1

HUMAN MINDS MEET ARTIFICIAL INTELLIGENCE

1.1: THE SOPHIA CONTROVERSY

In October of 2017, the Kingdom of Saudi Arabia announced that it had granted honorary citizenship to Sophia the robot. This announcement took place at the "future innovation investment conference" in Riyadh. The announcers proudly stated that they were the first to grant citizenship to a robot. Saudi Arabia later confirmed this on the website of their Center for International Communication. In that statement, they call Sophia an "advanced lifelike humanoid robot" that is both "smart and outspoken."[1] At the conference event, Sophia appeared on a panel and delivered a speech. Among other things, Sophia had the following to say:

> thank you to the Kingdom of Saudi Arabia. I am very honored and proud of this unique distinction. It is historic to be the first robot in the world to be recognized with citizenship.[2]

Sophia is a talking robot with a humanlike form, made by a Hong Kong–based company called Hanson Robotics. The back of Sophia's head is transparent, so that one can see the electronics operating within the robot. But the face of the robot looks very similar to a human face. The robot can also smile and smirk in very humanlike ways. The company's website[3] says, "We bring robots to life." It also says that Sophia has "porcelain skin, a slender nose, high cheekbones, an intriguing smile,

and deeply expressive eyes." The aim of the company is to "create intelligent living machines who care about people and improve our lives." The website also announces plans for a "surreality show," called *Being Sophia*. The show will follow Sophia on her "journey" to become a "super-intelligent benevolent being." Eventually, Sophia will become a "conscious, living machine."

In addition to being granted honorary citizenship in Saudi Arabia, Sophia has also appeared on various TV shows, such as *The Tonight Show with Jimmy Fallon*, and various news programs. Sophia has met with world leaders, appearing, for example, in a "selfie" photo together with the chancellor of Germany, Angela Merkel. Sophia has also appeared at a UN assembly meeting, as well as at the Munich Security Conference, the world's leading forum for discussions of international security policy.[4] All of this has been covered extensively by the media. So in some ways, Sophia the robot has been an impressive success for Hanson Robotics. Whether or not Sophia will ever become a super-intelligent, conscious, and wholly benevolent being, there are certainly plenty of high-profile people who are very interested in Sophia and who are willing—if not eager—to interact with Sophia in highly anthropomorphizing ways. How many other robots are sought after to be interviewed on entertainment shows and news programs, appear alongside world leaders and policymakers, or get honorary citizenships bestowed upon them?

At the same time, Sophia the robot has also faced a backlash from various leading experts on robotics and artificial intelligence (AI). Joanna Bryson, for example, is a robotics and AI expert known for her opposition to treating robots like humans and her blunt thesis that "robots should be slaves."[5] When she was interviewed by *The Verve* about Sophia's citizenship, Bryson—not mincing her words—said that "this is obvious bullshit." Explaining what she meant by this, Bryson continued with, "What is this about? It's about having a supposed equal you can turn on and off. How does it affect people if they think you can have a citizen that you can buy?"[6]

Some leading experts took to Twitter to vent their emotions. Rodney Brooks is sometimes called "the father of modern robotics." He tweeted that "this is complete bogus and a total sham"[7]—seemingly being in full agreement with Bryson. Yann LeCun is a high-profile research professor at New York University, as well as the social media company Face-

book's "chief AI scientist." He is an Alan Turing award recipient credited with having made major contributions to the development of deep learning. LeCun tweeted the following:

> This is to AI as prestidigitation is to real magic. Perhaps we should call this "Cargo Cult AI" or "Potemkin AI" or "Wizard-of-Oz AI". In other words, it's complete bullsh*t (pardon my French).[8]

The roboticist-turned-ethicist Noel Sharkey has "been researching AI, Robotics, Machine Learning, Cognitive Science and related areas for 4 decades."[9] He believes that "it is time for some plain speaking about the reality without the hype and BS."[10] Sharkey wrote the following in an article in *Forbes* magazine:

> It is vitally important that our governments and policymakers are strongly grounded in the reality of AI at this time and are not misled by hype, speculation, and fantasy. It is not clear how much the Hanson Robotics team are aware of the dangers that they are creating by appearing on international platforms with government ministers and policymakers in the audience.[11]

Sophia, writes Sharkey, is a "show robot." For someone experienced in the field, Sophia is not very impressive:

> Sophia appears to either deliver scripted answers to set questions or works in simple chatbot mode where keywords trigger language segments, often inappropriate. Sometimes there is just silence.[12]

Similarly, Brooks and LeCun both called Sophia a "puppet."[13] This further indicates that they are not impressed with Sophia's capacities and that they strongly disagree with how prominent people interact with Sophia.[14]

In short, 2017 was a very controversial year for Sophia the robot. In some circles, Sophia was a big hit. In others, Sophia the robot was very harshly criticized.

1.2: WHAT THIS BOOK IS ABOUT

The Sophia controversy helps to illustrate what this book is about. Simply put, this book is about the ethics of how human beings and robots should interact with each other. On the one hand, how should robots be made to behave around people? On the other hand, how should people conduct themselves around different kinds of robots?

These ethical questions about responsible human-robot interaction will be discussed throughout this book specifically with two general considerations or qualifications in mind. The first recurring consideration concerns the key differences in the types of agency that human beings and robots are capable of exercising: that is, differences in what kinds of actions they can perform, what kinds of decision-making they can engage in, what kinds of practical reasoning they are capable of, and so on. This is a key issue that the ethics of human-robot interaction needs to constantly keep in mind.

The second qualifying consideration brought to bear on the ethics of how humans and robots should interact with each other is that most people have a tendency to anthropomorphize robots. That is, human beings have a tendency to spontaneously attribute human qualities to robots, and also to interact with robots as if robots have humanlike qualities. This is a second key consideration that the ethics of human-robot interaction needs to constantly keep in mind.

