ARTIFICIAL INTELLIGENCE

WHAT EVERYONE NEEDS TO KNOW®

JERRY KAPLAN

OXFORD
UNIVERSITY PRESS

26. Thomas C. Scott-Phillips, "Evolutionary Psychology and the Origins of Language," *Journal of Evolutionary Psychology* 8(4) (2010):289–307, https://thomscottphillips.files.wordpress.com/2014/08/scott-phillips-2010-ep-and-language-origins.pdf.
27. Noam Chomsky, "Three Factors in Language Design," *Linguistic Inquiry* 36, no. 1 (2005): 1–22, http://www.biolinguistics.uqam.ca/Chomsky_05.pdf.
28. The actual process by which computer languages are executed is more nuanced than implied here. Some are "compiled"—translated into a so-called lower-level language in advance of being executed, while others are "interpreted" a little at a time as they are needed.
29. For links and an introduction to statistical machine translation, see http://www.statmt.org.

4

PHILOSOPHY OF ARTIFICIAL INTELLIGENCE

What is the philosophy of AI?

You might wonder why a field like AI seems to attract so much controversy. After all, other engineering disciplines—such as civil, mechanical, or electrical engineering—aren't typically the target of vociferous criticism from various branches of the humanities. Largely, these wounds are self-inflicted, as some practitioners, whether due to naiveté or in an effort to draw attention and funding, have made highly public overly broad claims for the generality of their results and optimistic forecasts of the future trajectory of the field.[1] That said, AI does pose real challenges to philosophical and religious doctrine about human uniqueness and our place in the universe. Intelligent machines offer the potential to shine an objective light on fundamental questions about the nature of our minds, the existence of free will, and whether nonbiological agents can be said to be alive. The prospect of actually settling many deep historical debates is both exciting and a little scary for those who ponder such issues. In the end, many of these issues come down to basic beliefs we have about ourselves, some of which resist scientific explanation (such as the existence of the human soul), or the Cartesian idea that mental events are somehow distinct from and independent of the physical world (dualism).

These intellectual questions are compounded by more pedestrian fears that AI may threaten the livelihoods if not the actual lives of many people. This concern, though legitimate, is fanned by the recurring theme in fiction and film of robot rebellion, dating back at least to the 1920 play by Czech playwright Karel Čapek, *R.U.R.*, also called *Rossum's Universal Robots*, which is credited with inventing the term *robot* (after the Czech word *robota*, meaning forced labor).[2]

In short, the philosophy of AI asks the question of whether computers, machines in general, or for that matter anything that is not of natural origin can be said to have a mind, and/or to think. The answer, simply put, depends on what you mean by "mind" and "think." The debate has raged on in various forms—unabated and unresolved—for decades, with no end in sight.

Here's some of the colorful history and arguments put forth by proponents and critics of the idea that machines can or do possess thinking minds.

What is "strong" versus "weak" AI?

I won't review the litany of claims made by AI researchers, but the most controversial of these can be summarized as a variant of what's called the "strong" versus the "weak" view on AI. In short, strong AI posits that machines do or ultimately will have minds, while weak AI asserts that they merely simulate, rather than duplicate, real intelligence. (The terms are sometimes misused, in my opinion, to describe the distinction between systems that exhibit general intelligent behavior versus those that are limited to a narrow domain, functioning as electronic idiot savants.) Stated another way, the distinction is between whether machines can be truly intelligent or simply able to act "as if" they are intelligent.

To demonstrate how confusing this matter can be, in this chapter I will attempt to convince you that you simultaneously hold contradictory views on this subject. If you do, it doesn't mean that you are crazy or muddled in your thinking; instead, I believe it indicates that we simply don't have an accepted intellectual framework sufficient to resolve this conflict—at least not yet. You and I may not, but I'm hopeful that at some point in the future, our children will.

Can a computer "think"?

