26 PHILOSOPHICAL FOUNDATIONS

In which we consider what it means to think and whether artifacts could and should ever do so.

Philosophers have been around far longer than computers and have been trying to resolve some questions that relate to AI: How do minds work? Is it possible for machines to act intelligently in the way that people do, and if they did, would they have real, conscious minds? What are the ethical implications of intelligent machines?

WEAK AI
STRONG AI

First, some terminology: the assertion that machines could act *as if* they were intelligent is called the **weak AI** hypothesis by philosophers, and the assertion that machines that do so are *actually* thinking (not just *simulating* thinking) is called the **strong AI** hypothesis.

Most AI researchers take the weak AI hypothesis for granted, and don't care about the strong AI hypothesis—as long as their program works, they don't care whether you call it a simulation of intelligence or real intelligence. All AI researchers should be concerned with the ethical implications of their work.

26.1 WEAK AI: CAN MACHINES ACT INTELLIGENTLY?

The proposal for the 1956 summer workshop that defined the field of Artificial Intelligence (McCarthy *et al.*, 1955) made the assertion that "Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it." Thus, AI was founded on the assumption that weak AI is possible. Others have asserted that weak AI is impossible: "Artificial intelligence *pursued within the cult of computationalism* stands not even a ghost of a chance of producing durable results" (Sayre, 1993).

Clearly, whether AI is impossible depends on how it is defined. In Section 1.1, we defined AI as the quest for the best agent program on a given architecture. With this formulation, AI is by definition possible: for any digital architecture with k bits of program storage there are exactly 2^k agent programs, and all we have to do to find the best one is enumerate and test them all. This might not be feasible for large k, but philosophers deal with the theoretical, not the practical.

Our definition of AI works well for the engineering problem of finding a good agent, given an architecture. Therefore, we're tempted to end this section right now, answering the title question in the affirmative. But philosophers are interested in the problem of comparing two architectures—human and machine. Furthermore, they have traditionally posed the question not in terms of maximizing expected utility but rather as, "**Can machines think**?"

CAN MACHINES THINK?

CAN SUBMARINES SWIM?

The computer scientist Edsger Dijkstra (1984) said that "The question of whether *Machines Can Think* ... is about as relevant as the question of whether *Submarines Can Swim*." The American Heritage Dictionary's first definition of *swim* is "To move through water by means of the limbs, fins, or tail," and most people agree that submarines, being limbless, cannot swim. The dictionary also defines *fly* as "To move through the air by means of wings or winglike parts," and most people agree that airplanes, having winglike parts, can fly. However, neither the questions nor the answers have any relevance to the design or capabilities of airplanes and submarines; rather they are about the usage of words in English. (The fact that ships *do* swim in Russian only amplifies this point.). The practical possibility of "thinking machines" has been with us for only 50 years or so, not long enough for speakers of English to settle on a meaning for the word "think"—does it require "a brain" or just "brain-like parts."

Alan Turing, in his famous paper "Computing Machinery and Intelligence" (1950), suggested that instead of asking whether machines can think, we should ask whether machines can pass a behavioral intelligence test, which has come to be called the **Turing Test**. The test is for a program to have a conversation (via online typed messages) with an interrogator for five minutes. The interrogator then has to guess if the conversation is with a program or a person; the program passes the test if it fools the interrogator 30% of the time. Turing conjectured that, by the year 2000, a computer with a storage of 10^9 units could be programmed well enough to pass the test. He was wrong—programs have yet to fool a sophisticated judge.

TURING TEST

On the other hand, many people *have* been fooled when they didn't know they might be chatting with a computer. The ELIZA program and Internet chatbots such as MGONZ (Humphrys, 2008) and NATACHATA have fooled their correspondents repeatedly, and the chatbot CYBERLOVER has attracted the attention of law enforcement because of its penchant for tricking fellow chatters into divulging enough personal information that their identity can be stolen. The Loebner Prize competition, held annually since 1991, is the longest-running Turing Test-like contest. The competitions have led to better models of human typing errors.

Turing himself examined a wide variety of possible objections to the possibility of intelligent machines, including virtually all of those that have been raised in the half-century since his paper appeared. We will look at some of them.

26.1.1 The argument from disability

The "argument from disability" makes the claim that "a machine can never do *X*." As examples of *X*, Turing lists the following:

> Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behavior as man, do something really new.

In retrospect, some of these are rather easy—we're all familiar with computers that "make mistakes." We are also familiar with a century-old technology that has had a proven ability to "make someone fall in love with it"—the teddy bear. Computer chess expert David Levy predicts that by 2050 people will routinely fall in love with humanoid robots (Levy, 2007). As for a robot falling in love, that is a common theme in fiction,[1] but there has been only limited speculation about whether it is in fact likely (Kim *et al.*, 2007). Programs do play chess, checkers and other games; inspect parts on assembly lines, steer cars and helicopters; diagnose diseases; and do hundreds of other tasks as well as or better than humans. Computers have made small but significant discoveries in astronomy, mathematics, chemistry, mineralogy, biology, computer science, and other fields. Each of these required performance at the level of a human expert.

Given what we now know about computers, it is not surprising that they do well at combinatorial problems such as playing chess. But algorithms also perform at human levels on tasks that seemingly involve human judgment, or as Turing put it, "learning from experience" and the ability to "tell right from wrong." As far back as 1955, Paul Meehl (see also Grove and Meehl, 1996) studied the decision-making processes of trained experts at subjective tasks such as predicting the success of a student in a training program or the recidivism of a criminal. In 19 out of the 20 studies he looked at, Meehl found that simple statistical learning algorithms (such as linear regression or naive Bayes) predict better than the experts. The Educational Testing Service has used an automated program to grade millions of essay questions on the GMAT exam since 1999. The program agrees with human graders 97% of the time, about the same level that two human graders agree (Burstein *et al.*, 2001).

It is clear that computers can do many things as well as or better than humans, including things that people believe require great human insight and understanding. This does not mean, of course, that computers use insight and understanding in performing these tasks—those are not part of *behavior*, and we address such questions elsewhere—but the point is that one's first guess about the mental processes required to produce a given behavior is often wrong. It is also true, of course, that there are many tasks at which computers do not yet excel (to put it mildly), including Turing's task of carrying on an open-ended conversation.

26.1.2 The mathematical objection

It is well known, through the work of Turing (1936) and Gödel (1931), that certain mathematical questions are in principle unanswerable by particular formal systems. Gödel's incompleteness theorem (see Section 9.5) is the most famous example of this. Briefly, for any formal axiomatic system F powerful enough to do arithmetic, it is possible to construct a so-called Gödel sentence $G(F)$ with the following properties:

- $G(F)$ is a sentence of F, but cannot be proved within F.
- If F is consistent, then $G(F)$ is true.

[1] For example, the opera Coppélia (1870), the novel *Do Androids Dream of Electric Sheep?* (1968), the movies *AI* (2001) and *Wall-E* (2008), and in song, Noel Coward's 1955 version of *Let's Do It: Let's Fall in Love* predicted "probably we'll live to see machines do it." He didn't.

Philosophers such as J. R. Lucas (1961) have claimed that this theorem shows that machines are mentally inferior to humans, because machines are formal systems that are limited by the incompleteness theorem—they cannot establish the truth of their own Gödel sentence—while humans have no such limitation. This claim has caused decades of controversy, spawning a vast literature, including two books by the mathematician Sir Roger Penrose (1989, 1994) that repeat the claim with some fresh twists (such as the hypothesis that humans are different because their brains operate by quantum gravity). We will examine only three of the problems with the claim.

