26 PHILOSOPHICAL FOUNDATIONS

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

As we mentioned in Chapter 1, philosophers have been around for much 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 minds? What are the ethical implications of intelligent machines? For the first 25 chapters of this book, we have considered questions from AI itself; now we consider the philosopher's agenda for one chapter.

First, some terminology: the assertion that machines could possibly act intelligently (or, perhaps better, 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 (as opposed to *simulating* thinking) is called the **strong AI** hypothesis.

WEAK AI

STRONG AI

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?

Some philosophers have tried to prove that AI is impossible; that machines cannot possibly act intelligently. Some have used their arguments to call for a stop to AI research:

> Artificial intelligence *pursued within the cult of cornputationalism* stands not even a ghost of a chance of producing durable results ... it is time to divert the efforts of AI researchers—and the considerable monies made available for their support—into avenues other than the computational approach. (Sayre, 1993)

Clearly, whether AI is impossible depends on how it is defined. In essence, AI is the quest for the best agent program on a given architecture. With this formulation, AI is by definition possible: for any digital architecture consisting of k bits of 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 as, "**Can** machines think?" Unfortunately, this question is ill-defined. To see why, consider the following questions:

CAN MACHINES THINK?

- Can machines fly?
- Can machines swim?

Most people agree that the answer to the first question is yes, airplanes can fly, but the answer to the second is no; boats and submarines do move through the water, but we do not call that swimming. However, neither the questions nor the answers have any impact at all on the working lives of aeronautic and naval engineers or on the users of their products. The answers have very little to do with the design or capabilities of airplanes and submarines, and much more to do with the way we have chosen to use words. The word "swim" in English has come to mean "to move along in the water by movement of body parts," whereas the word "fly" has no such limitation on the means of locomotion.[1] 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."

Alan Turing, in his famous paper "Computing Machinery and Intelligence" (Turing, 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 5 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, but he was wrong. Some people **have** been fooled for 5 minutes; for example, the ELIZA program and the Internet chatbot called MGONZ have fooled humans who didn't realize they might be talking to a program, and the program ALICE fooled one judge in the 2001 Loebner Prize competition. But no program has come close to the 30% criterion against trained judges, and the field of AI as a whole has paid little attention to Turing tests.

Turing also 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.

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:

[1] In Russian, the equivalent of "swim" ***does*** apply to ships.

> 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.

Turing had to use his intuition to guess what would be possible in the future, but we have the luxury of looking back at what computers have already done. It is undeniable that computers now do many things that previously were the domain of humans alone. Programs play chess, checkers and other games, inspect parts on assembly lines, check the spelling of word processing documents, 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.

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 $\mathbf{F}$ 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 $\mathbf{F}$, but cannot be proved within $\mathbf{F}$.
- If F is consistent, then $G(F)$ is true.

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, Godel'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. (The sentence cannot be false, because if it were then Lucas could not consistently assert it, so it would 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 10 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 mathematical reasoning plays more than a peripheral role in what it means to be intelligent.

Third, and most importantly, 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 itself contain 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.

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. (See Chapter 10.)

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), *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-FashionedAI," or GOFAI, a term coined by 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:

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 Chapter 19 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. As we saw in Chapter 19, 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, they 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.

In sum, many of the issues Dreyfus has focused on—background commonsense knowledge, the qualification problem, uncertainty, learning, compiled forms of decision making, the importance of considering situated agents rather than disembodied inference engines—have by now been incorporated into standard intelligent agent design. In our view, this is evidence of AI's progress, not of its impossibility.

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."

POLITE CONVENTION

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.

Turing 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, not in whether someone else pronounces them to be real or simulated. On the other hand, many philosophers are keenly interested in the question. To help understand it, we will consider the question of whether other artifacts are considered real.

In 1848, artificial urea was synthesized for the first time, by Frederick Wohler. This was important because it proved that organic and inorganic chemistry could be united, a question that had been hotly debated. Once the synthesis was accomplished, chemists agreed that artificial urea ***was*** urea, because it had all the right physical properties. Similarly, artificial sweeteners are undeniably sweeteners, and artificial insernination (the other AI) is undeniably insemination. On the other hand, artificial flowers are not flowers, and Daniel Dennett points out that artificial Chateau Latour wine would not be Chateau Latour wine, even if it was chemically indistinguishable, simply because it was not made in the right place in the right way. Nor is an artificial Picasso painting a Picasso painting, no matter what it looks like.