The noted English mathematician Alan Turing considered this question in a 1950 essay entitled "Computing Machinery and Intelligence."[3] In it, he proposes, essentially, to put the issue to a vote. Constructing what he calls the "imitation game," he imagines an interrogator in a separate room, communicating with a man and a woman only through written communication (preferably typed), attempting to guess which interlocutor is the man and which is the woman. The man tries to fool the interrogator into thinking he is the woman, leaving the woman to proclaim her veracity (in vain, as Turing notes) in an attempt to help the interrogator make the correct identifications. Turing then invites the reader to imagine substituting a machine for the man, and a man for the woman. (The imitation game is now widely called the Turing Test.)[4]

Leaving aside the remarkable psychological irony of this famously homosexual scientist tasking the man with convincing the interrogator that he is a woman, not to mention his placing the man in the role of deceiver and the woman as truth teller, he goes on to ask whether it's plausible that the machine could ever win this game against a man. (That is, the machine is tasked with fooling the interrogator into thinking it is the man, while the man is telling the truth about who he is.) Contrary to the widely held belief that Turing was proposing an "entrance exam" to determine whether machines had come of age and become intelligent, he was actually speculating that our common use of the term *think* would eventually stretch sufficiently to be appropriately applied to certain machines or programs of adequate capability. His estimate of when this might occur was the end of the twentieth century, a remarkably accurate guess considering that we now routinely refer to computers as "thinking," mostly when we are waiting impatiently for them to respond. In his words, "The original question, 'Can machines think?' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."[5]

Is Turing right? Is this question too meaningless to deserve discussion? (And thus, by implication, this discussion is a waste of time?) Obviously, it depends on what we mean by "think."

We might consider thinking to be the ability to manipulate symbols to reason from initial assumptions to conclusions. From this perspective, it should be noncontroversial that computer programs, as we currently interpret them, are capable of such manipulations and therefore are capable of thinking. But surely just stirring up a brew of symbols isn't sufficient—it has to mean something or do something. Otherwise, there's not much justification for distinguishing one computer program from another, and any program that we interpret as processing symbols—no matter now trivial—would qualify as thinking, which doesn't seem right. But how does a computer program mean or do something?

The branch of philosophy and linguistics that deals with such questions, semiotics, studies the use of symbols for reasoning and communication. A distinction is commonly made between syntax, the rules for arranging and manipulating symbols, and semantics, the meaning of the symbols and rules. While syntax is pretty easy is to understand, semantics is not—even the experts don't agree on what "meaning" means. Most theories propose that meaning requires some way of relating the symbols themselves to the things they denote in the real world.

A quick example might help. You may think of numbers by themselves as having meaning, but they don't. To visualize why, consider the following symbols !, @, #, and $ as connected by an operator, +, that you can use to combine any pair of symbols from the set (=) into another symbol in the set:

```
! + ! = @
! + @ = #
@ +! = #
! + # = $
# +! = $
@ + @ = $
```

Now you can play a little game of starting with a set of symbols and tracing it through the above rules to see where you wind up. Sounds like a good way to keep your five-year-old occupied for a few minutes, but it doesn't exactly command your attention as expressing a fundamental truth about the structure of our universe—until you substitute different symbols, leaving everything else the same:

```
1 + 1 = 2
1 + 2 = 3
2 + 1 = 3
1 + 3 = 4
3 + 1 = 4
2 +2 = 4
```

Suddenly, everything makes sense. We all know what 1, 2, 3 and 4 mean, except for the minor inconvenience that they don't actually mean anything more or less than !, @, #, and $ do. They derive their meaning from how we connect them to other concepts or real-world objects. If we connect $ with any collection of four things, an expanded set of the above rules is exceedingly useful for solving certain problems of great practical significance. You can sit around manipulating symbols all day long and it doesn't mean a thing. In other words, loosely speaking, what you think doesn't matter, until you actually do something.

And to do something requires some connection between the actor manipulating the symbol system and something external to that actor. In the case of computer programs, this could (for instance) be figuring out how much you owe the phone company this month, the movement of a chess piece (physically or virtually), or a robot picking up a pencil. Only in this context can you say that the symbol manipulations have meaning.

Now, common arithmetic is one thing, but a vastly expanded concept of symbols and rules is a reasonable description of just about any computer program at some level, even if it's possible to make other interpretations of those same programs. It's an incredible eye-opener for most computer science majors when they first discover that all the math they ever learned in high school is simply a special case of some surprisingly easy to understand more general rules.[6]

So some critics of AI, most notably John Searle, professor of philosophy at the University of California at Berkeley, rightfully observe that computers, by themselves, can't "think" in this sense at all, since they don't actually mean or do anything—at best, they manipulate symbols. We're the ones associating their computations with the external world. But Searle goes further. He points out that even saying that computers are manipulating symbols is a stretch. Electrons may be floating around in circuits, but we are the ones interpreting this activity as symbol manipulation.