First, Gödel's incompleteness theorem applies only to formal systems that are powerful enough to do arithmetic. This includes Turing machines, and Lucas's claim is in part based on the assertion that computers are Turing machines. This is a good approximation, but is not quite true. Turing machines are infinite, whereas computers are finite, and any computer can therefore be described as a (very large) system in propositional logic, which is not subject to Gödel's incompleteness theorem. Second, an agent should not be too ashamed that it cannot establish the truth of some sentence while other agents can. Consider the sentence

> J. R. Lucas cannot consistently assert that this sentence is true.

If Lucas asserted this sentence, then he would be contradicting himself, so therefore Lucas cannot consistently assert it, and hence it must be true. We have thus demonstrated that there is a sentence that Lucas cannot consistently assert while other people (and machines) can. But that does not make us think less of Lucas. To take another example, no human could compute the sum of a billion 10 digit numbers in his or her lifetime, but a computer could do it in seconds. Still, we do not see this as a fundamental limitation in the human's ability to think. Humans were behaving intelligently for thousands of years before they invented mathematics, so it is unlikely that formal mathematical reasoning plays more than a peripheral role in what it means to be intelligent.

Third, and most important, even if we grant that computers have limitations on what they can prove, there is no evidence that humans are immune from those limitations. It is all too easy to show rigorously that a formal system cannot do X, and then claim that humans *can* do X using their own informal method, without giving any evidence for this claim. Indeed, it is impossible to *prove* that humans are not subject to Gödel's incompleteness theorem, because any rigorous proof would require a formalization of the claimed unformalizable human talent, and hence refute itself. So we are left with an appeal to intuition that humans can somehow perform superhuman feats of mathematical insight. This appeal is expressed with arguments such as "we must assume our own consistency, if thought is to be possible at all" (Lucas, 1976). But if anything, humans are known to be inconsistent. This is certainly true for everyday reasoning, but it is also true for careful mathematical thought. A famous example is the four-color map problem. Alfred Kempe published a proof in 1879 that was widely accepted and contributed to his election as a Fellow of the Royal Society. In 1890, however, Percy Heawood pointed out a flaw and the theorem remained unproved until 1977.

26.1.3 The argument from informality

One of the most influential and persistent criticisms of AI as an enterprise was raised by Turing as the "argument from informality of behavior." Essentially, this is the claim that human behavior is far too complex to be captured by any simple set of rules and that because computers can do no more than follow a set of rules, they cannot generate behavior as intelligent as that of humans. The inability to capture everything in a set of logical rules is called the **qualification problem** in AI.

QUALIFICATION PROBLEM

The principal proponent of this view has been the philosopher Hubert Dreyfus, who has produced a series of influential critiques of artificial intelligence: *What Computers Can't Do* (1972), the sequel *What Computers Still Can't Do* (1992), and, with his brother Stuart, *Mind Over Machine* (1986).

The position they criticize came to be called "Good Old-Fashioned AI," or GOFAI, a term coined by philosopher John Haugeland (1985). GOFAI is supposed to claim that all intelligent behavior can be captured by a system that reasons logically from a set of facts and rules describing the domain. It therefore corresponds to the simplest logical agent described in Chapter 7. Dreyfus is correct in saying that logical agents are vulnerable to the qualification problem. As we saw in Chapter 13, probabilistic reasoning systems are more appropriate for open-ended domains. The Dreyfus critique therefore is not addressed against computers *per se*, but rather against one particular way of programming them. It is reasonable to suppose, however, that a book called *What First-Order Logical Rule-Based Systems Without Learning Can't Do* might have had less impact.

Under Dreyfus's view, human expertise does include knowledge of some rules, but only as a "holistic context" or "background" within which humans operate. He gives the example of appropriate social behavior in giving and receiving gifts: "Normally one simply responds in the appropriate circumstances by giving an appropriate gift." One apparently has "a direct sense of how things are done and what to expect." The same claim is made in the context of chess playing: "A mere chess master might need to figure out what to do, but a grandmaster just sees the board as demanding a certain move . . . the right response just pops into his or her head." It is certainly true that much of the thought processes of a present-giver or grandmaster is done at a level that is not open to introspection by the conscious mind. But that does not mean that the thought processes do not exist. The important question that Dreyfus does not answer is *how* the right move gets into the grandmaster's head. One is reminded of Daniel Dennett's (1984) comment,

> It is rather as if philosophers were to proclaim themselves expert explainers of the methods of stage magicians, and then, when we ask how the magician does the sawing-the-lady-in-half trick, they explain that it is really quite obvious: the magician doesn't really saw her in half; he simply makes it appear that he does. "But how does he do *that*?" we ask. "Not our department," say the philosophers.

Dreyfus and Dreyfus (1986) propose a five-stage process of acquiring expertise, beginning with rule-based processing (of the sort proposed in GOFAI) and ending with the ability to select correct responses instantaneously. In making this proposal, Dreyfus and Dreyfus in effect move from being AI critics to AI theorists—they propose a neural network architecture

organized into a vast "case library," but point out several problems. Fortunately, all of their problems have been addressed, some with partial success and some with total success. Their problems include the following:

1. Good generalization from examples cannot be achieved without background knowledge. They claim no one has any idea how to incorporate background knowledge into the neural network learning process. In fact, we saw in Chapters 19 and 20 that there are techniques for using prior knowledge in learning algorithms. Those techniques, however, rely on the availability of knowledge in explicit form, something that Dreyfus and Dreyfus strenuously deny. In our view, this is a good reason for a serious redesign of current models of neural processing so that they *can* take advantage of previously learned knowledge in the way that other learning algorithms do.
2. Neural network learning is a form of supervised learning (see Chapter 18), requiring the prior identification of relevant inputs and correct outputs. Therefore, they claim, it cannot operate autonomously without the help of a human trainer. In fact, learning without a teacher can be accomplished by **unsupervised learning** (Chapter 20) and **reinforcement learning** (Chapter 21).
3. Learning algorithms do not perform well with many features, and if we pick a subset of features, "there is no known way of adding new features should the current set prove inadequate to account for the learned facts." In fact, new methods such as support vector machines handle large feature sets very well. With the introduction of large Web-based data sets, many applications in areas such as language processing (Sha and Pereira, 2003) and computer vision (Viola and Jones, 2002a) routinely handle millions of features. We saw in Chapter 19 that there are also principled ways to generate new features, although much more work is needed.
4. The brain is able to direct its sensors to seek relevant information and to process it to extract aspects relevant to the current situation. But, Dreyfus and Dreyfus claim, "Currently, no details of this mechanism are understood or even hypothesized in a way that could guide AI research." In fact, the field of active vision, underpinned by the theory of information value (Chapter 16), is concerned with exactly the problem of directing sensors, and already some robots have incorporated the theoretical results obtained. STANLEY's 132-mile trip through the desert (page 28) was made possible in large part by an active sensing system of this kind.

In sum, many of the issues Dreyfus has focused on—background commonsense knowledge, the qualification problem, uncertainty, learning, compiled forms of decision making—are indeed important issues, and have by now been incorporated into standard intelligent agent design. In our view, this is evidence of AI's progress, not of its impossibility.

One of Dreyfus' strongest arguments is for situated agents rather than disembodied logical inference engines. An agent whose understanding of "dog" comes only from a limited set of logical sentences such as "$Dog(x) \Rightarrow Mammal(x)$" is at a disadvantage compared to an agent that has watched dogs run, has played fetch with them, and has been licked by one. As philosopher Andy Clark (1998) says, "Biological brains are first and foremost the control systems for biological bodies. Biological bodies move and act in rich real-world

EMBODIED COGNITION

surroundings." To understand how human (or other animal) agents work, we have to consider the whole agent, not just the agent program. Indeed, the **embodied cognition** approach claims that it makes no sense to consider the brain separately: cognition takes place within a body, which is embedded in an environment. We need to study the system as a whole; the brain augments its reasoning by referring to the environment, as the reader does in perceiving (and creating) marks on paper to transfer knowledge. Under the embodied cognition program, robotics, vision, and other sensors become central, not peripheral.

