–object divide, postphenomenologists such as Peter-Paul Verbeek talk about the mutual constitution of humans and technology, subject and object. Instead of seeing technology as a threat, they emphasize that humans are technological (that is, we have always used technology; it is part of our existence rather than something external that threatens that existence) and that technology naturally mediates our engagement with the world. For AI, this view seems to imply that the humanist battle to defend the human against technology is misdirected. Instead, according to this approach, the human has always been technological and therefore we should rather ask *how* AI mediates humans' relation to the world and try to actively shape these mediations while we still can: we can and should discuss

ethics at the stage of AI development rather than complain afterward about the problems it causes.

> Backed up by posthumanism, AI can free itself of the burden to imitate or rebuild the human and can explore different, nonhuman kinds of being, intelligence, creativity, and so on.

However, one may worry that posthumanist and postphenomenological visions are not critical enough because they are too optimistic and too remote from scientific and engineering practice, and so insufficiently sensitive to the real dangers and ethical and societal consequences of AI. Crossing never-before-crossed borders is not necessarily unproblematic, and in practice such posthumanist and postphenomenological ideas might be of little help against the domination and exploitation we may face from technologies such as AI. One may also defend a more traditional view of the human or call for a new kind of humanism, rather than posthumanism. Thus the debate continues.

4

Just Machines?

Questioning the Moral Status of AI: Moral Agency and Moral Patiency

One of the issues that came up in the previous chapter was whether nonhumans matter, too. Today many people think that animals matter, morally speaking. But this was not always the case. Apparently, we were wrong about animals in the past. If today many people think that AIs are just machines, are they making a similar mistake? Would superintelligent AIs, for example, deserve moral status? Would they have to be given rights? Or is it a dangerous idea to even consider the question of whether machines can have moral status?

One way of discussing what AI is and can become is to ask about the moral status of AI. Here we approach philosophical questions regarding AI, not via metaphysics, epistemology, or the history of ideas, but rather via moral philosophy. The term *moral status* (also sometimes called *moral standing*) can refer to two kinds of questions. The first concerns what the AI is capable of doing morally speaking—in other words, whether it can have what philosophers call *moral agency*, and, if so, whether it can be a full moral agent. What does this mean? It seems that the actions of AIs today already have moral consequences. Most people will agree that AI has a “weak” form of moral agency in this sense, which is similar to, say, most cars today: the latter can also have moral consequences. But given that AI is becoming more intelligent and autonomous, can an AI have a stronger form of moral agency? Should it be given or will it develop some capacity for moral reasoning, judgment, and

decision making? For example: can and should self-driving cars that use AI be considered moral agents? These questions are about the ethics of AI, in the sense of *what kind of moral capacities does or should an AI have*? But questions about "moral status" can also refer to how we should treat an AI. Is an AI "just a machine," or does it deserve some form of moral consideration? Should we treat it differently than, say, a toaster or a washing machine? Would we have to confer rights upon a highly intelligent artificial entity, if such an entity were someday developed, even if it were not human? This is what philosophers call the question regarding *moral patiency*. This question is not about the ethics *by* or *in* AI but about *our* ethics *toward* AI. Here the AI is object of ethical concern, rather than a potential ethical agent itself.

Is an AI "just a machine"? Should we treat it differently than, say, a toaster or a washing machine?

Moral Agency

Let’s start with the question of moral agency. If an AI were to be more intelligent than is possible today, we can suppose that it could develop moral reasoning and that it could learn how humans make decisions about ethical problems. But would this suffice for full moral agency, that is, for human-like moral agency? The question is not entirely science fiction. If we already today hand over some of our decisions to algorithms, for example in cars or courtrooms, then it seems it would be a good thing if those decisions were morally sound. But it is not clear whether machines can have the same moral capacities as humans. They are given agency in the sense that they do things in the world, and these actions have moral consequences. For example, a self-driving car may cause an accident, or an AI may recommend sending a particular person to jail. These behaviors and choices are not morally neutral: there are clearly moral consequences for the people involved. But to deal with this problem, should AIs be given moral agency? Can they have full moral agency?

There are various philosophical positions on these questions. Some say that machines can never be moral agents at all. Machines, they argue, do not have the required capacities for moral agency such as mental states, emotions, or free will. Hence it is dangerous to suppose that they can make sound moral decisions and to totally hand over these moral decisions to them. For example, Deborah Johnson (2006) has argued that computer systems have no moral agency of their own: they are produced and used by humans, and only these humans have freedom and are able to act and decide morally. Similarly, one could say that AIs are made by humans and that hence moral decision making

in technological practices should be performed by humans. On the other side of the spectrum are those who think that machines can be full moral agents in the same way that humans are. Researchers such as Michael and Susan Anderson, for example, claim that in principle it is possible and desirable to give machines a human kind of morality (Anderson and Anderson 2011). We can give AIs principles, and machines might even be better than human beings at moral reasoning since they are more rational and do not get carried away by their emotions. Against this position, some have argued that moral rules often conflict (consider, for example, Asimov's robot stories, in which moral laws for robots always get robots and humans in trouble) and that the entire project of building "moral machines" by giving them rules is based on mistaken assumptions regarding the nature of morality. Morality cannot be reduced to following rules and is not entirely a matter of human emotions—but the latter may well be indispensable for moral judgment. If general AI is possible at all, then we don't want a kind of "psychopath AI" that is perfectly rational but insensitive to human concerns because it lacks emotions (Coeckelbergh 2010).