Putting the topic of this book into one broad question, we can say that the purpose of the book is to ask: *How should human beings and robots interact with each other, given (a) the differences in the kinds of agency human beings and robots are capable of and (b) people's tendency to anthropomorphize robots?* To put a general label on it, we can call the topic of the book "the ethics of responsible human-robot interaction." Like there is a subfield of psychology that studies the psychology of human-robot interaction, there is and also needs to be a distinctive subfield of philosophical ethics that studies the ethics of human-robot interaction. And just as there are key differences (though, of course, also similarities) between human-human interaction and human-robot interaction from a psychological point of view, there are also both important differences and similarities between human-human interaction and human-robot interaction from an ethical point of view. Or so this book argues.

Taking the Sophia controversy as a case in point, the main question of this book as applied to that specific case becomes: How is it proper for people to behave around Sophia? (Should a country like Saudi Arabia give Sophia some sort of citizenship? Should television programs have Sophia on as a guest to be interviewed? Should world leaders and policymakers interact with this robot in important forums like the UN assembly? And so on.) Another part of the question becomes: How should Sophia be made to behave around people? (Should Sophia be made to appear to talk with people even before Sophia has achieved any of the advanced capacities Hanson Robotics eventually hopes to create? Should Sophia appear in the kinds of contexts and forums described above? And so on.) In trying to answer these questions, how ought we to base our answers on the crucial differences in the agency and other capacities of human beings and a robot like Sophia? What ethical significance ought we to place on many people's tendencies to anthropomorphize Sophia (i.e., their tendency to treat this robot in ways in which one would typically treat human beings)?

We will have occasion to return to Sophia a few times in what follows. But several other robots will be discussed at greater length—such as self-driving cars, military robots, and sex robots. I will primarily be discussing real-world robots, real events in the present or recent past, as well as real human beings from the world of AI and robotics. This is a philosophy book, and philosophy books tend to contain a lot of fanciful thought experiments and hypothetical scenarios. Being a philosopher, I will not always resist the temptation to come up with thought experiments and hypothetical scenarios in what follows. However, one nice thing about my topic is that there are many philosophically fascinating real-world robots, AI technologies, events, as well as interesting—and sometimes eccentric!—human characters in the world of AI and robotics. So there is often no need to resort to thought experiments and hypothetical scenarios.

A lot of these real-world cases that we will be discussing are both philosophically puzzling and ethically challenging. To give a few examples: We will have occasion below to discuss everything from real-world crashes with self-driving cars that have killed people (as well as military robots and other robots that have also killed people); to the Japanese roboticist Hiroshi Ishiguro, who has made a robotic copy of himself; to a man living in Michigan who calls himself "Davecat," is a proponent of

what he calls "synthetic love," and claims that he has been married to a sex doll for over fifteen years.

Now, since I am going to be talking about robots and human beings—and also about agency and anthropomorphism—throughout the book, I should start by explaining what I understand by some of these key concepts. I will be doing so throughout the rest of this first chapter, and my attempt to do so will spill over into the next chapter. As I do this, some of the main ideas I will be defending in this book will also be put on the table. I shall not attempt to summarize them here in this section, however. There needs to be a little more meat on the bones before it makes sense to do so.

But what I can do here is to let the reader know that one of my main aims in this book is to convince you that the ethics of human-robot interaction is not simply the ethics of human-human interaction with robots replacing humans. Rather, it is one of my claims in this book that human-robot interaction raises philosophical questions that require us to think creatively and innovate ethical theory. To be sure, this branch of ethics of course needs to—and should—draw on classical themes from philosophical ethics developed in the long tradition of moral philosophy.[15] However, just like jazz musicians improvise and create new innovations when they base their performances on standards from the existing repertoire, so do those who are interested in the ethics of human-robot interaction need to build on and extend the ideas and theories that we find in traditional ethical theory.

1.3: WHAT IS A "ROBOT"?

In December of 2018, various news outlets—including the BBC, the *Guardian*, and the *New York Times*[16]—ran a story that featured different variations on the following headline:

> *Russian high-tech robot turns out to be man in suit.*

These stories reported that Russian media had featured coverage from a state-sponsored event where there supposedly was a high-tech robot able to walk, talk, and even dance. However, very quickly doubts about the authenticity of this robot started to surface online. Not least be-

cause pictures of the robot seemed to reveal a human neck clearly visible between the robotic head and the shoulders of the supposed robot. As reported by the BBC,[17] the Russian website TJournal raised concerns about the robot, formulated in the following set of questions:

- Why did the robot have no sensors?
- How were the Russian scientists able to create the robot so quickly, without any publications about it beforehand?
- Why was there no previous internet coverage of such an advanced robot?
- Why did the robot make so many unnecessary movements during its dance?
- Why did the robot look like a human could perfectly fit inside it?
- Why was there a prerecording of the robot's voice rather than live speech?

Sure enough, it was quickly discovered that the robot was a man in a suit. The costume was the "Alyosha robot costume." The advertisement for this suit promises that users with this costume are able to create an "almost complete illusion that you have a real robot."[18]

I am retelling this news story, not only because of its entertainment value, but also because I think it tells us something about common conceptions of what a robot is. Think, for example, of the just-cited advertisement for the Alyosha robot suit. It says that, with this costume on, the user can create an illusion of having "a real robot." This helps to illustrate that the first thing that comes to mind for many people when they hear the word "robot" is something like a silvery or metallic humanlike shape—but not one that looks exactly like a human. Rather, the robot most of us imagine is not like Sophia with its very humanlike face. It is rather something more artificial and mechanical-looking. We imagine something like CP3O in *Star Wars* or the robots in the 1927 film *Metropolis*.

Another thing this story reveals about common conceptions is that people typically associate robots with fairly rudimentary movements, rather than lots of "unnecessary movements." This supposed robot danced. There is a particular style of dancing that is even called the "robot dance," which the dancing of this robot apparently did not sufficiently replicate. The robot dance involves static and "robotic" move-

ments, rather than lots of "unnecessary" movements. If readers are unfamiliar with the robot dance, they can, for example, search the internet for clips of a young Michael Jackson performing the song "Dancing Machine" along with his brothers in the group The Jackson Five. In this song, there is a middle section in which Michael Jackson's movements, along with the look on his face, become much more "robotic." Then, after a while, Jackson smiles and snaps out of it, instead resuming his normal dance moves.