26.2 STRONG AI: CAN MACHINES REALLY THINK?

Many philosophers have claimed that a machine that passes the Turing Test would still not be *actually* thinking, but would be only a *simulation* of thinking. Again, the objection was foreseen by Turing. He cites a speech by Professor Geoffrey Jefferson (1949):

> Not until a machine could write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it.

Turing calls this the argument from **consciousness**—the machine has to be aware of its own mental states and actions. While consciousness is an important subject, Jefferson's key point actually relates to **phenomenology**, or the study of direct experience: the machine has to actually feel emotions. Others focus on **intentionality**—that is, the question of whether the machine's purported beliefs, desires, and other representations are actually "about" something in the real world.

Turing's response to the objection is interesting. He could have presented reasons that machines can in fact be conscious (or have phenomenology, or have intentions). Instead, he maintains that the question is just as ill-defined as asking, "Can machines think?" Besides, why should we insist on a higher standard for machines than we do for humans? After all, in ordinary life we never have *any* direct evidence about the internal mental states of other humans. Nevertheless, Turing says, "Instead of arguing continually over this point, it is usual to have the polite convention that everyone thinks."

Turing argues that Jefferson would be willing to extend the polite convention to machines if only he had experience with ones that act intelligently. He cites the following dialog, which has become such a part of AI's oral tradition that we simply have to include it:

> HUMAN: In the first line of your sonnet which reads "shall I compare thee to a summer's day," would not a "spring day" do as well or better?
> MACHINE: It wouldn't scan.
> HUMAN: How about "a winter's day." That would scan all right.
> MACHINE: Yes, but nobody wants to be compared to a winter's day.
> HUMAN: Would you say Mr. Pickwick reminded you of Christmas?
> MACHINE: In a way.
> HUMAN: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

> MACHINE: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

One can easily imagine some future time in which such conversations with machines are commonplace, and it becomes customary to make no linguistic distinction between "real" and "artificial" thinking. A similar transition occurred in the years after 1848, when artificial urea was synthesized for the first time by Frederick Wöhler. Prior to this event, organic and inorganic chemistry were essentially disjoint enterprises and many thought that no process could exist that would convert inorganic chemicals into organic material. Once the synthesis was accomplished, chemists agreed that artificial urea *was* urea, because it had all the right physical properties. Those who had posited an intrinsic property possessed by organic material that inorganic material could never have were faced with the impossibility of devising any test that could reveal the supposed deficiency of artificial urea.

For thinking, we have not yet reached our 1848 and there are those who believe that artificial thinking, no matter how impressive, will never be real. For example, the philosopher John Searle (1980) argues as follows:

> No one supposes that a computer simulation of a storm will leave us all wet ... Why on earth would anyone in his right mind suppose a computer simulation of mental processes actually had mental processes? (pp. 37–38)

While it is easy to agree that computer simulations of storms do not make us wet, it is not clear how to carry this analogy over to computer simulations of mental processes. After all, a Hollywood simulation of a storm using sprinklers and wind machines *does* make the actors wet, and a video game simulation of a storm *does* make the simulated characters wet. Most people are comfortable saying that a computer simulation of addition is addition, and of chess is chess. In fact, we typically speak of an *implementation* of addition or chess, not a *simulation*. Are mental processes more like storms, or more like addition?

Turing's answer—the polite convention—suggests that the issue will eventually go away by itself once machines reach a certain level of sophistication. This would have the effect of *dissolving* the difference between weak and strong AI. Against this, one may insist that there is a *factual* issue at stake: humans do have real minds, and machines might or might not. To address this factual issue, we need to understand how it is that humans have real minds, not just bodies that generate neurophysiological processes. Philosophical efforts to solve this **mind–body problem** are directly relevant to the question of whether machines could have real minds.

MIND–BODY PROBLEM

The mind–body problem was considered by the ancient Greek philosophers and by various schools of Hindu thought, but was first analyzed in depth by the 17th-century French philosopher and mathematician René Descartes. His *Meditations on First Philosophy* (1641) considered the mind's activity of thinking (a process with no spatial extent or material properties) and the physical processes of the body, concluding that the two must exist in separate realms—what we would now call a **dualist** theory. The mind–body problem faced by dualists is the question of how the mind can control the body if the two are really separate. Descartes speculated that the two might interact through the pineal gland, which simply begs the question of how the mind controls the pineal gland.

DUALISM

MONISM
PHYSICALISM

The **monist** theory of mind, often called **physicalism**, avoids this problem by asserting the mind is not separate from the body—that mental states *are* physical states. Most modern philosophers of mind are physicalists of one form or another, and physicalism allows, at least in principle, for the possibility of strong AI. The problem for physicalists is to explain how physical states—in particular, the molecular configurations and electrochemical processes of the brain—can simultaneously be **mental states**, such as being in pain, enjoying a hamburger, knowing that one is riding a horse, or believing that Vienna is the capital of Austria.

MENTAL STATES

26.2.1 Mental states and the brain in a vat

Physicalist philosophers have attempted to explicate what it means to say that a person—and, by extension, a computer—is in a particular mental state. They have focused in particular on **intentional states**. These are states, such as believing, knowing, desiring, fearing, and so on, that refer to some aspect of the external world. For example, the knowledge that one is eating a hamburger is a belief *about* the hamburger and what is happening to it.

INTENTIONAL STATE

If physicalism is correct, it must be the case that the proper description of a person's mental state is *determined* by that person's brain state. Thus, if I am currently focused on eating a hamburger in a mindful way, my instantaneous brain state is an instance of the class of mental states "knowing that one is eating a hamburger." Of course, the specific configurations of all the atoms of my brain are not essential: there are many configurations of my brain, or of other people's brain, that would belong to the same class of mental states. The key point is that the same brain state could not correspond to a fundamentally distinct mental state, such as the knowledge that one is eating a banana.

The simplicity of this view is challenged by some simple thought experiments. Imagine, if you will, that your brain was removed from your body at birth and placed in a marvelously engineered vat. The vat sustains your brain, allowing it to grow and develop. At the same time, electronic signals are fed to your brain from a computer simulation of an entirely fictitious world, and motor signals from your brain are intercepted and used to modify the simulation as appropriate.[2] In fact, the simulated life you live replicates exactly the life you would have lived, had your brain not been placed in the vat, including simulated eating of simulated hamburgers. Thus, you could have a brain state identical to that of someone who is really eating a real hamburger, but it would be literally false to say that you have the mental state "knowing that one is eating a hamburger." You aren't eating a hamburger, you have never even experienced a hamburger, and you could not, therefore, have such a mental state.

This example seems to contradict the view that brain states determine mental states. One way to resolve the dilemma is to say that the content of mental states can be interpreted from two different points of view. The "**wide content**" view interprets it from the point of view of an omniscient outside observer with access to the whole situation, who can distinguish differences in the world. Under this view, the content of mental states involves both the brain state and the environment history. **Narrow content**, on the other hand, considers only the brain state. The narrow content of the brain states of a real hamburger-eater and a brain-in-a-vat "hamburger"-"eater" is the same in both cases.

WIDE CONTENT
NARROW CONTENT

[2] This situation may be familiar to those who have seen the 1999 film *The Matrix*.

Wide content is entirely appropriate if one's goals are to ascribe mental states to others who share one's world, to predict their likely behavior and its effects, and so on. This is the setting in which our ordinary language about mental content has evolved. On the other hand, if one is concerned with the question of whether AI systems are really thinking and really do have mental states, then narrow content is appropriate; it simply doesn't make sense to say that whether or not an AI system is really thinking depends on conditions outside that system. Narrow content is also relevant if we are thinking about designing AI systems or understanding their operation, because it is the narrow content of a brain state that determines what will be the (narrow content of the) next brain state. This leads naturally to the idea that what matters about a brain state—what makes it have one kind of mental content and not another—is its functional role within the mental operation of the entity involved.