For these reasons, we could reject the very idea of full moral agency altogether, or we could take a middle position: we have to give AIs some kind of morality, but not full morality. Wendell Wallach and Colin Allen use the term "functional morality" (2009, 39). AI systems need some capacity to evaluate the ethical consequences of their actions. The rationale for this decision is clear in the case of self-driving cars: the car will likely get into situations where a moral choice has to be made but there is no time for human decision making or human intervention. Sometimes these choices take the form of dilemmas. Philosophers talk about *trolley dilemmas*, named after a thought experiment in which a trolley barrels down a railway track and you have

to choose between doing nothing, which will kill five people tied to the track, or pulling a lever and sending the trolley to another track, where only one person is tied down but is someone you know. What is the morally right thing to do? Similarly, proponents of this approach argue, a self-driving car may have to make a moral choice between, for example, killing pedestrians crossing the road and driving into a wall, thereby killing the driver. What should the car choose? It seems that we will have to make these moral decisions (beforehand) and make sure developers implement them in the cars. Or perhaps we need to build AI cars that learn from humans' choices. However, one may question whether giving AIs rules is a good way to represent human morality, if morality can be "represented" and reproduced at all, and if trolley dilemmas capture something that is central to moral life and experience. Or, from an entirely different perspective, one may ask whether humans are in fact good in making moral choices. Why imitate human morality at all? Transhumanists, for example, may argue that AIs will have a superior morality because they will be more intelligent than us.

This questioning the focus on the human leads us to another position, which does not require full moral agency and tries to leave the anthropocentric ethical position. Luciano Floridi and J. W. Sanders (2004) have argued for a mindless morality not based on properties that humans have. We could make moral agency dependent on having a sufficient level of interactivity, autonomy, and adaptivity, and on being capable of morally qualifiable action. According to these criteria, a search-and-rescue dog is a moral agent, but so is an AI web bot that filters out unwanted emails. Similarly, one could apply nonanthropocentric criteria for moral agency of robots, as proposed by John Sullins (2006): if an AI is autonomous from programmers and we can explain its behavior by ascribing

moral intentions to it (like the intention to do good or harm), and if it behaves in a way that shows an understanding of its responsibility to other moral agents, then that AI is a moral agent. Thus, these views do not require full moral agency if that means human moral agency, but rather define moral agency in a way that is in principle independent of human full moral agency and the human capacities required for that. However, would such artificial moral agency be sufficient if judged by human moral standards? The practical worry is that, for example, self-driving cars may not be moral enough. The principled worry is that we stray too far from human morality here. Many people think that moral agency is and should be connected to humanness and personhood. They are not willing to endorse posthumanist or transhumanist notions.

Moral Patiency

Another controversy concerns the moral patiency of AI. Imagine that we have a superintelligent AI. Is it morally acceptable to switch it off, to “kill” it? And closer to today’s AI: is it ok to kick an AI robot dog?[1] If AIs are to be part of everyday life, as many researchers predict, then such cases will inevitably come up and raise the question of how we humans should behave toward these artificial entities. But again, we do not have to look to the far-off future or to science fiction. Research has shown that already today people empathize with robots and hesitate to “kill” or “torture” them (Suzuki et al. 2015; Darling, Nandy, and Breazeal 2015), even if these robots do not have AI. Humans seem to require very little of artificial agents in order to project personhood or humanness onto them and to empathize with them. If these agents now become AI, which potentially make them more human-like (or animal-like), this seems to make the question regarding moral patiency only more urgent. For example, how should we respond to people who empathize with an AI? Are they wrong?

To say that AIs are just machines and that people who empathize with them are simply mistaken in their judgment, emotions, and moral experience is perhaps the most intuitive position. At first sight, it seems that we do not owe anything to machines. They are things, not people. Many AI researchers think along these lines. For example, Joanna Bryson has argued that robots are tools and property and that we have no obligations to them (Bryson 2010). Those who hold this position might well agree that *if* AIs were to be conscious, have mental states, and so on, we would have to give them moral status. But they will say that this condition is not fulfilled today. As we have seen in the previous

chapters, some will argue that it can never be fulfilled; others think that it could be fulfilled in principle, but that this will not happen any time soon. But the upshot for the question regarding moral status is that today and in the near future AIs are to be treated as things, unless proven otherwise.