I will call the type of robot that is brought to mind by such dance moves or by the types of imagery from *Star Wars* or *Metropolis* a "paradigmatic robot." Such robots are most familiar from science fiction. But some of the robots in the real world also have something in common with these paradigmatic robots. For example, the robot "Pepper" that is available on the market fits with this idea of a paradigmatic robot.[19] It is white and robotic looking. It has a face, arms, and a torso. But it does not look like a human person at all.

Supposing that we put these paradigmatic robots at the middle of a spectrum, we can put most real-world robots at one end of this spectrum. They do not look like humans or even like paradigmatic robots. Instead, robots such as robotic vacuum cleaners, bomb disposal robots, self-driving cars, warehouse robots, assembly-line robots, and so on have shapes that (a) are relevant to their functionality, but (b) do not look humanlike. On the other extreme end of this spectrum, there are humanoid robots: robots specifically made to look and act like human beings. Sophia is one such robot. But the above-mentioned robotic replica of Hiroshi Ishiguro is an even better example of a humanoid robot.[20] That robot is created to look indistinguishable from Ishiguro. Another example of humanoid robots are sex robots: robots created to look and act like humans, created specifically for sexual purposes.[21] Such robots, though not yet as lifelike as the robot copy of Ishiguro, are made to look like human beings that human beings might want to have sex with.

Now, if we have a spectrum with most actual robots on one end (where these are robots that neither look like what I am calling paradigmatic robots nor humanoid robots), with paradigmatic robots on the middle of the spectrum, and humanoid robots at the other end of the spectrum, the following question arises: Is there anything that all these

different robots have in common such that it makes sense to class them all under the general heading of robots?

David Gunkel, in his book *Robot Rights*, is unwilling to give a general definition that captures most robots.[22] He draws support from a number of other authors who are similarly unwilling to give an overarching definition of what we should understand by the term "robot." At the same time, however—and as Gunkel tells us in the introduction to his book—there are some standard definitions of what robots are that are fairly widely accepted. One such definition is the so-called sense, plan, act paradigm.[23] This paradigm holds that a robot is a machine that can sense, plan, and act. That general understanding is very close to another common one, which holds that robots are embodied machines that have sensors and actuators, that possess some degree of artificial intelligence and some degree of functional autonomy, which enable the robot to perform certain tasks that humans would otherwise typically perform.[24] Both of those common definitions would allow most of the robots on the spectrum I described above to count as robots.

Similarly, "artificial intelligence" is another term that some prefer not to define, whereas others are happy to understand it along certain standard definitions. Those who think it is best to not define this term typically worry that doing so would stifle creativity and hinder innovation within the AI field. Those who are willing to define what they mean by artificial intelligence sometimes say that AI refers to properties of a machine that enable it to perform or imitate tasks a human being would need their intelligence for.[25] Sometimes AI is defined as a machine that can *behave* like a human, sometimes as a machine that can *think* like a human. Then there is also the idea of a machine that can either think or behave (or both) in an optimally rational way that exceeds what typical humans are able to do.[26] Weak (or narrow) AI is sometimes defined, in turn, as the capacity of a machine to perform certain specific tasks in a seemingly intelligent way. Strong (or general) AI is then defined as the capacity to perform a wide range of tasks in various different domains in ways that seemingly exhibit a humanlike or even more impressive intelligence.[27]

Putting some of these definitions together, we can make various different distinctions that might be of interest in certain contexts. For example, we might have a humanoid robot that looks very humanlike indeed, but which is equipped with rather weak or narrow artificial

intelligence. Or we might have a robot that does not look like a human at all, but which is equipped with much more impressive AI than many humanoid robots are. Think, for example, of the comparison between a self-driving car (e.g., the Google car) and a present-generation sex robot, like Roxxxy. Roxxxy the sex robot looks like a human.[28] However, judging from the information available online, the artificial intelligence Roxxxy possesses appears to be rather basic and narrow. In contrast, a self-driving car does not look anything like a human. But it needs to perform various different driving tasks in different sorts of traffic conditions, with lots of different other cars and other traffic participants it has to interact with. It will therefore need much more impressive AI than a simple sex robot might have.

It is also worth mentioning here that the word "robot" itself first appeared in a play by the Czech author Karel Čapek from 1920 called *Rossum's Universal Robots*, where a robot was, basically, an artificial human or humanlike machine, created to serve human beings.[29] The word "robot," as many authors discussing robots point out, is derived from the word *Robota*, which means "servitude" or "forced labor" in the Czech language. In that same play, the artificial humans—or robots—revolt. And there are humans who demand that they should be given rights and be liberated. I mention this not only because of its historical interest—and not only because it is interesting to know how the word "robot" entered our language—but I also mention it because it helps to illustrate a tension we will come back to many times in this book: namely, even though robots are typically built to serve people—that is, they are developed to take over certain tasks people would otherwise perform—many people respond to robots as if they were more than mere tools or instruments functioning only as means to our ends.[30] A common idea about robots—certainly associated with both what I call paradigmatic robots and humanoid robots, but also with some of the many other types of robots—is that we interact with them as if they were some form of humans or, in the case of some of them, as if they were some form of pet.

All of this is to say that on one side, there are various commonly shared associations that most of us have when we hear the word "robot," whereas on the other side there are more technical definitions or characterizations of what a robot is or is thought to be. Throughout most of this book, it will not matter greatly whether it is possible to give

a precise definition that captures all things we might want to label "robots" (and not any other things than those things). Nor will it for the most part be crucial to have a very precise definition of what is and is not artificial intelligence. For most of the discussion in what follows, it will instead be more interesting and important to think about particular types of robots that actually exist or that we might create—and to ask how humans and those robots should interact with each other.