26.2.2 Functionalism and the brain replacement experiment

FUNCTIONALISM

The theory of **functionalism** says that a mental state is any intermediate causal condition between input and output. Under functionalist theory, any two systems with isomorphic causal processes would have the same mental states. Therefore, a computer program could have the same mental states as a person. Of course, we have not yet said what "isomorphic" really means, but the assumption is that there is some level of abstraction below which the specific implementation does not matter.

The claims of functionalism are illustrated most clearly by the brain replacement experiment. This thought experiment was introduced by the philosopher Clark Glymour and was touched on by John Searle (1980), but is most commonly associated with roboticist Hans Moravec (1988). It goes like this: Suppose neurophysiology has developed to the point where the input–output behavior and connectivity of all the neurons in the human brain are perfectly understood. Suppose further that we can build microscopic electronic devices that mimic this behavior and can be smoothly interfaced to neural tissue. Lastly, suppose that some miraculous surgical technique can replace individual neurons with the corresponding electronic devices without interrupting the operation of the brain as a whole. The experiment consists of gradually replacing all the neurons in someone's head with electronic devices.

We are concerned with both the external behavior and the internal experience of the subject, during and after the operation. By the definition of the experiment, the subject's external behavior must remain unchanged compared with what would be observed if the operation were not carried out.[3] Now although the presence or absence of consciousness cannot easily be ascertained by a third party, the subject of the experiment ought at least to be able to record any changes in his or her own conscious experience. Apparently, there is a direct clash of intuitions as to what would happen. Moravec, a robotics researcher and functionalist, is convinced his consciousness would remain unaffected. Searle, a philosopher and biological naturalist, is equally convinced his consciousness would vanish:

> You find, to your total amazement, that you are indeed losing control of your external behavior. You find, for example, that when doctors test your vision, you hear them say "We are holding up a red object in front of you; please tell us what you see." You want

[3] One can imagine using an identical "control" subject who is given a placebo operation, for comparison.

> to cry out "I can't see anything. I'm going totally blind." But you hear your voice saying in a way that is completely out of your control, "I see a red object in front of me." ... your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same. (Searle, 1992)

One can do more than argue from intuition. First, note that, for the external behavior to remain the same while the subject gradually becomes unconscious, it must be the case that the subject's volition is removed instantaneously and totally; otherwise the shrinking of awareness would be reflected in external behavior—"Help, I'm shrinking!" or words to that effect. This instantaneous removal of volition as a result of gradual neuron-at-a-time replacement seems an unlikely claim to have to make.

Second, consider what happens if we do ask the subject questions concerning his or her conscious experience during the period when no real neurons remain. By the conditions of the experiment, we will get responses such as "I feel fine. I must say I'm a bit surprised because I believed Searle's argument." Or we might poke the subject with a pointed stick and observe the response, "Ouch, that hurt." Now, in the normal course of affairs, the skeptic can dismiss such outputs from AI programs as mere contrivances. Certainly, it is easy enough to use a rule such as "If sensor 12 reads 'High' then output 'Ouch.' " But the point here is that, because we have replicated the functional properties of a normal human brain, we assume that the electronic brain contains no such contrivances. Then we must have an explanation of the manifestations of consciousness produced by the electronic brain that appeals only to the functional properties of the neurons. *And this explanation must also apply to the real brain, which has the same functional properties.* There are three possible conclusions:

1. The causal mechanisms of consciousness that generate these kinds of outputs in normal brains are still operating in the electronic version, which is therefore conscious.
2. The conscious mental events in the normal brain have no causal connection to behavior, and are missing from the electronic brain, which is therefore not conscious.
3. The experiment is impossible, and therefore speculation about it is meaningless.

Although we cannot rule out the second possibility, it reduces consciousness to what philosophers call an **epiphenomenal** role—something that happens, but casts no shadow, as it were, on the observable world. Furthermore, if consciousness is indeed epiphenomenal, then it cannot be the case that the subject says "Ouch" *because it hurts*—that is, because of the conscious experience of pain. Instead, the brain must contain a second, unconscious mechanism that is responsible for the "Ouch."

EPIPHENOMENON

Patricia Churchland (1986) points out that the functionalist arguments that operate at the level of the neuron can also operate at the level of any larger functional unit—a clump of neurons, a mental module, a lobe, a hemisphere, or the whole brain. That means that if you accept the notion that the brain replacement experiment shows that the replacement brain is conscious, then you should also believe that consciousness is maintained when the entire brain is replaced by a circuit that updates its state and maps from inputs to outputs via a huge lookup table. This is disconcerting to many people (including Turing himself), who have the intuition that lookup tables are not conscious—or at least, that the conscious experiences generated during table lookup are not the same as those generated during the operation of a

system that might be described (even in a simple-minded, computational sense) as accessing and generating beliefs, introspections, goals, and so on.

26.2.3 Biological naturalism and the Chinese Room

BIOLOGICAL NATURALISM

A strong challenge to functionalism has been mounted by John Searle's (1980) **biological naturalism**, according to which mental states are high-level emergent features that are caused by low-level physical processes *in the neurons*, and it is the (unspecified) properties of the neurons that matter. Thus, mental states cannot be duplicated just on the basis of some program having the same functional structure with the same input–output behavior; we would require that the program be running on an architecture with the same causal power as neurons. To support his view, Searle describes a hypothetical system that is clearly running a program and passes the Turing Test, but that equally clearly (according to Searle) does not *understand* anything of its inputs and outputs. His conclusion is that running the appropriate program (i.e., having the right outputs) is not a *sufficient* condition for being a mind.

The system consists of a human, who understands only English, equipped with a rule book, written in English, and various stacks of paper, some blank, some with indecipherable inscriptions. (The human therefore plays the role of the CPU, the rule book is the program, and the stacks of paper are the storage device.) The system is inside a room with a small opening to the outside. Through the opening appear slips of paper with indecipherable symbols. The human finds matching symbols in the rule book, and follows the instructions. The instructions may include writing symbols on new slips of paper, finding symbols in the stacks, rearranging the stacks, and so on. Eventually, the instructions will cause one or more symbols to be transcribed onto a piece of paper that is passed back to the outside world.

So far, so good. But from the outside, we see a system that is taking input in the form of Chinese sentences and generating answers in Chinese that are as "intelligent" as those in the conversation imagined by Turing.[4] Searle then argues: the person in the room does not understand Chinese (given). The rule book and the stacks of paper, being just pieces of paper, do not understand Chinese. Therefore, there is no understanding of Chinese. *Hence, according to Searle, running the right program does not necessarily generate understanding.*

Like Turing, Searle considered and attempted to rebuff a number of replies to his argument. Several commentators, including John McCarthy and Robert Wilensky, proposed what Searle calls the systems reply. The objection is that asking if the human in the room understands Chinese is analogous to asking if the CPU can take cube roots. In both cases, the answer is no, and in both cases, according to the systems reply, the entire system *does* have the capacity in question. Certainly, if one asks the Chinese Room whether it understands Chinese, the answer would be affirmative (in fluent Chinese). By Turing's polite convention, this should be enough. Searle's response is to reiterate the point that the understanding is not in the human and cannot be in the paper, so there cannot be any understanding. He seems to be relying on the argument that a property of the whole must reside in one of the parts. Yet

[4] The fact that the stacks of paper might contain trillions of pages and the generation of answers would take millions of years has no bearing on the *logical* structure of the argument. One aim of philosophical training is to develop a finely honed sense of which objections are germane and which are not.

water is wet, even though neither H nor O_2 is. The real claim made by Searle rests upon the following four axioms (Searle, 1990):

1. Computer programs are formal (syntactic).
2. Human minds have mental contents (semantics).
3. Syntax by itself is neither constitutive of nor sufficient for semantics.
4. Brains cause minds.