It's worth mentioning a more subtle argument put forth by some prominent thinkers, such as M. Ross Quillian.[7] While the symbols themselves may be devoid of any semantics, perhaps the meaning arises out of their relationships to other symbols, just as the definition of a word in a dictionary is expressed in terms of other words. While I regard this as an important insight and step forward, it seems insufficient. Aliens reading a dictionary could certainly glean a great deal about the nature of language, but it isn't going to give them a satisfactory understanding of what love is, for instance. Machine learning algorithms suffer from the same conceptual (though not practical) shortcoming—they reflect the complexity of the real world, but without some connection to that world, it's just so much unmoored structure.

Searle's arguments, and related ones by others, all make perfect intuitive sense until you apply them to people. We take it for granted that people think. But what's the difference between ideas swirling around in your brain and bytes zipping around in a computer? In both cases, information is going in, represented in some form that can plausibly be called symbolic (discrete nerve signals from your eyes, for example), getting processed, and coming back out (nerve signals to your hand to press keys on your keyboard, resulting in a spreadsheet of total monthly sales).

Searle argues that these must, in fact, be different things, but we just don't understand yet what the brain is doing. (He wisely abstains from speculating on what the actual difference is.)[8] It's important to understand what he's not saying. He isn't positing some magical property of the human mind that transcends science—his feet are firmly planted on the ground with a belief in the physical world as (mostly) deterministic, subject to measurement and rational explanation. He's just saying that something is happening in our brains that we don't understand yet, and that when we do (which he accepts as likely), it will pave the way for a satisfying explanation of what he believes are uniquely human phenomena—not just "thinking" but also consciousness, the feeling of experiencing things (what philosophers call "qualia"), sentience, and so on. He also isn't arguing that a computer program can never perform any particular task—be that to paint beautiful paintings, discover laws of nature, or console you on the loss of a loved one. But he believes that the program is *simulating* thinking, not *duplicating* the process that occurs in human minds when they engage in these activities. To Searle, a player piano isn't doing the same thing as a master musician when performing a Rachmaninoff concerto, even if it sounds the same. In short, Searle is saying that when it comes to computers, at least as they exist today, no one is home.

Despite the ongoing efforts of generations of AI researchers to explain away Searle's observations, in my opinion his basic point is right.[9] Computer programs, taken by themselves, don't really square with our commonsense intuition about what it means to think. They are "simply" carrying out logical, deterministic sequences of actions, no matter how complex, changing their internal configurations from one state to another. But here's where we get into trouble: if you believe that our brains are little more than symbol manipulators composed of biological material, then you are naturally forced to conclude that your brain, by itself, can't think either. Disconnect it from the outside world, and it would be doing just what a computer does. But that doesn't square with our commonsense intuition that even if we sit in a dark, quiet room, deprived of all input and output, we can still sit there and think. We can't have it both ways: if symbol manipulation is the basis of intelligence, either both people and machines can think (in principle, if not in practice today), or neither can.

But if you prefer to maintain the comforting conceit that we are special—different from machines in some fundamental way yet to be determined (as Searle believes), or that we are imbued with some mystical qualities quite distinct from the rest of the natural world, then you can cling to the notion that thinking is uniquely human, and machines are simply pretenders to our cognitive abilities. It's your choice. But before you make up your mind, bear in mind that there's an accumulating body of evidence chipping away at our seemingly obvious intuitions about our most quintessentially human abilities—for example, that we actually have free will.

Can a computer have free will?

Virtually everyone believes that humans, and possibly some animals, have free will, but can a machine or a computer also have free will? To answer this question, it's necessary to have some notion of what we mean by free will. There is a long

intellectual and religious history of debate about the nature and existence of free will. (Wikipedia has an excellent article reviewing the various schools of thought and major arguments.)[10] Usually what we mean is that we have the ability to make considered choices, possibly swayed but not determined by forces outside of ourselves. So the first thing to observe is that we make a distinction between inside and outside: to understand free will, we have to wrap a box around what is "us" to separate it from what is "not us." But that alone is not enough. Inside the box, we must be free to consider our options without undue influence so we can make a thoughtful choice, without having a particular conclusion preordained or forced upon us. An important consequence of this concept is that our decisions must not, in principle, be predictable. If they were, we wouldn't really be making a free choice.