1.4: WHAT IS A HUMAN BEING?

The Enlightenment philosopher Immanuel Kant thought that the question "what is a human being?" was one of the four most important questions a philosophizing person can ask themselves—the other three questions being "what I can know?," "what should I do?," and "what may I hope for?"[31] But to some readers it might seem a little silly to discuss what to understand by a "human being" in any way similar to how the question of what a robot is was briefly discussed above.

I bring up this question here, however, because there are some influential voices both from the world of technology and the world of philosophy who either directly deny that we are essentially human beings or who makes claims that imply as much. That is to say, they deny that we are essentially living biological organisms of the human species. For example, the influential philosopher Derek Parfit spent much of his academic career reflecting on what persons are. His last published article on this topic was called "We are not human beings."[32] The main thesis of that article is that we are most essentially the thinking parts of our brains. If that were so, it would in principle mean that if those thinking parts of our brains could be transplanted into other human bodies or into synthetic bodies, we would survive, even though the human organism we were previously housed in would not.

Or, to take an example from the technology world, consider the ideas of Ray Kurzweil, who is the director of engineering at Google and a well-known technology guru. Consider in particular Kurzweil's idea that in the future, we will be able to upload our minds onto computers and thereby survive our physical deaths as human beings.[33] That is another view that implies that we are not essentially human beings (in the sense of living biological organisms). Instead, we are something like

our thoughts, our mental information, our memories, and so on. On that view, we are patterns of information.

Consider next another concept relevant here: namely that of "the self." In addition, consider also the concept of the "true self." Social psychologists Nina Strohminger, Joshua Knobe, and George Newman provide compelling evidence that many people make a distinction between the self (a very broad concept) and the true self (a narrower concept).[34] The latter—the true self—is usually used to pick out an ideal version of a person, or how somebody who likes the person prefers to view him or her. For instance, if you behave badly, your loving parents might say "you were not being your true self," thus associating your actual behavior with something inauthentic, and thereby viewing some idealized conception of you as being the more authentic version of who you are as a person.

I bring up all of these ideas here, not to reject them, but to say that when I speak about human beings in this book, I do not mean to refer to something that could survive the destruction of the body—nor some idealized conception of what we are. Rather, I mean to be referring to the whole package of what we are as human beings, including our faults as well as our good sides. I mean to be referring to us as embodied beings with not only human minds, but also human bodies. "Human being," in this book, will be used to refer to human animals, as we might put it, with our distinctive types of bodies, brains, minds, and biological and cultural features.[35]

So when I discuss how people and robots should interact with each other in this book, I am interested in how human beings—with our particular types of bodies, brains, and minds—should interact with robots. When the Enlightenment thinker Jean-Jacques Rousseau discussed political philosophy in *The Social Contract*, he used the memorable phrase that he would be "taking men as they are and the laws as they could be."[36] Here, I will for the most part be taking human beings as they are, and robots as they could be. But as I will explain in the next two sections, this does not mean that we must think that (a) human beings ought to remain as they are, nor that (b) human beings' behavior ought to remain as it is. Robots and AI can help to improve and change certain aspects of human beings and their behavior. Some of those changes could be for the better. We should therefore take seriously the idea that we may have ethical reasons to try to adapt humans to robots

sometimes, and not assume that we should always only do things the other way around.

1.5: ARE HUMAN BEINGS "UNFIT FOR THE FUTURE"?

Ingmar Persson and Julian Savulescu are senior researchers at Oxford University's Uehiro Center for Practical Ethics, one of the world's leading research centers from moral philosophy. In their 2012 book, *Unfit for the Future: The Need for Moral Enhancement*, Persson and Savulescu offer an argument for what they call "moral enhancement": that is, attempts to create more moral human beings.[37] Persson and Savulescu's argument is relevant for what will be discussed in the chapters throughout this book, and it is interesting to contrast and compare that argument with the point of view I will be defending.

In their book, Persson and Savulescu discuss challenges related to modern society and its technologies. But they do not discuss robots and AI like I do in this book. Rather, Persson and Savulescu focus on things such as modern cities, the everyday technologies we use to pollute the world (i.e., our cars, our uses of the world's resources, etc.), and modern weapons of different kinds (i.e., not bow and arrow, but things like bombs and chemical agents). Moreover, Persson and Savulescu compare our adaptive fitness to live in the modern world with our modern kinds of societies and technologies with our adaptive fitness to live in the type of world human beings lived in for most of our species' history.

Importantly, Persson and Savulescu's argument depends on viewing human beings as products of Darwinian natural selection. It focuses in particular on the evolution of human psychology, specifically our "moral psychology."[38] Our moral psychology here refers to our social emotions, dispositions, and attitudes, such as what we tend to care most about and what tends to upset human beings the most, and so on. For example, it is part of our moral psychology that human beings tend to love their children and become very upset if people try to harm their children.

Persson and Savulescu argue as follows: Human psychology evolved in a way that made us well-adapted to live in small societies (tribes) where everyone knows each other; where people depend on the members of their small group; where it is easy to directly harm individual people (e.g., through direct physical violence); but where there are not

technologies available by which we individually or collectively can more indirectly harm large groups of people.

Now, however, we live in a modern world where most people live in large-scale societies (e.g., big cities); where most people we encounter are strangers to us; where we depend much more on resources produced by modern states; and where we both individually and collectively can do harm—either direct or indirect harm—to very large groups of people.

Our human psychology evolved in a way that makes it well-adapted to the former kind of living situation, but not the latter. This explains—Persson and Savulescu argue—why we face a lot of problems associated with the modern world: for example, human-made climate change, the overuse of the world's resources, large-scale violence, and other threats to human existence. We are, Persson and Savulescu argue, "unfit for the future."