From the first three axioms Searle concludes that programs are not sufficient for minds. In other words, an agent running a program *might* be a mind, but it is not *necessarily* a mind just by virtue of running the program. From the fourth axiom he concludes "Any other system capable of causing minds would have to have causal powers (at least) equivalent to those of brains." From there he infers that any artificial brain would have to duplicate the causal powers of brains, not just run a particular program, and that human brains do not produce mental phenomena solely by virtue of running a program.

The axioms are controversial. For example, axioms 1 and 2 rely on an unspecified distinction between syntax and semantics that seems to be closely related to the distinction between narrow and wide content. On the one hand, we can view computers as manipulating syntactic symbols; on the other, we can view them as manipulating electric current, which happens to be what brains mostly do (according to our current understanding). So it seems we could equally say that brains are syntactic.

Assuming we are generous in interpreting the axioms, then the conclusion—that programs are not sufficient for minds—*does* follow. But the conclusion is unsatisfactory—all Searle has shown is that if you explicitly deny functionalism (that is what his axiom 3 does), then you can't necessarily conclude that non-brains are minds. This is reasonable enough—almost tautological—so the whole argument comes down to whether axiom 3 can be accepted. According to Searle, the point of the Chinese Room argument is to provide intuitions for axiom 3. The public reaction shows that the argument is acting as what Daniel Dennett (1991) calls an **intuition pump**: it amplifies one's prior intuitions, so biological naturalists are more convinced of their positions, and functionalists are convinced only that axiom 3 is unsupported, or that in general Searle's argument is unconvincing. The argument stirs up combatants, but has done little to change anyone's opinion. Searle remains undeterred, and has recently started calling the Chinese Room a "refutation" of strong AI rather than just an "argument" (Snell, 2008).

INTUITION PUMP

Even those who accept axiom 3, and thus accept Searle's argument, have only their intuitions to fall back on when deciding what entities are minds. The argument purports to show that the Chinese Room is not a mind *by virtue of running the program*, but the argument says nothing about how to decide whether the room (or a computer, some other type of machine, or an alien) is a mind *by virtue of some other reason*. Searle himself says that some machines do have minds: humans are biological machines with minds. According to Searle, human brains may or may not be running something like an AI program, but if they are, that is not the reason they are minds. It takes more to make a mind—according to Searle, something equivalent to the causal powers of individual neurons. What these powers are is left unspecified. It should be noted, however, that neurons evolved to fulfill *functional* roles—creatures

with neurons were learning and deciding long before consciousness appeared on the scene. It would be a remarkable coincidence if such neurons just happened to generate consciousness because of some causal powers that are irrelevant to their functional capabilities; after all, it is the functional capabilities that dictate survival of the organism.

In the case of the Chinese Room, Searle relies on intuition, not proof: just look at the room; what's there to be a mind? But one could make the same argument about the brain: just look at this collection of cells (or of atoms), blindly operating according to the laws of biochemistry (or of physics)—what's there to be a mind? Why can a hunk of brain be a mind while a hunk of liver cannot? That remains the great mystery.

26.2.4 Consciousness, qualia, and the explanatory gap

CONSCIOUSNESS

Running through all the debates about strong AI—the elephant in the debating room, so to speak—is the issue of **consciousness**. Consciousness is often broken down into aspects such as understanding and self-awareness. The aspect we will focus on is that of *subjective experience*: why it is that it *feels* like something to have certain brain states (e.g., while eating a hamburger), whereas it presumably does not feel like anything to have other physical states (e.g., while being a rock). The technical term for the intrinsic nature of experiences is **qualia** (from the Latin word meaning, roughly, "such things").

QUALIA

Qualia present a challenge for functionalist accounts of the mind because different qualia could be involved in what are otherwise isomorphic causal processes. Consider, for example, the **inverted spectrum** thought experiment, which the subjective experience of person X when seeing red objects is the same experience that the rest of us experience when seeing green objects, and vice versa. X still calls red objects "red," stops for red traffic lights, and agrees that the redness of red traffic lights is a more intense red than the redness of the setting sun. Yet, X's subjective experience is just different.

INVERTED SPECTRUM

Qualia are challenging not just for functionalism but for all of science. Suppose, for the sake of argument, that we have completed the process of scientific research on the brain—we have found that neural process P_{12} in neuron N_{177} transforms molecule A into molecule B, and so on, and on. There is simply no currently accepted form of reasoning that would lead from such findings to the conclusion that the entity owning those neurons has any particular subjective experience. This **explanatory gap** has led some philosophers to conclude that humans are simply incapable of forming a proper understanding of their own consciousness. Others, notably Daniel Dennett (1991), avoid the gap by denying the existence of qualia, attributing them to a philosophical confusion.

EXPLANATORY GAP

Turing himself concedes that the question of consciousness is a difficult one, but denies that it has much relevance to the practice of AI: "I do not wish to give the impression that I think there is no mystery about consciousness ... But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper." We agree with Turing—we are interested in creating programs that behave intelligently. The additional project of making them conscious is not one that we are equipped to take on, nor one whose success we would be able to determine.

26.3 The Ethics and Risks of Developing Artificial Intelligence

So far, we have concentrated on whether we *can* develop AI, but we must also consider whether we *should*. If the effects of AI technology are more likely to be negative than positive, then it would be the moral responsibility of workers in the field to redirect their research. Many new technologies have had unintended negative side effects: nuclear fission brought Chernobyl and the threat of global destruction; the internal combustion engine brought air pollution, global warming, and the paving-over of paradise. In a sense, automobiles are robots that have conquered the world by making themselves indispensable.

All scientists and engineers face ethical considerations of how they should act on the job, what projects should or should not be done, and how they should be handled. See the handbook on the *Ethics of Computing* (Berleur and Brunnstein, 2001). AI, however, seems to pose some fresh problems beyond that of, say, building bridges that don't fall down:

- People might lose their jobs to automation.
- People might have too much (or too little) leisure time.
- People might lose their sense of being unique.
- AI systems might be used toward undesirable ends.
- The use of AI systems might result in a loss of accountability.
- The success of AI might mean the end of the human race.

We will look at each issue in turn.

People might lose their jobs to automation. The modern industrial economy has become dependent on computers in general, and select AI programs in particular. For example, much of the economy, especially in the United States, depends on the availability of consumer credit. Credit card applications, charge approvals, and fraud detection are now done by AI programs. One could say that thousands of workers have been displaced by these AI programs, but in fact if you took away the AI programs these jobs would not exist, because human labor would add an unacceptable cost to the transactions. So far, automation through information technology in general and AI in particular has created more jobs than it has eliminated, and has created more interesting, higher-paying jobs. Now that the canonical AI program is an "intelligent agent" designed to assist a human, loss of jobs is less of a concern than it was when AI focused on "expert systems" designed to replace humans. But some researchers think that doing the complete job is the right goal for AI. In reflecting on the 25th Anniversary of the AAAI, Nils Nilsson (2005) set as a challenge the creation of human-level AI that could pass the employment test rather than the Turing Test—a robot that could learn to do any one of a range of jobs. We may end up in a future where unemployment is high, but even the unemployed serve as managers of their own cadre of robot workers.

People might have too much (or too little) leisure time. Alvin Toffler wrote in *Future Shock* (1970), "The work week has been cut by 50 percent since the turn of the century. It is not out of the way to predict that it will be slashed in half again by 2000." Arthur C. Clarke (1968b) wrote that people in 2001 might be "faced with a future of utter boredom, where the main problem in life is deciding which of several hundred TV channels to select."