Now, you might assume that computers cannot have free will because they are different from us in two key respects. First, they work according to well-understood engineering principles and so can always be predicted. Second, they can't really be said to consider choices in the same sense that people do. The problem is, both of these assertions are questionable at best.

Let's start by digging into the concept of predictability. For the purposes of this discussion I'm going to assume, as most people do (at least in contemporary Western cultures), that the physical world operates in accordance with certain laws of nature, whether or not we know or can know what those laws are. This is not to say that everything is predetermined—indeed, randomness may in fact be a fundamental part of nature. But randomness is just that—random, not a free pass for things to happen in accordance with some grander plan or principle that is somehow outside of the laws of nature. Otherwise those plans would simply be part of the laws. In other words, there is no such thing as magic. Further, I'm going to assume that your mind arises from your brain, and your brain is a physical object subject to the laws of nature.

What exactly your mind is, or how it arises from the brain, doesn't matter for this discussion, as long as you accept that it does. Another way to say this is that given a particular state of mind, there will be an equally distinct state of the brain—two different incompatible thoughts or beliefs can't arise from a single physical arrangement of matter and energy in your brain. I'm not aware of any objective evidence to the contrary, but that doesn't mean for certain that these assumptions are correct—indeed, much of the historical debate over free will focuses on precisely these assumptions, so to some degree I've baked in my conclusions by taking these positions.

Now imagine that we put you in a room, police interrogation style, with a one-way mirror on the wall so a group of very smart future scientists can observe everything about you—including the state and behavior of every neuron in your brain. We then ask you to say out loud either "red" or "blue." But before you do, we challenge the scientists to predict which you are going to pick. Running their tests, simulation models, and whatever else they want, they demonstrate that they can correctly predict what you are going to say 100 percent of the time. From this, they proudly announce that you do not have free will—after all, no matter how hard you try, you can't fool them.

But you beg to differ, and demand an opportunity to demonstrate that, in fact, you are not so dull and predictable. First, you try to decide what you're going to pick, then explicitly change your mind. This doesn't work, because, of course, the scientists are able to predict that you are going to do this. But then you get an idea. You discover that if you sit very quietly, you can hear the scientists discussing their predictions. So the next time they ask you to pick a color, you listen in on their deliberations and learn what they have predicted. Then you simply pick the other color. Stymied by your inventiveness, they incorporate this into their models—that not only do you get to pick, but also that you have access to their prediction before you do so. There's nothing uncertain or unclear about

this new wrinkle, but to their surprise, their enhanced model doesn't work. No matter how they try, you can still prove them wrong by picking the other color.

So how did you show them up? By expanding the "box" between the inside and outside of your thoughts—in this case, to include them. In short, if the box is big enough, what's inside it cannot in all circumstances predict what it will do, even though something completely outside the box can (in principle, as far as we know). As long as you can enlarge the box to include the prediction, no such prediction can always be correct.

Now, there's nothing in this argument that can't apply as well to a machine as to you. We can build a robot that does exactly what you did. No matter how we program that robot to make decisions, no matter how predictable that robot is, as long as it has access to an outside forecast of its own actions, that forecast can't always be correct. The robot can simply wait for that forecast, then do the opposite. So a sufficiently capable robot can't always be predicted, where "sufficiently capable" means it has access to the attempt to predict what it will do.

This is an example of what computer scientists call an undecidable problem—there is no effective algorithm that can solve the problem completely (meaning that it gives a correct answer in all cases). Note that this is an entirely different concept from the more widely known and similarly named uncertainty principle in physics, which states that your knowledge of both the position and momentum of a particle is limited in precision and inversely related. Undecidable problems really do exist. Probably the most famous one was formulated by none other than Alan Turing, and it is called the "halting problem." The halting problem is easy to state: can you write a program A that will examine any other program B along with its input and tell you whether or not B will eventually stop running? In other words, can A tell if B will ever finish and produce an answer? Turing showed that no such program A can exist, using an argument similar to the one above.[11]

So in practice, what actually happens? The program doesn't make a mistake—that is, give you a wrong answer. Instead, it simply never stops running. In the case of our future scientists, no matter how clever their prediction process, in some cases it will simply never reach a conclusion as to whether you are going to pick red or blue. This doesn't mean you don't get to pick your answer, just that they can't always tell in advance what you are going to pick. The scientists might cry foul, noting that they are never wrong, which is true. But you counter that never being wrong is not the same thing as being able to reliably predict your behavior.