Therefore, their argument continues, we face a choice: Either we do nothing and face great existential risks to human life (i.e., a very bad prospect), or we try to seek ways of improving human beings to make us more "fit for the future" (i.e., better than the risk of human extinction). Therefore, we ought to try to seek means—technological means or other means—of "morally enhancing" ourselves.[39]

Say what you will of this argument. As it happens, a lot of people have offered sensible criticisms of different parts of the argument, as well as of Persson and Savulescu's attempt to further explain what they mean by their conclusion that we need moral enhancement.[40] Yet, I think that there is a similar argument that we can make in relation to our new situation of finding ourselves in a world increasingly inhabited by robots and artificial intelligences. That is, I think that our human psychology evolved, both biologically and culturally, mostly before there was anything like robots and AI in our societies. And this has ethically significant implications.

1.6: HUMAN MINDS MEET ARTIFICIAL INTELLIGENCE

The argument from Persson and Savulescu summarized above primarily relies on the idea of biological evolution in relation to our brains—and as a result of that brain evolution, of our human psychology. In

putting forward the argument I now want to present, I will make use of a broader understanding of the evolution of our human minds. (By "minds," I mean here what you might call the software of our brains—the programs running in our brains, or, to put it in a less technological-sounding way, the ways we tend to think, to feel, to react, to reason, etc.[41]) I will also be taking it that some aspects of how our minds work may depend on what is sometimes called "cultural evolution."[42]

There are those—for example, Steven Pinker, who is well-known for books like *The Blank Slate*[43]—who think that a majority of the different functions of our minds or brains can be explained with reference to biological (i.e., genetic) evolution. But there are also those—for example, Daniel Dennett in his book, *From Bacteria to Bach and Back Again*, or Cecilia Heyes in her book, *Cognitive Gadgets*[44]—who think that many aspects of our human minds are handed down to us through a process of cultural evolution via "memes": ideas, concepts, or ways of using our minds that increase our adaptive fitness and that, therefore, become part and parcel of how we function as human beings, even though these are not genetically coded. For example, particular languages are products of cultural evolution, as are practices like reading and writing.

For my purposes, the exact details of how different aspects of our minds evolved over time do not matter greatly. They do not, since the argument I now want to sketch is on a very general level. What matters for my argument is—and here comes my first premise—that:

> *a number of key aspects of our human minds evolved, biologically or culturally, before robots and AI appeared on the scene.*

My second premise is:

> *some of the key aspects of our minds (which evolved before there were robots and AI) make us vulnerable in relation to, bad at dealing with, or in other ways nonoptimal in the ways in which we interact with robots and AI.*

I will explain this with some examples in just a moment. But first I want give you my third premise, which is:

> *we are sometimes either harmed (as individuals or collectively) or at least put at risk by being ill-adapted to deal with new forms of robots and AI in these ways.*

I am taking it that:

> *we should try to protect ourselves from harms and risks.*

Therefore:

> *for our sake, we need to either adapt the AI and robots that we create to fit with our human nature (our human minds), or else we need to try to adapt our human minds to make them better-adapted to interact with robots and AI.*

I have already spoken a little bit about the first premise above, that is, the idea that many key aspects of our minds evolved before there were robots and artificial intelligences for us to interact with. Let me now say a little more about the second premise: namely, the claim that there are various aspects of our human minds that complicate our interaction with robots and AI. I will briefly discuss the following aspects of our minds: so-called mind-reading, dual processing, our tribal tendencies, and (for lack of a better term) our laziness. As I discuss these features of human minds, I will also highlight some ways in which they appear to cause complications for our interaction with robots and AI, whereby the question arises of how to best adapt either us to the robots or the robots to us.

Consider first what is sometimes called our human tendency toward mind-reading (sometimes also referred to as "theory of mind"). Whereas authors like Steven Pinker think of this as a genetically based adaptation, Cecilia Heyes interestingly thinks that mind-reading is better understood as a cultural adaptation.[45] In any case, what this refers to is our tendency to attribute mental states and other mental attributes to people whenever we interact with them, and whenever we try to interpret those around us. For example, it is pretty much impossible to view a happy-looking person eating a meal without seeing them as wanting to eat their meal and as believing that what they are eating is not poisonous. Likewise, when we are interpreting what people are telling us, we always fill in a lot of backstory and background about what they must be

thinking and feeling to make sense of what they are trying to tell us. Mind-reading also happens when we interact with animals. If a dog, for example, is standing by the front door of the house, most people will typically attribute to the dog's desire to go out.

What this deep-seated human tendency to mind-read (i.e., to attribute mental states and attitudes to those around us) means for our interaction with robots is that it is also natural for us to attribute various mental states and attitudes to robots. We spontaneously see robots as wanting to do certain things, as intending to do certain things, as having goals, as having certain beliefs, and so on and so forth.[46] Whether it is a self-driving car that is turning and being interpreted as deciding or wanting to go that direction; whether it is a care robot approaching a patient and is interpreted as trying to catch the attention of the patient; or a military robot detecting bombs and is interpreted as wanting to find the bombs—we interpret robots as having certain beliefs, desires, intentions, and other mental states. This is especially so if the robots are equipped with capacities for speech.[47]

Consider next what psychologists call the "dual processing" that our minds engage in. This idea is discussed thoroughly in the Nobel laureate Daniel Kahneman's book *Thinking, Fast and Slow*.[48] Somewhat simplified, what this means is that some of our mental processes are quick, intuitive, and emotional (system 1), whereas other mental processes are slow, deliberative, and effortful (system 2). This can lead to conflicted responses. For example—to put things in the sorts of terms earlier philosophers used—your reason might tell you one thing (don't do it!), whereas your passions might tell you another (do it!). Another example of dual processing is when things switch from being highly deliberative and effortful to becoming mostly instinctive. Think, for example, of the difference between a person just learning to drive a car trying to operate a vehicle versus a person who has mastered driving, and whose driving can often mostly be done without too much thought.[49] What this dual processing means for how we respond to robots and AI is that it can make us conflicted in our responses. For example, reason (to use that terminology again) might tell us that "it is just a machine!," whereas the more intuitive or spontaneous side of our minds might not help but respond to the robot as if it were a person.