The only one of these predictions that has come close to panning out is the number of TV channels. Instead, people working in knowledge-intensive industries have found themselves part of an integrated computerized system that operates 24 hours a day; to keep up, they have been forced to work *longer* hours. In an industrial economy, rewards are roughly proportional to the time invested; working 10% more would tend to mean a 10% increase in income. In an information economy marked by high-bandwidth communication and easy replication of intellectual property (what Frank and Cook (1996) call the "Winner-Take-All Society"), there is a large reward for being slightly better than the competition; working 10% more could mean a 100% increase in income. So there is increasing pressure on everyone to work harder. AI increases the pace of technological innovation and thus contributes to this overall trend, but AI also holds the promise of allowing us to take some time off and let our automated agents handle things for a while. Tim Ferriss (2007) recommends using automation and outsourcing to achieve a four-hour work week.

People might lose their sense of being unique. In *Computer Power and Human Reason*, Weizenbaum (1976), the author of the ELIZA program, points out some of the potential threats that AI poses to society. One of Weizenbaum's principal arguments is that AI research makes possible the idea that humans are automata—an idea that results in a loss of autonomy or even of humanity. We note that the idea has been around much longer than AI, going back at least to *L'Homme Machine* (La Mettrie, 1748). Humanity has survived other setbacks to our sense of uniqueness: *De Revolutionibus Orbium Coelestium* (Copernicus, 1543) moved the Earth away from the center of the solar system, and *Descent of Man* (Darwin, 1871) put *Homo sapiens* at the same level as other species. AI, if widely successful, may be at least as threatening to the moral assumptions of 21st-century society as Darwin's theory of evolution was to those of the 19th century.

AI systems might be used toward undesirable ends. Advanced technologies have often been used by the powerful to suppress their rivals. As the number theorist G. H. Hardy wrote (Hardy, 1940), "A science is said to be useful if its development tends to accentuate the existing inequalities in the distribution of wealth, or more directly promotes the destruction of human life." This holds for all sciences, AI being no exception. Autonomous AI systems are now commonplace on the battlefield; the U.S. military deployed over 5,000 autonomous aircraft and 12,000 autonomous ground vehicles in Iraq (Singer, 2009). One moral theory holds that military robots are like medieval armor taken to its logical extreme: no one would have moral objections to a soldier wanting to wear a helmet when being attacked by large, angry, axe-wielding enemies, and a teleoperated robot is like a very safe form of armor. On the other hand, robotic weapons pose additional risks. To the extent that human decision making is taken out of the firing loop, robots may end up making decisions that lead to the killing of innocent civilians. At a larger scale, the possession of powerful robots (like the possession of sturdy helmets) may give a nation overconfidence, causing it to go to war more recklessly than necessary. In most wars, at least one party is overconfident in its military abilities—otherwise the conflict would have been resolved peacefully.

Weizenbaum (1976) also pointed out that speech recognition technology could lead to widespread wiretapping, and hence to a loss of civil liberties. He didn't foresee a world with terrorist threats that would change the balance of how much surveillance people are willing to

accept, but he did correctly recognize that AI has the potential to mass-produce surveillance. His prediction has in part come true: the U.K. now has an extensive network of surveillance cameras, and other countries routinely monitor Web traffic and telephone calls. Some accept that computerization leads to a loss of privacy—Sun Microsystems CEO Scott McNealy has said "You have zero privacy anyway. Get over it." David Brin (1998) argues that loss of privacy is inevitable, and the way to combat the asymmetry of power of the state over the individual is to make the surveillance accessible to all citizens. Etzioni (2004) argues for a balancing of privacy and security; individual rights and community.

The use of AI systems might result in a loss of accountability. In the litigious atmosphere that prevails in the United States, legal liability becomes an important issue. When a physician relies on the judgment of a medical expert system for a diagnosis, who is at fault if the diagnosis is wrong? Fortunately, due in part to the growing influence of decision-theoretic methods in medicine, it is now accepted that negligence cannot be shown if the physician performs medical procedures that have high *expected* utility, even if the *actual* result is catastrophic for the patient. The question should therefore be "Who is at fault if the diagnosis is unreasonable?" So far, courts have held that medical expert systems play the same role as medical textbooks and reference books; physicians are responsible for understanding the reasoning behind any decision and for using their own judgment in deciding whether to accept the system's recommendations. In designing medical expert systems as agents, therefore, the actions should be thought of not as directly affecting the patient but as influencing the physician's behavior. If expert systems become reliably more accurate than human diagnosticians, doctors might become legally liable if they *don't* use the recommendations of an expert system. Atul Gawande (2002) explores this premise.

Similar issues are beginning to arise regarding the use of intelligent agents on the Internet. Some progress has been made in incorporating constraints into intelligent agents so that they cannot, for example, damage the files of other users (Weld and Etzioni, 1994). The problem is magnified when money changes hands. If monetary transactions are made "on one's behalf" by an intelligent agent, is one liable for the debts incurred? Would it be possible for an intelligent agent to have assets itself and to perform electronic trades on its own behalf? So far, these questions do not seem to be well understood. To our knowledge, no program has been granted legal status as an individual for the purposes of financial transactions; at present, it seems unreasonable to do so. Programs are also not considered to be "drivers" for the purposes of enforcing traffic regulations on real highways. In California law, at least, there do not seem to be any legal sanctions to prevent an automated vehicle from exceeding the speed limits, although the designer of the vehicle's control mechanism would be liable in the case of an accident. As with human reproductive technology, the law has yet to catch up with the new developments.

The success of AI might mean the end of the human race. Almost any technology has the potential to cause harm in the wrong hands, but with AI and robotics, we have the new problem that the wrong hands might belong to the technology itself. Countless science fiction stories have warned about robots or robot–human cyborgs running amok. Early examples

include Mary Shelley's *Frankenstein, or the Modern Prometheus* (1818)[5] and Karel Capek's play *R.U.R.* (1921), in which robots conquer the world. In movies, we have *The Terminator* (1984), which combines the cliches of robots-conquer-the-world with time travel, and *The Matrix* (1999), which combines robots-conquer-the-world with brain-in-a-vat.

It seems that robots are the protagonists of so many conquer-the-world stories because they represent the unknown, just like the witches and ghosts of tales from earlier eras, or the Martians from *The War of the Worlds* (Wells, 1898). The question is whether an AI system poses a bigger risk than traditional software. We will look at three sources of risk.

First, the AI system's state estimation may be incorrect, causing it to do the wrong thing. For example, an autonomous car might incorrectly estimate the position of a car in the adjacent lane, leading to an accident that might kill the occupants. More seriously, a missile defense system might erroneously detect an attack and launch a counterattack, leading to the death of billions. These risks are not really risks of AI systems—in both cases the same mistake could just as easily be made by a human as by a computer. The correct way to mitigate these risks is to design a system with checks and balances so that a single state-estimation error does not propagate through the system unchecked.

Second, specifying the right utility function for an AI system to maximize is not so easy. For example, we might propose a utility function designed to *minimize human suffering*, expressed as an additive reward function over time as in Chapter 17. Given the way humans are, however, we'll always find a way to suffer even in paradise; so the optimal decision for the AI system is to terminate the human race as soon as possible—no humans, no suffering. With AI systems, then, we need to be very careful what we ask for, whereas humans would have no trouble realizing that the proposed utility function cannot be taken literally. On the other hand, computers need not be tainted by the irrational behaviors described in Chapter 16. Humans sometimes use their intelligence in aggressive ways because humans have some innately aggressive tendencies, due to natural selection. The machines we build need not be innately aggressive, unless we decide to build them that way (or unless they emerge as the end product of a mechanism design that encourages aggressive behavior). Fortunately, there are techniques, such as apprenticeship learning, that allows us to specify a utility function by example. One can hope that a robot that is smart enough to figure out how to terminate the human race is also smart enough to figure out that that was not the intended utility function.

Third, the AI system's learning function may cause it to evolve into a system with unintended behavior. This scenario is the most serious, and is unique to AI systems, so we will cover it in more depth. I. J. Good wrote (1965),

ULTRAINTELLIGENT MACHINE

> Let an **ultraintelligent machine** be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.