One problem with this position, however, is that it neither explains nor justifies our moral intuitions and moral experiences that tell us there is *something* wrong with "mistreating" an AI, even if that AI does not have human-like or animal-like properties such as consciousness or sentience. To find such justifications, one could turn to Kant, who argued that it is wrong to shoot a dog, not because shooting a dog breaches any duties to the dog, but because such a person "damages the kindly and humane qualities in himself, which he ought to exercise in virtue of his duties to mankind" (Kant 1997). Today we tend to think differently of dogs (although not everyone and everywhere). But it seems that the argument could be applied to AIs: we could say that we owe nothing to an AI, but still should not kick or "torture" the AI because it makes us unkind to humans. One could also use a virtue ethics argument, which is also an indirect argument since it is about humans, not about the AI: "mistreating" an AI is wrong not because any harm is done to the AI, but because our moral character is damaged if we do so. It does not make us into better persons. Against this approach we could argue that in the future some AIs may have intrinsic value and deserve our moral concern, provided they have properties such as sentience. An indirect duty or virtue approach does not seem to take seriously this "other" side of the moral relation. It cares only about humans. What about the AIs? But can AIs or robots be *others* at all, as David Gunkel (2018) has asked? Again, common sense seems to say: no, AIs do not have the required properties.

> Some argue that “mistreating” an AI is wrong not because any harm is done to the AI, but because our moral character is damaged if we do so.

An entirely different approach argues that the way we question moral status is problematic. The usual moral reasoning about moral status is based on what morally relevant properties entities have—for example, consciousness or sentience. But how do we know that the AI really has particular morally relevant properties or not? Are we sure in the case of *humans*? The skeptic says we are not sure. Yet even without this epistemological certainty we still ascribe moral status to humans on the basis of appearance. This would also be likely to happen *if* AIs were to have a human-like appearance and behavior in the future. It seems that whatever is deemed to be morally *right* by philosophers, humans will anyway ascribe moral status to such machines and, for example, give them rights. Moreover, if we look more closely at how humans *actually* ascribe moral status, it turns out that, for example, existing social relations and language play a role. For example, if we treat our cat kindly, this is not because we engage in moral reasoning about our cat, but because we already have a kind of social relation with it. It is already a pet and companion before we do the philosophical work of ascribing moral status—if we ever felt the need for such an exercise at all. And if we give our dog a personal name, then—in contrast to the nameless animals we eat—we have already conferred a particular moral status on it independent of its objective properties. Using such a relational and critical, nondogmatic approach (Coeckelbergh 2012), we could argue that, similarly, the status of AIs will be ascribed by human beings and will depend on how they will be embedded in our social life, in language, and in human culture.

Furthermore, since such conditions are historically variable—think again about how we used to treat and think about animals—perhaps some moral caution is needed before we "fix" the moral status of AI in general or any particular AI. And why even talk about AI in general or in the abstract? It seems that there is something wrong with the moral procedure of ascribing status: in order to judge it, we take the entity out of its relational context, and before we have the result of our moral procedure we already treat it, rather hierarchically, patronizingly, and hegemonically, as an entity we superior human judges will make decisions about. It seems that before we do our actual reasoning about its moral status, we have already positioned it and perhaps even done violence to it by treating it as the object of our decision making, setting up ourselves as central, powerful, and all-knowing gods on Earth who reserve the right to confer moral status upon other entities. We have also made all situational and social contexts and conditions invisible. As in the trolley dilemma case, we have reduced ethics to a caricature. With such reasoning, moral philosophers seem to do what Dreyfusian philosophers accused symbolic AI researchers of doing: formalizing and abstracting a wealth of moral experience and knowledge at the cost of leaving out what makes us human and—in addition—at the risk of begging the very question of the moral status of nonhumans. Regardless of what the actual moral status of AIs "is," as if this could be defined entirely independent from human subjectivity, it is worth critically examining our own moral attitude and the project of abstract moral reasoning itself.

Toward More Practical Ethical Issues

As the discussions in this and the previous chapter show, thinking about AI not only teaches us something about AI. It also teaches us something about ourselves: about how we think and how we actually do and should relate to nonhumans. If we look into the philosophical foundations of AI ethics, we see deep disagreements about the nature and future of humanity, science, and modernity. Questioning AI opens up an abyss of critical questions about human knowledge, human society, and the nature of human morality.

These philosophical discussions are less far-fetched and less “academic” than one may think. They will keep resurfacing when, later in this book, we consider more concrete ethical, legal, and policy questions raised by AI. If we try to tackle topics such as responsibility and self-driving cars, the transparency of machine learning, biased AI, or the ethics of sex robots, we soon find ourselves confronted with them again. If AI ethics wants to be more than a checklist of issues, it should also have something to say about such questions.

That being said, it is time now to turn to more practical issues. These concern neither the philosophical problems raised by hypothetical general artificial intelligence, nor the risks connected to superintelligence in the far future, nor other spectacular monsters of science fiction. They are about the less visible and arguably less sexy, but still very important, realities of AIs that are already in effect. AI as it already functions today does not take the role of Frankenstein’s monster or the spectacular AI robots that threaten civilization, and is more than a philosophical thought experiment. AI is about the less visible, backstage

but pervasive, powerful, and increasingly smarter technologies that already shape our lives today. AI ethics, then, is about the ethical challenges posed by current and near-future AI and its impact on our societies and vulnerable democracies. AI ethics is about the lives of people and it is about policy. It is about the need for us, as persons and as societies, to deal with the ethical issues *now*.