So, it's not the case that a deterministic machine, whose behavior is completely specified and understood, can always be predicted. Any given state of a computer program may transition to its next state in an entirely predictable way, but surprisingly, we can't simply string knowledge of these states together to get a complete picture of what the program will ultimately do. And the same, of course, is true for you—in particular, you can never accurately predict your own behavior. It's possible that this is why we have the strong intuition that we have free will, but this is simply an interesting hypothesis, not a proven fact. Other possibilities are that our subjective sense of free will has arisen to serve some yet to be identified evolutionary purpose(s), like desiring sweets or being attracted to the opposite sex.

Now let's turn to the question of what it means for you to make a decision of your own volition. Just because you can make a choice doesn't mean you have free will. For instance, you could flip a coin to decide.

One of the clearest and most concise critiques of relying on chance to provide the wiggle room needed to explain free will is by contemporary thinker Sam Harris.[12] He argues that the whole idea that you can make a meaningful deliberate choice independent of outside or prior influences simply doesn't make any sense. He asks you to imagine two worlds. Both are exactly the same right up until you make a decision of

your own free will, then they diverge by virtue of your choice. In one, you choose red and in the other you choose blue. Now, in what sense did you intentionally pick one rather than the other? Your thinking was exactly the same up until that precise moment, yet somehow you made a different choice. But, you might counter, you made up your own mind. Harris would reply, based on what? Something led up to your decision, presumably internal mental deliberations—otherwise your decision was simply determined by some process that, though possibly random, does not reflect anything resembling what we mean by deliberative intent. But that means that the "red" and "blue" worlds had already diverged before you decided. So let's move the starting line back to when you began to think about the problem—maybe that's when you exercised free will. But at that point you hadn't decided anything at all—in fact, you hadn't even begun to think about it. Harris concludes, reasonably enough, that free will in the sense of intentional choice, unfettered and undetermined by previous events, is nothing more than an illusion.

Now let's look at the question of how computers make decisions. Unlike people, we have a really good idea of how they work. Nonetheless, they can make choices without relying on randomness. They can weigh evidence, apply knowledge and expertise, make decisions in the face of uncertainty, take risks, modify their plans based on new information, observe the results of their own actions, reason (in the case of symbolic processing), or use what could reasonably be called intuition (for instance, by employing machine learning to inform their actions in the absence of any deeper understanding of causal relationships). And as IBM's Watson illustrates, they are capable of using metaphor and analogy to solve problems. Now, all of my descriptions superimpose somewhat anthropomorphic interpretations on what they are doing, but that's no less reasonable than describing your deliberations even though your thoughts are ultimately represented by some particular states of your brain.

Up until fairly recently, the idea that we could have access to our own internal reflections was simply a pipe dream, so philosophers could plausibly presume that there might be something magical, mysterious, or nonphysical about our mental processes. But experimental psychologists have unearthed new and disquieting evidence that our brains make decisions before our minds are consciously aware of them, just as they regulate our blood pressure without our conscious intervention. For instance, in 2008 a group of researchers asked test subjects to freely choose whether to push a button with their left or right hands. Using an fMRI brain scanner, they were able to predict which hand the subjects would use up to ten seconds before the subjects consciously made the decision.[13] So what does this say about the box we need to draw around "us" versus the external world? As we learn more and more about how our brains—as opposed to our minds—actually work, our private, mental world would seem to be shrinking into invisibility.

So if there's no such thing as free will, why should you ever try to do anything, for instance, to lose weight? Sam Harris goes on to make the interesting observation that you may not have any meaningful choice as to whether to diet or not, but one thing for sure is that if you don't try, you won't succeed. So even if free will does not exist, it doesn't get you off the hook for trying—that just goes hand in hand with actually doing.