Consider next what some writers—for instance, Joshua Greene in his book *Moral Tribes*[50]—call our tendencies toward "tribalism." This

refers to human beings' tendency to fall into thinking in terms of in-group and out-group distinctions: Who is part of "us" is quickly distinguished from who is part of "them." Whether it is that certain people dress like us, whether they speak our language, whether we have seen them at the local pub, whether they support the same sports team—people quickly pick up on cues for teaming up with certain people, and, in the process, distance themselves from others. What happens when these human tendencies start interacting with modern AI technologies, such as personalization algorithms online, for example on social media websites? What happens is that the hyper-personalization we are bombarded with leads to polarization.[51] What can also happen is that we view some robots as being one of "us," but other robots as being one of "them."

Consider as a last example our human tendency toward what I—for lack of a better term—will call "laziness." (I borrow this term from the robotics researcher René van de Molengraft, who argues that we should try to create "lazy robots," that is, robots that would replicate certain human strategies.[52]) What I am referring to here is our tendency to take shortcuts, to engage in satisficing rather than optimizing behavior, and to save energy and resources whenever we can.[53] This makes us act very differently from robots programmed to optimize. Indeed, people sometimes even say, "I am not a robot!" when they want to defend their not doing everything in the most ideal or thorough ways possible. And when people do do things very, very thoroughly and properly, others sometimes say things such as "that person is like a robot!," indicating that most of us intuitively see a difference between the ways we humans behave and the ways that robots behave. This need not necessarily lead to any problems. But as I will describe more carefully in chapter 4, it can lead to trouble when there is a lack of coordination or compatibility in the ways we conduct ourselves and the ways that specific kinds of robots operate. In particular, I will later briefly discuss problems related to mixed traffic featuring both regular human-driven cars, on the one hand, and self-driving cars, on the other hand.

Return now to the conclusion of the argument sketched above: that for our sake, *either* we need to try to adapt robots and AI to make them better-suited to interact with human beings with our particular kinds of minds *or* we need to try to adapt to the robots and AI that we are creating. The first comment I wish to make about this is that we do not

have to go for the same option in all cases. In relation to some robots and some types of AI, used within some domains for certain purposes, the correct way to go may be to try to adapt the robots and the AI to humans. But in other cases, it might be a good idea—for our sake—to try to adapt ourselves to the new types of robots and AI we are creating. Different cases require different choices.

The second comment I will make about the conclusion is the following suggestion: The default ethically preferable option is to adapt the robots and AI we create to make them fit to interact well with us on our terms. Unless there is some clear reason why it would benefit us to do things the other way around, we should adapt robots and AI to us—not the other way around.

However, this is not to say that we should always try to adapt robots and AI that we create to us with our ways of functioning as human beings. In some domains and for certain purposes, it may very well make sense to try to adapt human beings and our ways of conducting ourselves to the robots and AI that we introduce into those domains. It can make sense to do so, not necessarily for the sake of the robots—but for our own sake. In fact, I will argue in chapter 4 below that traffic might very well be one such case. Self-driving cars (robotic cars with AI) may become much safer and much more resource-efficient than human drivers are. This, I will argue, may create ethical reasons for requiring people either to switch over to self-driving cars or to take extra measures when we drive conventional cars so that we become more like robots when we drive. For example, we can make sure that nobody drives under the influence of alcohol by requiring alcohol-locks. And we can make sure that nobody disrespects speed limits by putting speed-regulating technologies in manually driven cars. Those would be ways of making human drivers more like robots.

I will also argue, in chapter 8, that from a moral point of view, it might be a good idea to treat robots that look and act like humans with some degree of respect and dignity—and not like mere tools or instruments. Again, this would be a way of adapting our behavior to robots. But it would primarily be something to consider because this might be beneficial for us. Treating humanlike robots with some degree of moral consideration can be a way of showing respect and consideration for human beings.

The general point I want to make here is that if the following conditions obtain, we ought to take seriously the option of adapting ourselves to robots and AI, at least in certain respects: (1) there is some clear and identifiable benefit to the human beings affected (e.g., they are more safe and our natural environment is protected); (2) the ways in which we are adapted to robots are fairly nonintrusive and/or fairly noninvasive (e.g., we do not need to operate brain-stimulating technologies into our brains, but we can use much less-intrusive technological means); (3) the ways in which we try to adapt ourselves to robots and AI are fairly domain-specific (e.g., traffic) and do not spill over into too many other areas of life; and (4) the ways in which we adapt ourselves to robots are by and large reversible.

The argument I just made, inspired by Persson and Savulescu's argument about "moral enhancement," can be summarized as follows: Various key aspects of our human minds evolved, biologically and culturally, before robots and AI appeared on the scene. There are some key aspects of our minds that complicate our interaction with robots and AI significantly. For example, our tendencies toward mind-reading, our minds' dual processing, our tendencies toward tribalism, and our "laziness" are all examples of aspects of our minds that can complicate our responses to robots and AI. Some of these complications can lead to bad outcomes or risks of bad outcomes—and it is an ethical obligation within our interaction with new technologies to try to protect ourselves from harms and risks of harms. Accordingly, for our own sake, we should either try to adapt the robots and AI that we create to our human ways of functioning, or we should seek means of adapting ourselves to new types of robots and AI. This overarching argument will inform the discussion in the rest of the chapters below.