[5] As a young man, Charles Babbage was influenced by reading *Frankenstein*.

TECHNOLOGICAL SINGULARITY

The "intelligence explosion" has also been called the **technological singularity** by mathematics professor and science fiction author Vernor Vinge, who writes (1993), "Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended." Good and Vinge (and many others) correctly note that the curve of technological progress (on many measures) is growing exponentially at present (consider Moore's Law). However, it is a leap to extrapolate that the curve will continue to a singularity of near-infinite growth. So far, every other technology has followed an S-shaped curve, where the exponential growth eventually tapers off. Sometimes new technologies step in when the old ones plateau; sometimes we hit hard limits. With less than a century of high-technology history to go on, it is difficult to extrapolate hundreds of years ahead.

Note that the concept of ultraintelligent machines assumes that intelligence is an especially important attribute, and if you have enough of it, all problems can be solved. But we know there are limits on computability and computational complexity. If the problem of defining ultraintelligent machines (or even approximations to them) happens to fall in the class of, say, NEXPTIME-complete problems, and if there are no heuristic shortcuts, then even exponential progress in technology won't help—the speed of light puts a strict upper bound on how much computing can be done; problems beyond that limit will not be solved. We still don't know where those upper bounds are.

Vinge is concerned about the coming singularity, but some computer scientists and futurists relish it. Hans Moravec (2000) encourages us to give every advantage to our "mind children," the robots we create, which may surpass us in intelligence. There is even a new word—**transhumanism**—for the active social movement that looks forward to this future in which humans are merged with—or replaced by—robotic and biotech inventions. Suffice it to say that such issues present a challenge for most moral theorists, who take the preservation of human life and the human species to be a good thing. Ray Kurzweil is currently the most visible advocate for the singularity view, writing in *The Singularity is Near* (2005):

TRANSHUMANISM

> The Singularity will allow us to transcend these limitations of our biological bodies and brain. We will gain power over our fates. Our mortality will be in our own hands. We will be able to live as long as we want (a subtly different statement from saying we will live forever). We will fully understand human thinking and will vastly extend and expand its reach. By the end of this century, the nonbiological portion of our intelligence will be trillions of trillions of times more powerful than unaided human intelligence.

Kurzweil also notes the potential dangers, writing "But the Singularity will also amplify the ability to act on our destructive inclinations, so its full story has not yet been written."

If ultraintelligent machines are a possibility, we humans would do well to make sure that we design their predecessors in such a way that they design themselves to treat us well. Science fiction writer Isaac Asimov (1942) was the first to address this issue, with his three laws of robotics:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given to it by human beings, except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

These laws seem reasonable, at least to us humans.[6] But the trick is how to implement these laws. In the Asimov story *Roundabout* a robot is sent to fetch some selenium. Later the robot is found wandering in a circle around the selenium source. Every time it heads toward the source, it senses a danger, and the third law causes it to veer away. But every time it veers away, the danger recedes, and the power of the second law takes over, causing it to veer back towards the selenium. The set of points that define the balancing point between the two laws defines a circle. This suggests that the laws are not logical absolutes, but rather are weighed against each other, with a higher weighting for the earlier laws. Asimov was probably thinking of an architecture based on control theory—perhaps a linear combination of factors—while today the most likely architecture would be a probabilistic reasoning agent that reasons over probability distributions of outcomes, and maximizes utility as defined by the three laws. But presumably we don't want our robots to prevent a human from crossing the street because of the nonzero chance of harm. That means that the negative utility for harm to a human must be much greater than for disobeying, but that each of the utilities is finite, not infinite.

FRIENDLY AI

Yudkowsky (2008) goes into more detail about how to design a **Friendly AI**. He asserts that friendliness (a desire not to harm humans) should be designed in from the start, but that the designers should recognize both that their own designs may be flawed, and that the robot will learn and evolve over time. Thus the challenge is one of mechanism design—to define a mechanism for evolving AI systems under a system of checks and balances, and to give the systems utility functions that will remain friendly in the face of such changes.

We can't just give a program a static utility function, because circumstances, and our desired responses to circumstances, change over time. For example, if technology had allowed us to design a super-powerful AI agent in 1800 and endow it with the prevailing morals of the time, it would be fighting today to reestablish slavery and abolish women's right to vote. On the other hand, if we build an AI agent today and tell it to evolve its utility function, how can we assure that it won't reason that "Humans think it is moral to kill annoying insects, in part because insect brains are so primitive. But human brains are primitive compared to my powers, so it must be moral for me to kill humans."

Omohundro (2008) hypothesizes that even an innocuous chess program could pose a risk to society. Similarly, Marvin Minsky once suggested that an AI program designed to solve the Riemann Hypothesis might end up taking over all the resources of Earth to build more powerful supercomputers to help achieve its goal. The moral is that even if you only want your program to play chess or prove theorems, if you give it the capability to learn and alter itself, you need safeguards. Omohundro concludes that "Social structures which cause individuals to bear the cost of their negative externalities would go a long way toward ensuring a stable and positive future," This seems to be an excellent idea for society in general, regardless of the possibility of ultraintelligent machines.

[6] A robot might notice the inequity that a human is allowed to kill another in self-defense, but a robot is required to sacrifice its own life to save a human.

We should note that the idea of safeguards against change in utility function is not a new one. In the *Odyssey*, Homer (ca. 700 B.C.) described Ulysses' encounter with the sirens, whose song was so alluring it compelled sailors to cast themselves into the sea. Knowing it would have that effect on him, Ulysses ordered his crew to bind him to the mast so that he could not perform the self-destructive act. It is interesting to think how similar safeguards could be built into AI systems.

Finally, let us consider the robot's point of view. If robots become conscious, then to treat them as mere "machines" (e.g., to take them apart) might be immoral. Science fiction writers have addressed the issue of robot rights. The movie *A.I.* (Spielberg, 2001) was based on a story by Brian Aldiss about an intelligent robot who was programmed to believe that he was human and fails to understand his eventual abandonment by his owner–mother. The story (and the movie) argue for the need for a civil rights movement for robots.

26.4 SUMMARY

This chapter has addressed the following issues:

- Philosophers use the term **weak AI** for the hypothesis that machines could possibly behave intelligently, and **strong AI** for the hypothesis that such machines would count as having actual minds (as opposed to simulated minds).
- Alan Turing rejected the question "Can machines think?" and replaced it with a behavioral test. He anticipated many objections to the possibility of thinking machines. Few AI researchers pay attention to the Turing Test, preferring to concentrate on their systems' performance on practical tasks, rather than the ability to imitate humans.
- There is general agreement in modern times that mental states are brain states.
- Arguments for and against strong AI are inconclusive. Few mainstream AI researchers believe that anything significant hinges on the outcome of the debate.
- Consciousness remains a mystery.
- We identified six potential threats to society posed by AI and related technology. We concluded that some of the threats are either unlikely or differ little from threats posed by "unintelligent" technologies. One threat in particular is worthy of further consideration: that ultraintelligent machines might lead to a future that is very different from today—we may not like it, and at that point we may not have a choice. Such considerations lead inevitably to the conclusion that we must weigh carefully, and soon, the possible consequences of AI research.

BIBLIOGRAPHICAL AND HISTORICAL NOTES

Sources for the various responses to Turing's 1950 paper and for the main critics of weak AI were given in the chapter. Although it became fashionable in the post-neural-network era

to deride symbolic approaches, not all philosophers are critical of GOFAI. Some are, in fact, ardent advocates and even practitioners. Zenon Pylyshyn (1984) has argued that cognition can best be understood through a computational model, not only in principle but also as a way of conducting research at present, and has specifically rebutted Dreyfus's criticisms of the computational model of human cognition (Pylyshyn, 1974). Gilbert Harman (1983), in analyzing belief revision, makes connections with AI research on truth maintenance systems. Michael Bratman has applied his "belief-desire-intention" model of human psychology (Bratman, 1987) to AI research on planning (Bratman, 1992). At the extreme end of strong AI, Aaron Sloman (1978, p. xiii) has even described as "racialist" the claim by Joseph Weizenbaum (1976) that intelligent machines can never be regarded as persons.