To summarize, it's not clear whether, or what, it means for you to have free will—lots of smart people find it plausible that your sense of choice is nothing more than an illusion. Presumably your brain, as a physical object, plays by the same rules as the rest of the physical world, and so may be subject to inspection and analysis. And if your mind arises from your brain, at some level it too must operate according to some laws of nature, whether we understand those laws yet or not. Introducing randomness into the picture doesn't get around this problem, and neither does the peculiar fact that lots of deterministic processes are nonetheless not subject to prediction, even in principle. Finally, there's no reason other than wishful thinking to suggest that machines are in this regard any different from us. This is not to say that people and machines are equivalent in all respects—they clearly aren't. But when it comes to making choices, so far, at least, there aren't good reasons to believe they operate according to different natural or scientific principles.

So we're left with the following conclusion: either both people and computers can have free will, or neither can—at least until we discover some evidence to the contrary. Take your pick.

Can a computer be conscious?

As with free will, satisfying definitions of consciousness are notoriously elusive. The more we seem to learn about brain science, the more problematic the abstract notion of consciousness becomes. Some researchers tie consciousness to the role of emotional states and physical embodiment. Others have developed evidence that blocking communications across various parts of the brain will cause consciousness to cease. Studies of patients in vegetative states suggest that consciousness is not entirely black or white but can be somewhere in between, resulting in limited awareness and ability to respond to external events. Antonio Damasio, a cognitive neuroscientist at the University of Southern California, has developed an influential theory called the "somatic marker hypothesis," which in part proposes that broad linkages across our brains and bodies are the basis of sentience.[14] Giulio Tononi, who holds the Distinguished Chair in Consciousness Science at the University of Wisconsin–Madison, believes that consciousness arises from the wide integration of information within the brain.[15]

Until we have an objective way to define and test for human consciousness other than by simply observing others, there's no rational basis for believing that people are conscious but machines cannot be. But it's equally unjustified to assert that machines can be conscious. At the present time there's no credible way to establish whether computers and animals—or other people, for that matter—experience consciousness the same way we feel that we do.

This is a serious problem. Most of us would agree that hurting or killing a conscious being against its will is morally wrong. But what if it isn't conscious? I can build a machine that objects strongly to being turned off, but does that make doing so wrong? (I will discuss this issue further in the next section.)

That said, my personal opinion is that the notion of consciousness, or subjective experience more generally, simply doesn't apply to machines. I've certainly seen no evidence of it to date. And without some definitional guideposts to point to how we might even address the question, I'm lost. It's likely that machines will, at the very least, *behave* as if they are conscious, leaving us with some difficult choices about the consequences. And our children, who likely will grow up being tenderly cared for by patient, selfless, insightful machines, may very well answer this question differently than we might today.

Can a computer "feel"?

You might have noticed a common thread so far: that the answers to our questions hinged largely on whether you regard words like *intelligence*, *thinking*, and *feeling* as connoting something sacrosanct about humans (or at least biological creatures), or whether you are comfortable expanding their applicability to certain artifacts.

In this regard, our own language is working against us. The challenge posed by AI is how to describe, and therefore how to understand and reason about, a phenomenon never before encountered in human experience—computational devices capable of perception, reasoning, and complex actions. But the words that seem to most closely fit these new developments are colored with implications about humanity's uniqueness. To put this in perspective, it's been a few hundred years or so since we last faced a serious challenge to our beliefs about our place in the universe—the theory that we descended from less capable creatures. In some quarters, this proposal did not go down well. Yet today there is widespread (though not universal) acceptance of and comfort with the idea that we originated not through some sudden divine act of intentional creation but through the process of natural selection as noted by Darwin, among others.

Okay, we're animals—so what? It turns out that this seemingly simple shift in categories is a much bigger deal than you might expect. It ignited a raging debate that is far from settled, and AI is poised to open a new frontier in that war of words. At issue is what moral obligations, if any, do we have toward other living creatures? All of a sudden, they became distant relatives, not just resources put on earth for our convenience and use. Fundamental to that question is whether other animals feel pain, and whether we have the right to inflict it on them.