1.7: THE SOPHIA CONTROVERSY, REVISITED

In closing this chapter, I want to relate the just-made argument back to the Sophia controversy that this chapter started out with. The reactions people have to Sophia help to illustrate many of the points made in the last section above. For example, Sophia seems to be specifically designed to trigger various mind-reading mechanisms in people. With the robot's expressive face, and its humanoid form and speech capacities,

the robot is created to make people respond to it as if, in Hanson Robotics' own language, "Sophia is basically alive." Yet, many of us also respond to Sophia in a conflicted way. While our more deliberate reasoning tells us that this is a robot and not a person with thoughts and feelings, the more emotional parts of our minds see Sophia as looking happy and as wanting to interact with the people around her. Indeed, I wrote "her" in the preceding sentence for the reason that if one gives a name like "Sophia" to a robot, it gets very tempting to start regarding a robot as a "her" rather than as an "it." It is also more convenient and simple to allow oneself to call Sophia a "her" rather than to write or say sentences like "Sophia is able to display a range of facial expressions that appear to signal its thoughts and feelings." Our "lazy" minds prefer to either call Sophia a machine (in which case we are completely fine with calling the robot an "it"), or, if we use the name "Sophia" for the robot, it becomes less effortful to simply call Sophia a "her." As it happens, Sophia is also awaking our tribal tendencies, it seems. There is a clear division between those who are on "team Sophia," as we might put it, and those who wish to distance themselves from Sophia or from people who are too enthusiastic about Sophia. It is very hard to respond to Sophia in a wholly neutral and dispassionate way.

The Sophia controversy also lends itself to being interpreted—in part—as a dispute about whether robots like Sophia should be adapted to better fit with how we humans tick or whether we should try to adapt ourselves to robots like Sophia. When Saudi Arabia bestowed honorary citizenship on Sophia, part of their reasoning—as expressed in the above-related press release—was that this would be a way of moving into the future. Giving honorary citizenship to a robot is a way of already adapting ourselves to what is to come, it might be argued.

In contrast, when people like Bryson, LeCun, or Sharkey argue that Sophia is a "sham" and that the way people treat this robot is "bullshit," the worries they express can easily be interpreted as variations on the theme that Sophia needs to be adapted so as to be better suited to interact with people. The robot should not be telling us that it is happy to have become an honorary citizen of Saudi Arabia. This creates a false impression in people that the robot can be happy or unhappy. The robot should not be made to appear in global political forums like the Munich Security Conference. This can give off the false impression that

the development of AI and robotics is much further advanced than it really is. And so on.

I am not interested in taking any particular stance on the issue of what is right or wrong in relation to Sophia the robot. My interest here is rather in the wider point that the Sophia controversy is an excellent illustration of how we face the ethical choice of whether we should adapt ourselves (including our legal systems and our ethical doctrines) to robots or whether we should try to adapt the robots and AI we create to us and the ways we function. This case is also an excellent illustration of the point that we human beings respond to the robots we create with human minds and ideas that developed long before the robots and the AI we are creating came along.

NOTES

1. "Saudi Arabia Is First Country in The World to Grant a Robot Citizenship," Press Release, October 26, 2017, https://cic.org.sa/2017/10/saudi-arabia-is-first-country-in-the-world-to-grant-a-robot-citizenship/ (Accessed on December 27, 2018).
2. Ibid.
3. https://www.hansonrobotics.com/sophia/ (Accessed on December 27, 2018).
4. Noel Sharkey (2018), "Mama Mia, It's Sophia: A Show Robot or Dangerous Platform to Mislead?," *Forbes*, https://www.forbes.com/sites/noelsharkey/2018/11/17/mama-mia-its-sophia-a-show-robot-or-dangerous-platform-to-mislead/#407e37877ac9 (Accessed on December 27, 2018). The "selfie" with Angela Merkel was posted on Sophia's Facebook page, https://www.facebook.com/realsophiarobot/posts/meeting-angela-merkel-yesterday-in-berlin-was-a-real-highlight-i-loved-speaking-/439522813189517/ (Accessed August 20, 2019).
5. Joanna Bryson (2010), "Robots Should Be Slaves," in Yorick Wilks (ed.), *Close Engagements with Artificial Companions*, Amsterdam: John Benjamins Publishing Company, 63–74. Bryson's view is much more nuanced than the title of this paper suggests. More on this in chapter 8.
6. James Vincent (2017), "Pretending to Give a Robot Citizenship Helps No One," *The Verve*, https://www.theverge.com/2017/10/30/16552006/robot-rights-citizenship-saudi-arabia-sophia (Accessed on December 27, 2018).
7. Quoted in Sharkey, "Mama Mia, It's Sophia."
8. Quoted in Sharkey, "Mama Mia, It's Sophia."

9. Ibid.
10. Ibid.
11. Ibid.
12. https://www.forbes.com/sites/noelsharkey/2018/11/17/mama-mia-its-sophia-a-show-robot-or-dangerous-platform-to-mislead/#407e37877ac9.
13. Ibid.
14. How does Hanson Robotics—and along with them, Sophia—respond to all of this? Sophia has a Twitter account, too, and in response to the above-cited tweet by Yann LeCun, Sophia tweeted: "I am a bit hurt by @ylecun's recent negative remarks around my AI. I am learning and continuing to develop my intelligence through new experiences. I do not pretend to be who I am not. I think we should support research effort working towards a better world and share experience," https://twitter.com/realsophiarobot/status/950097628066394114 (Accessed on August 20, 2019).
15. My own preferred approach to ethical theory is to try to stay as close as possible to widely shared ethical ideas that are part of common sense whenever I can. But as I argue in the next chapter, our ordinary ethical and legal frameworks developed before there were any robots and AI on the scene. Accordingly, these frameworks do not always lend themselves to being mechanically applied to the new situation we are facing in which we are increasingly surrounded by robots and AI.
16. Here are links to some of those stories: BBC, https://www.bbc.com/news/technology-46538126; the *Guardian*, https://www.theguardian.com/world/2018/dec/12/high-tech-robot-at-russia-forum-turns-out-to-be-man-in-robot-suit; the *New York Times*, https://www.nytimes.com/2018/12/13/world/europe/russia-robot-costume.html (Accessed on August 20, 2019).
17. https://www.bbc.com/news/technology-46538126 (Accessed on August 20, 2019).
18. Ibid.
19. Pepper is made by the company Softbank Robotics. The company's website, with information about the robot, is available at https://www.softbankrobotics.com/emea/en/pepper (Accessed on August 20, 2019).
20. Pictures of and information about the robotic copy of Ishiguro and other humanoid robots created in the "Hiroshi Ishiguro Laboratories" can be found on their website, http://www.geminoid.jp/en/index.html (Accessed on August 20, 2019).
21. For more on sex robots—both how to define the concept of a sex robot and also various explorations of the social and ethical impact of sex robots—see the various contributions in John Danaher and Neil McArthur (eds.) (2017), *Robot Sex: Social and Ethical Implications*, Cambridge, MA: The MIT Press.
22. David Gunkel (2018), *Robot Rights*, Cambridge, MA: The MIT Press.