Proponents of the importance of embodiment in cognition include the philosophers Merleau-Ponty, whose *Phenomenology of Perception* (1945) stressed the importance of the body and the subjective interpretation of reality afforded by our senses, and Heidegger, whose *Being and Time* (1927) asked what it means to actually be an agent, and criticized all of the history of philosophy for taking this notion for granted. In the computer age, Alva Noe (2009) and Andy Clark (1998, 2008) propose that our brains form a rather minimal representation of the world, use the world itself in a just-in-time basis to maintain the illusion of a detailed internal model, use props in the world (such as paper and pencil as well as computers) to increase the capabilities of the mind. Pfeifer *et al.* (2006) and Lakoff and Johnson (1999) present arguments for how the body helps shape cognition.

The nature of the mind has been a standard topic of philosophical theorizing from ancient times to the present. In the *Phaedo*, Plato specifically considered and rejected the idea that the mind could be an "attunement" or pattern of organization of the parts of the body, a viewpoint that approximates the functionalist viewpoint in modern philosophy of mind. He decided instead that the mind had to be an immortal, immaterial soul, separable from the body and different in substance—the viewpoint of dualism. Aristotle distinguished a variety of souls (Greek $\psi\upsilon\chi\eta$) in living things, some of which, at least, he described in a functionalist manner. (See Nussbaum (1978) for more on Aristotle's functionalism.)

Descartes is notorious for his dualistic view of the human mind, but ironically his historical influence was toward mechanism and physicalism. He explicitly conceived of animals as automata, and he anticipated the Turing Test, writing "it is not conceivable [that a machine] should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as even the dullest of men can do" (Descartes, 1637). Descartes's spirited defense of the animals-as-automata viewpoint actually had the effect of making it easier to conceive of humans as automata as well, even though he himself did not take this step. The book *L'Homme Machine* (La Mettrie, 1748) did explicitly argue that humans are automata.

Modern analytic philosophy has typically accepted physicalism, but the variety of views on the content of mental states is bewildering. The identification of mental states with brain states is usually attributed to Place (1956) and Smart (1959). The debate between narrow-content and wide-content views of mental states was triggered by Hilary Putnam (1975), who introduced so-called **twin earths** (rather than brain-in-a-vat, as we did in the chapter) as a device to generate identical brain states with different (wide) content.

TWIN EARTHS

Functionalism is the philosophy of mind most naturally suggested by AI. The idea that mental states correspond to classes of brain states defined functionally is due to Putnam (1960, 1967) and Lewis (1966, 1980). Perhaps the most forceful proponent of functionalism is Daniel Dennett, whose ambitiously titled work *Consciousness Explained* (Dennett, 1991) has attracted many attempted rebuttals. Metzinger (2009) argues there is no such thing as an objective *self*, that consciousness is the subjective appearance of a world. The inverted spectrum argument concerning qualia was introduced by John Locke (1690). Frank Jackson (1982) designed an influential thought experiment involving Mary, a color scientist who has been brought up in an entirely black-and-white world. *There's Something About Mary* (Ludlow *et al.*, 2004) collects several papers on this topic.

Functionalism has come under attack from authors who claim that they do not account for the *qualia* or "what it's like" aspect of mental states (Nagel, 1974). Searle has focused instead on the alleged inability of functionalism to account for intentionality (Searle, 1980, 1984, 1992). Churchland and Churchland (1982) rebut both these types of criticism. The Chinese Room has been debated endlessly (Searle, 1980, 1990; Preston and Bishop, 2002). We'll just mention here a related work: Terry Bisson's (1990) science fiction story *They're Made out of Meat*, in which alien robotic explorers who visit earth are incredulous to find thinking human beings whose minds are made of meat. Presumably, the robotic alien equivalent of Searle believes that he can think due to the special causal powers of robotic circuits; causal powers that mere meat-brains do not possess.

Ethical issues in AI predate the existence of the field itself. I. J. Good's (1965) ultraintelligent machine idea was foreseen a hundred years earlier by Samuel Butler (1863). Written four years after the publication of Darwin's *On the Origins of Species* and at a time when the most sophisticated machines were steam engines, Butler's article on *Darwin Among the Machines* envisioned "the ultimate development of mechanical consciousness" by natural selection. The theme was reiterated by George Dyson (1998) in a book of the same title.

The philosophical literature on minds, brains, and related topics is large and difficult to read without training in the terminology and methods of argument employed. The *Encyclopedia of Philosophy* (Edwards, 1967) is an impressively authoritative and very useful aid in this process. *The Cambridge Dictionary of Philosophy* (Audi, 1999) is a shorter and more accessible work, and the online *Stanford Encyclopedia of Philosophy* offers many excellent articles and up-to-date references. The *MIT Encyclopedia of Cognitive Science* (Wilson and Keil, 1999) covers the philosophy of mind as well as the biology and psychology of mind. There are several general introductions to the philosophical "AI question" (Boden, 1990; Haugeland, 1985; Copeland, 1993; McCorduck, 2004; Minsky, 2007). *The Behavioral and Brain Sciences*, abbreviated *BBS*, is a major journal devoted to philosophical and scientific debates about AI and neuroscience. Topics of ethics and responsibility in AI are covered in the journals *AI and Society* and *Journal of Artificial Intelligence and Law*.

EXERCISES

26.1 Go through Turing's list of alleged "disabilities" of machines, identifying which have been achieved, which are achievable in principle by a program, and which are still problematic because they require conscious mental states.

26.2 Find and analyze an account in the popular media of one or more of the arguments to the effect that AI is impossible.

26.3 In the brain replacement argument, it is important to be able to restore the subject's brain to normal, such that its external behavior is as it would have been if the operation had not taken place. Can the skeptic reasonably object that this would require updating those neurophysiological properties of the neurons relating to conscious experience, as distinct from those involved in the functional behavior of the neurons?

26.4 Suppose that a Prolog program containing many clauses about the rules of British citizenship is compiled and run on an ordinary computer. Analyze the "brain states" of the computer under wide and narrow content.

26.5 Alan Perlis (1982) wrote, "A year spent in artificial intelligence is enough to make one believe in God". He also wrote, in a letter to Philip Davis, that one of the central dreams of computer science is that "through the performance of computers and their programs we will remove all doubt that there is only a chemical distinction between the living and nonliving world." To what extent does the progress made so far in artificial intelligence shed light on these issues? Suppose that at some future date, the AI endeavor has been completely successful; that is, we have build intelligent agents capable of carrying out any human cognitive task at human levels of ability. To what extent would that shed light on these issues?

26.6 Compare the social impact of artificial intelligence in the last fifty years with the social impact of the introduction of electric appliances and the internal combustion engine in the fifty years between 1890 and 1940.

26.7 I. J. Good claims that intelligence is the most important quality, and that building ultraintelligent machines will change everything. A sentient cheetah counters that "Actually speed is more important; if we could build ultrafast machines, that would change everything," and a sentient elephant claims "You're both wrong; what we need is ultrastrong machines." What do you think of these arguments?

26.8 Analyze the potential threats from AI technology to society. What threats are most serious, and how might they be combated? How do they compare to the potential benefits?

26.9 How do the potential threats from AI technology compare with those from other computer science technologies, and to bio-, nano-, and nuclear technologies?

26.10 Some critics object that AI is impossible, while others object that it is *too* possible and that ultraintelligent machines pose a threat. Which of these objections do you think is more likely? Would it be a contradiction for someone to hold both positions?