The logical starting point for determining if animals feel pain is to consider how similar or different they are from us. There is an extensive scientific literature studying the physiological manifestations of pain in animals, mainly focusing on how much their reactions mirror our own.[16] As you might expect, the more closely related those animals are to humans, the more congruent their reactions. But despite this growing body of knowledge, the plain fact is that no one knows for sure. Advocates for animal rights, such as Peter Singer, point out that you can't even know for sure whether other people feel pain, though most of us, with the possible exception of psychopaths and solipsists, accept this as true. In his words: "We . . . know that the nervous systems of other animals were not artificially constructed—as a robot might be artificially constructed—to mimic the pain behavior of humans. The nervous systems of animals evolved as our own did, and in fact the evolutionary history of human beings and other animals, especially mammals, did not diverge until the central features of our nervous systems were already in existence."[17]

Many animal rights advocates take a better-safe-than-sorry approach to this question. What are the consequences of treating animals *as if* they feel pain versus the consequences of assuming they do not? In the former case, we merely impose some potentially unnecessary inconveniences and costs on ourselves, whereas in the latter case, we risk causing extreme and enduring suffering. But the underlying assumption in this debate is that the more similar animals are to us, the greater our moral obligation to act in what we perceive to be their independent interests.

Now let's apply this logic to machines. It's relatively simple to build a robot that flinches, cries out, and/or simply says, "Ouch, that hurts" when you pinch it. But as Peter Singer points out, does that say anything about whether it feels pain? Because we are able to look beyond its reactions to its internal structure, the answer is no. It reacts that way because that's what we designed it to do, not because it feels pain. (In chapter 8 I will consider the benefits and dangers of anthropomorphizing our creations.) While some people form inappropriate attachments to their possessions, such as falling in love with their cars, most of us recognize this as a misplaced application of our nurturing instinct. The tools we build are, well, tools—to be used for our betterment as we see fit. Whether those tools are simple and inanimate, like a hammer, or more complex and active, like an air-conditioner, does not seem to bear on the question. These gadgets lack the requisite breath of life to deserve moral consideration. And there's little reason to see computers as any different in this regard. Since computers are so different from us (at least today) and are designed by us for specific purposes (as opposed to naturally occurring), it seems logical to say they don't, and most likely never will, have real feelings.

Now let me convince you of the exact opposite. Imagine that you (or your spouse) give birth to a beautiful baby girl—your only child. Unfortunately, shortly after her fifth birthday, she develops a rare degenerative neurological condition that causes her brain cells to die prematurely, one by one. Luckily for her (and you), by that time the state of the art in neurological prosthetics has advanced considerably, and she is offered a novel treatment. Once every few months, you can take her to the doctor for a scan and neuronal replacement of any brain cells that have ceased to fully function in the interim. These remarkable implants, an amalgam of microscopic circuits and wires powered by body heat, precisely mirror the active properties of natural neurons. In an ingenious technique that mimics the human immune system, they are inserted intravenously, then they home in on neurons in the final stages of death, dissolving and replacing then in situ. The results are spectacular—your little girl continues to grow and thrive, experiencing all the trials and triumphs associated with a normal childhood.

After many years of regular outpatient visits no more noteworthy than regular dental checkups, the doctor informs you that there is no longer any need to continue. You ask if this means she's cured, but the answer isn't quite what you expected—the doctor nonchalantly informs you that 100 percent of her neurons have been replaced. She's a fully functioning, vivacious, and passionate teenager—apparently with an artificial brain.

Her life proceeds normally until one day, as a young adult, she enters one of her musical compositions into a prestigious competition for emerging composers. Upon learning of her childhood disability, the other contestants petition the panel of judges to disqualify her on the basis that her piece violates one of the contest rules—that all entries be composed without the assistance of computers or other artificial aids. After an all-too-brief hearing, she is referred to a parallel contest track for computer music. It pains you deeply to see your daughter so devastated. How, she cries, is she any different from the player in the violin competition who has an artificial elbow due to a skiing accident, or the one whose corneal implants permit her to sight-read without glasses?

Whether or not you concur with the judges' decision, a sober consideration of the facts unbiased by your feeling of kinship forces you to admit that they at least have a point—your daughter's brain is a man-made computing device, even if it produces normal human behavior and development in every relevant respect. Nonetheless, you would be loath to conclude that she is nothing more than a clever artifact, incapable of real feelings, undeserving of moral considerations or human rights.