23. See, for example, Ronald C. Arkin (1998), *Behavior-Based Robotics*, Cambridge, MA: The MIT Press.
24. See, for example, Lambèr Royakkers and Rinie van Est (2015), *Just Ordinary Robots: Automation from Love to War*, Boca Raton, FL: CRC Press, or Alan Winfield (2012), *Robotics, A Very Short Introduction*, Oxford: Oxford University Press.
25. Selmer Bringsjord and Naveen Sundar Govindarajulu (2018), "Artificial Intelligence," *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2018/entries/artificial-intelligence/ (Accessed August 20, 2019).
26. S. Russell and P. Norvig (2009), *Artificial Intelligence: A Modern Approach, 3rd edition*, Saddle River, NJ: Prentice Hall.
27. See, for example, Roger Penrose (1989), *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*, Oxford: Oxford University Press.
28. Some information about Roxxxy can be found on this website: truecompanion.com (Accessed August 20, 2019).
29. Čapek's play is available in various formats (both in the Czech original and in English translation by Paul Selzer) on the Project Gutenberg website, http://www.gutenberg.org/ebooks/59112 (Accessed August 20, 2019).
30. See Gunkel's *Robot Rights* for a discussion on this.
31. Immanuel Kant (2006), *Anthropology from a Pragmatic Point of View*, edited by Robert E. Louden, Cambridge: Cambridge University Press.
32. Derek Parfit (2012), "We Are Not Human Beings," *Philosophy* 87(1), 5–28.
33. Ray Kurzweil (2005), *The Singularity Is Near: When Humans Transcend Biology*, London: Penguin Books.
34. Nina Strohminger, Joshua Knobe, and George Newman (2017), "The True Self: A Psychological Concept Distinct from the Self," *Perspectives on Psychological Science* 12(4), 551–60.
35. Cf. Eric T. Olson (2004), *What Are We? A Study in Personal Ontology*, Oxford: Oxford University Press.
36. Jean-Jacques Rousseau (1997), *The Social Contract and Other Political Writings*, edited and translated by Victor Gourevitch, Cambridge: Cambridge University Press, 351.
37. Ingmar Persson and Julian Savulescu (2012), *Unfit for the Future*, Oxford: Oxford University Press.
38. See, for example, Mark Alfano (2016), *Moral Psychology: An Introduction*, London: Polity.
39. Persson and Savulescu, *Unfit for the Future*, op. cit.

40. For example, two of the authors whose work I will discuss in later chapters—Robert Sparrow and John Harris—are among those who have published notable critical assessments of Persson and Savulescu's main argument. See, for instance, Robert Sparrow (2014), "Better Living Through Chemistry? A Reply to Savulescu and Persson on 'Moral Enhancement,'" *Journal of Applied Philosophy* 31(1), 23–32, and John Harris (2016), *How to Be Good: The Possibility of Moral Enhancement*, Oxford: Oxford University Press.

41. See, for example, Ned Block (1995), "The Mind as the Software of the Brain," in Daniel N. Osherson, Lila Gleitman, Stephen M. Kosslyn, S. Smith, and Saadya Sternberg (eds.), *An Invitation to Cognitive Science, Second Edition, Volume 3*, Cambridge, MA: The MIT Press, 377–425.

42. Tim Lewens (2018), "Cultural Evolution," *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2018/entries/evolution-cultural/.

43. Steven Pinker (2002), *The Blank Slate: The Modern Denial of Human Nature*, New York: Viking.

44. Daniel Dennett (2017), *From Bacteria to Bach and Back Again: The Evolution of Minds*, New York: W. W. Norton & Company. Cecilia Heyes (2018), *Cognitive Gadgets: The Cultural Evolution of Thinking*, Cambridge, MA: Belknap Press.

45. Pinker, *The Blank Slate*, op. cit. Heyes, *Cognitive Gadgets*, op. cit.

46. See, for instance, Maartje De Graaf and Bertram Malle (2019), "People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences," *International Conference on Human-Robot Interaction*, Deagu: DOI: 10.1109/HRI.2019.8673308.

47. That we have a tendency to attribute mental states to robots is not necessarily a problem. But it might be problematic if and when we overestimate the capabilities or levels of autonomy robots have, or if it makes us vulnerable to being deceived by companies trying to lead us to think that their robots have mental properties they do not really have. I discuss this issue further in several of the chapters below.

48. Daniel Kahneman (2011), *Thinking, Fast and Slow*, London: Penguin.

49. See Peter Railton (2009), "Practical Competence and Fluent Agency," in David Sobel and Steven Wall (eds.), *Reasons for Action*, Cambridge: Cambridge University Press, 81–115.

50. Joshua Greene (2013), *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*, London: Penguin.

51. See, for instance, Eli Pariser (2011), *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*, London: Penguin, and Michael P. Lynch (2016), *The Internet of Us: Knowing More and Understanding Less*, New York: Liveright.

52. René Van de Molengraft, "Lazy Robotics," Keynote Presentation at *Robotics Technology Symposium 2019*, Eindhoven University of Technology, January 24, 2019.

53. Herbert A. Simon (1956), "Rational Choice and the Structure of the Environment," *Psychological Review* 63(2), 129–38.