We can conclude that in some cases, the behavior of an artifact is important, while in others it is the artifact's pedigree that matters. Which one is important in which case seems to be a matter of convention. But for artificial minds, there is no convention, and we are left to rely on intuitions. The philosopher John Searle (1980) has a strong one:

> 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. Most people are comfortable saying that a computer simulation of addition is addition, and a computer simulation of a chess game is a chess game. Are mental processes more like storms, or more like addition or chess? Like Chateau Latour and Picasso, or like urea? That all depends on your theory of mental states and processes.

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; as long as the processes are isomorphic down to the this level, the same mental states will occur.

BIOLOGICAL NATURALISM

In contrast, the **biological naturalism** theory says that mental states are high-level emergent features that are caused by low-level neurological 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. The theory does not say why neurons have this causal power, nor what other physical instantiations might or might not have it.

To investigate these two viewpoints we will first look at one of the oldest problems in the philosophy of mind, and then turn to three thought experiments.

The mind–body problem

MIND–BODY PROBLEM

The **mind–body problem** asks how mental states and processes are related to bodily (specifically, brain) states and processes. As if that wasn't hard enough, we will generalize the problem to the "mind–architecture" problem, to allow us to talk about the possibility of machines having minds.

Why is the mind–body problem a problem? The first difficulty goes back to René Descartes, who considered how an immortal soul interacts with a mortal body and concluded that the soul and body are two distinct types of things—a **dualist** theory. The **monist** theory, often called **materialism,** holds that there are no such things as immaterial souls; only material objects. Consequently, mental states—such as being in pain, knowing that one is riding a horse, or believing that Vienna is the capital of Austria—are brain states. John Searle pithily

DUALISM

MONISM

MATERIALISM

sums up the idea with the slogan, "*Brains cause minds*."

FREE WILL

The materialist must face at least two serious obstacles. The first is the problem of **free will:** how can it be that a purely physical mind, whose every transformation is governed strictly by the laws of physics, still retains any freedom of choice? Most philosophers regard this problem as requiring a careful reconstitution of our naive notion of free will, rather than presenting any threat to materialism. The second problem concerns the general issue of **consciousness** (and related, but not identical, questions of **understanding** and **self-awareness).** Put simply, why is it that it *feels* like something to have certain brain states, whereas it presumably does not feel like anything to have other physical states (e.g., being a rock).

CONSCIOUSNESS

To begin to answer such questions, we need ways to talk about brain states at levels more abstract than specific configurations of all the atoms of the brain of a particular person at a particular time. For example, as I think about the capital of Austria, my brain undergoes myriad tiny changes from one picosecond to the next, but these do not constitute a *qualitative* change in brain state. To account for this, we need a notion of brain state *types,* under which we can judge whether two brain states belong to the same or different types. Various authors have various positions on what one means by *type* in this case. Almost everyone believes that if one takes a brain and replaces some of the carbon atoms by a new set of carbon atoms,[2] the mental state will not be affected. This is a good thing because real brains are continually replacing their atoms through metabolic processes, and yet this in itself does not seem to cause major mental upheavals.

INTENTIONAL STATE

Now let's consider a particular kind of mental state: the **propositional attitudes** (first discussed in Chapter 10), which are also known as **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 belief that Vienna is the capital of Austria is a belief about a particular city and its status. We will be asking whether it is possible for computers to have intentional states, so it helps to understand how to characterize such stales. For example, one might say that the mental state in which I desire a hamburger differs from the state in which I desire a pizza because hamburger and pizza are different things in the real world. That is to say, intentional states have a necessary connection to their objects in the external world. On the other hand, we argued just a few paragraphs back that mental states are brain states; hence the identity or non-identity of mental states should be determined by staying completely "inside the head," without reference to the real world. To examine this dilemma we turn to a thought experiment that attempts to separate intentional states from their external objects.

The "brain in a vat" experiment

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.[3] Then the brain could have the mental state

[2] Perhaps even atoms of a different isotope of carbon, as is sometimes done in brain-scanning experiments.

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

$DyingFor(Me,$ Hamburger) even though it has no body to feel hunger and no taste buds to experience taste, and there may be no hamburger in the real world. In that case, would this be the same mental state as one held by a brain in a body?

WIDE CONTENT

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. So under wide content the brain-in-a-vat beliefs are different from those of a "normal" person.

NARROW CONTENT

Narrow content considers only the internal subjective point of view, and under this view the beliefs would all be the same.