So where does this leave us? On the one hand, our intuitions lead us to believe that machines, no matter how sophisticated, raise no ethical concerns in their own right. On the other, we can't comfortably exclude certain entities from the community of living things based solely on what materials they are composed of. My personal opinion, not universally shared, is that what's at issue here is little more than a decision we get to make as to whom, or to what, we choose to extend the courtesy of our empathy. Our conviction that other people or animals feel, or the fact that we love our relatives more strongly than strangers, is simply nature's way of guiding our behavior toward its own peculiar ends, an argument won not through logic and persuasion but through instinct and impulse. Though today we might be justifiably proud of our computational creations, it's hard to imagine why we should care about their welfare and achievements other than for how they benefit us. But nature has a sneaky habit of getting its way. Can machines feel? Who cares? The important question is whether highly sophisticated self-reproducing adaptive devices, which we may be in the process of creating, might inherit the earth—regardless of our role in helping this happen. Like so many species before us, we may simply be a stepping-stone to something we can't comprehend.

Notes

1. Stuart Armstrong, Kaj Sotala, and Sean S. OhEigeartaigh, "The Errors, Insights and Lessons of Famous AI Predictions—and What They Mean for the Future," Future of Humanity Institute, University of Oxford, 2014, http://www.fhi.ox.ac.uk/wp-content/uploads/FAIC.pdf.
2. Karel Čapek and Claudia Novack-Jones, *R.U.R. (Rossum's Universal Robots)* (New York: Penguin Classics, 2004).
3. Turing, A.M. (1950). "Computing machinery and intelligence," Mind, 59, 433-460, http://www.loebner.net/Prizef/TuringArticle.html.
4. If you've heard a more politically correct sanitized version of the Turing Test, namely, that it's about a machine attempting to convince a human that it is human, I encourage you to read Turing's original paper.
5. Section 6 of Turing's paper.
6. In mathematics, the study of systems of symbols and rules like these is called "abstract algebra."
7. R. Quillian, "Semantic Memory" (PhD diss., Carnegie Institute of Technology, 1966), reprinted in Marvin Minsky, *Semantic Information Processing* (Cambridge, MA: MIT Press, 2003).
8. For a short, informal expression of Searle's views, see Zan Boag, "Searle: It Upsets Me When I Read the Nonsense Written by My Contemporaries," *NewPhilosopher*, January 25, 2014, http://www.newphilosopher.com/articles/john-searle-it-upsets-me-when-i-read-the-nonsense-written-by-my-contemporaries/.
9. John Preston and Mark Bishop, eds., *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence* (Oxford: Oxford University Press, 2002).
10. http://en.wikipedia.org/wiki/Free_will.
11. The gist of Turing's argument is that there are as many different computer programs as integers, but those programs taken together behave in as many different ways as there are rational numbers—and you can't count rationals with integers. Alan Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem," *Proceedings of the London Mathematical Society*, Vol. s2–42, Issue 1, (1937): 230–65.
12. Sam Harris, *Free Will* (New York: Free Press, 2012).
13. Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes, "Unconscious Determinants of Free Decisions in the

Human Brain," *Nature Neuroscience* 11 (2008): 543–45, http://www.nature.com/neuro/journal/v11/n5/abs/nn.2112.html.

14. For instance, see Antonio Damasio, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Boston: Harcourt, 1999).
15. Giulio Tononi, *Phi: A Voyage from the Brain to the Soul* (New York: Pantheon, 2012).
16. For an excellent and very concise review of this issue, see Lynne U. Sneddon, "Can Animals Feel Pain?" http://www.wellcome.ac.uk/en/pain/microsite/culture2.html.
17. Peter Singer, *Animal Liberation*, 2nd ed. (New York: Avon Books, 1990), page 10, http://www.animal-rights-library.com/texts-m/singer03.htm.

5

ARTIFICIAL INTELLIGENCE AND THE LAW

How will AI affect the law?

AI will significantly impact a wide variety of human activities and have a dramatic influence on many fields, professions, and markets. Any attempt to catalog these would necessarily be incomplete and go quickly out of date, so I will focus on just one as an illustration: the potential effects of AI on the nature, practice, and application of the law. In this review, I will cover how AI will change the practice of law as well as the way laws will be formulated and administered, and why the emergence of AI systems will require modification and extension of current legal concepts and principles. But bear in mind that a similar analysis can be done for a broad array of fields and activities, from prospecting to plate tectonics, accounting to mathematics, traffic management to celestial dynamics, press releases to poetry.

How will AI change the practice of law?

To understand how AI is likely to impact the practice of law, it's helpful to understand how it is currently practiced, at least in the United States. The American Bar Association (ABA), an influential trade organization, was formed in 1878 by seventy-five prominent lawyers from around the country, and today