QUALIA

The belief that a hamburger is delicious has a certain intrinsic nature—there is something that it is like to have this belief. Now we get into the realm of **qualia**, or intrinsic experiences (from the Latin word meaning, roughly, "such things"). Suppose, through some accident of retinal and neural wiring, that person X experiences as red the color that person Y perceives as green, and vice-versa. Then when both see the same traffic light they will act the same way, but the **experience** they have will be in some way different. Both may agree that the name for their experience is "the light is red," but the experiences feel different. It is not clear whether that means they are the same or different mental states.

We now turn to another thought experiment that gets at the question of whether physical objects other than human neurons can have mental states.

The brain prosthesis experiment

The brain prosthesis experiment was introduced in the mid-1970s by Clark Glymour and was touched on by John Searle (1980), but is most commonly associated with the work of 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 and then reversing the process to return the subject to his or her normal biological state.

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.[4] 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:

[4] One can imagine using an identical "control" subject who is given a placebo operation, so that the two behaviors can be compared.

> 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 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." ... [Y]our conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same. (Searle, 1992)

But one can do more than argue from intuition. First, note that, in order 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, it seems, only two 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.

EPIPHENOMENON

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 the brain must contain a second, unconscious mechanism that is responsible for the "Ouch."

Third, consider the situation after the operation has been reversed and the subject has a normal brain. Once again, the subject's external behavior must, by definition, be as if the operation had not occurred. In particular, we should be able to ask, "What was it like during the operation? Do you remember the pointed stick?" The subject must have accurate memories of the actual nature of his or her conscious experiences, including the qualia, despite the fact that, according to Searle there were no such experiences.

Searle might reply that we have not defined the experiment properly. If the real neurons are, say, put into suspended animation between the time they are extracted and the time they are replaced in the brain, then of course they will not "remember" the experiences during

the operation. To deal with this eventuality, we need to make sure that the neurons' state is updated to reflect the internal state of the artificial neurons they are replacing. If the supposed "nonfunctional" aspects of the real neurons then result in functionally different behavior from that observed with artificial neurons still in place, then we have a simple *reductio ad absurdum,* because that would mean that the artificial neurons are not functionally equivalent to the real neurons. (See Exercise 26.3 for one possible rebuttal to this argument.)

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 prosthesis 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 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. This would suggest that the brain prosthesis experiment cannot use whole-brain-at-once replacement if it is to be effective in guiding intuitions, but it does not mean that it must use one-atom-at-a-time replacement as Searle have us believe.

The Chinese room

Our final thought experiment is perhaps the most famous of all. It is due to John Searle (1980), who 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 obviously "intelligent" as those in the conversation imagined by Turing.[5] Searle then argues as follows: the person in the room does not understand Chinese (given). The rule book and the stacks of paper, being

[5] The fact that the stacks of paper might well be larger than the entire planet 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.

just pieces of paper, do not understand Chinese. Therefore, there is no understanding of Chinese going on. *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, although one can ask if the human in the room understands Chinese, this 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 further suggests that one could imagine the human memorizing the rule book and the contents of all the stacks of paper, so that there would be nothing to have understanding *except* the human; and again, when one asks the human (in English), the reply will be in the negative.

Now we are down to the real issues. The shift from paper to memorization is a red herring, because both forms are simply physical instantiations of a running program. The real claim made by Searle rests upon the following four axioms (Searle, 1990):

1. Computer programs are formal, syntactic entities.
2. Minds have mental contents, or semantics.
3. Syntax by itself is not sufficient for semantics.
4. Brains cause minds.

From the first three axioms he 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 prograim.

The conclusions that programs are not sufficient for minds *does* follow from the axioms, if you are generous in interpreting them. 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, 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). But the reaction to his argument shows that it provides intuitions only to those who were already inclined to accept the idea that mere programs cannot generate true understanding.

To reiterate, the aim of the Chinese Room argument is to refute strong AI—the claim that running the right sort of program necessarily results in a mind. It does this by exhibiting an apparently intelligent system running the right sort of program that is, according to Searle, *demonstrably* not a mind. Searle appeals to intuition, not proof, for this part: 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?

Furthermore, when Searle admits that materials other than neurons could in principle be a mind, he weakens his argument even further, for two reasons: first, one has only Searle's intuitions (or one's own) to say that the Chinese room is not a mind, and second, even if we decide the room is not a mind, that tells us nothing about whether a program running on some other physical medium (including a computer) might be a mind.

Searle allows the logical possibility that the brain is actually implementing an AI program of the traditional sort—but the same program running on the wrong kind of machine would not be a mind. Searle has denied that he believes that "machines cannot have minds," rather, he asserts that some machines *do* have minds—humans are biological machines with minds. We are left without much guidance as to what types of machines do or do not qualify.