Chapter 4

The Turing Test

In Chapter 1, I referred to a short paper by Alan Turing that appeared in 1950 in the philosophy journal *Mind*; its title was “Computing Machinery and Intelligence.” It would not be an exaggeration to say that this paper permanently transformed the way people think about the mind-body problem. In fact, there are many who believe that Turing solved the problem in one stroke. This seems to be particularly true of computer scientists working in the field of AI, many of whom seem to think that Turing wrote in his paper everything they will ever need to know about this philosophical issue. So Turing’s position has come to have a foundational status in the AI field.

This means no more, however, than that these scientists would like to get on with the nonfoundational work without troubling about vexing philosophical arguments and counterarguments. It is a common enough posture for scientists to adopt toward the foundational aspects of their disciplines. It does not mean that we should accept that the mind-body problem is “solved.”

What, then, did Turing write that had such a galvanizing effect? I will summarize the key points here, but the original article is not especially technical and should be read by anyone who is interested in this subject.

---
The Turing Test Itself

First, Turing described something that he called the “Imitation Game.” In the Imitation Game there are three players in separate rooms who communicate by teletype machines. (We may prefer to think in terms of more modern equipment, such as computer-terminal screens.) Let’s call them “Player A,” “Player B,” and “Player C.” Players A and B are not of the same sex; Player C can be a man or a woman. The object of the game is for Player C to try to guess who is a woman and who is a man pretending to be a woman. The only information that Player C can have about Players A and B, however, is what he can get by typing questions or comments to them and reading their responses. If a player is in fact a woman, she should just reply in a natural manner to what Player C writes. If he is a man, he should attempt to answer the questions in a way that will convince Player C that he is a woman.

Having thus described the Imitation Game, Turing went on to describe a variation on it. He wrote:

We now ask the question, “What will happen when a machine takes the part of A in this game?” Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original “Can machines think?”

In the extensive literature on this article, a number of simplified variations of the latter version of the game have been proposed. One common variation has only two players, A and B. Player A’s task is to guess whether B is a human or a computer pretending to be one. This variation of the Imitation Game is what has come to be called the “Turing Test.” It’s worth taking note of the fact, however, that there is a significant difference between Turing’s own version and the variation that has taken its place in the literature. As Turing describes it, the interrogator is still trying to guess the gender of his interlocutors; he is not trying to expose a mechanical impostor and is apparently unaware of this dimension of the game. In the newer version, the interlocutor knows that Player B is either a human or a machine, and that that is precisely what is at issue. Gender is irrelevant, except insofar as it might be a source of clues for detecting humanity or its absence.

Presumably, the newer version of the game—which I shall henceforth refer to simply as the Turing Test—is more demanding of the computer, since Player A’s line of questioning is likely to be more focused on the relevant issue: Is Player B a human or a computer? In fact, Turing’s own version is confusing. He asks if the human interrogator will “decide wrongly” as often as he would if A and B were both humans. But what counts as deciding wrongly here? If a single guess consists of attributing a gender to both A and B, then C will “decide wrongly” every time, since A is neither a man nor a woman. Player C has not been asked to comment on whether he is confident that both A and B are human beings. In essence, Turing’s version of the game asks C to notice whether the game is being played according to the rules at all.

Fortunately, in his subsequent commentary, Turing himself begins to treat the game as if it were a game in which a human tries to decide whether she is interacting with another human or a machine. It is interesting, however, that the transition to what is now called the Turing Test was not explicitly noted by Turing himself.

We need to consider what would count as winning this game, or passing the Turing Test. To do so, imagine a few additional conditions. Suppose, for example, that there is a time limit of, say, a half-hour for the test. In addition, suppose that a single “test” actually includes a whole series of such half-hour interviews, with different questioners. Given these conditions, we may say that a computer “passes” the Turing Test if it performs as well as an actual human being. That is, if it is as likely that human questioners will take the programmed computer to be a human as it is that they will take a randomly selected human to be a human, we may say that in general human questioners cannot distinguish it from a human.

As simple as even this sounds, a moment’s reflection shows that there are some slippery aspects to the Turing Test. For one thing, Player A’s—the interrogator’s—strategy and decision will be strongly influenced by her general beliefs about the capabilities of computers. To see
this, think about how difficult or easy it might be for a human being to fail the Turing Test. Some humans are not very good at expressing themselves verbally (especially in writing). If Player A thinks that there are, or might be, computers that are very close to being able to perform at a human level on this test, then she might be more likely to be duped by a somewhat inept human. If, on the other hand, she is quite sure that no computer can even come close to this level of performance, then she will probably be more tolerant of B’s errors. Furthermore, somebody who has given a lot of thought to artificial intelligence is likely to have many more ideas about the sorts of errors a programmed computer, as opposed to a human, would be prone to. The most demanding version of the Turing Test, then, would have a sophisticated human (in matters concerning AI) doing the interrogating.

The “Strong AI” Thesis

Turing’s point was simple: If a computer can pass the Turing Test (although Turing himself, of course, did not call it that), then we know all that we need to know in order to decide that it is intelligent.

Although Turing himself limited his remarks to the property called “intelligence” and the question of whether machines can “think,” others have been eager to generalize his conclusions to cover any and all mental states and properties. According to this extended thesis, the Turing Test—passing computer may be said to be not just intelligent but conscious, with a mental life, ideas, beliefs, desires, and whatever else goes with having a mind. Put another way, the thesis is that the ability to pass the Turing Test is a logically sufficient condition for having a mind.

Often, when first considering this thesis, people begin to complain that no computer could ever perform indistinguishably from a human on this or that aspect of the Turing Test. They may assert, for example, that no programmed computer could possibly “get” jokes, or puns, or metaphors, or whatever. While there is a point to entertaining such objections—and we shall do so later—it is important to understand that they do not address the central philosophical point. These objections pertain to the question of whether we shall ever have programmed computers that can pass the Turing Test. This is undoubtedly an interesting and important question, since it calls upon us to consider just what kinds of capabilities a machine would have to have in order to succeed, but it is still a side-issue. Let’s assume, for the sake of the argument, that eventually some programmed computer will pass the Turing Test. The real philosophical issue is not whether we want to go along with Turing and conclude that the fact that some computer passes the Turing Test is a sufficient condition for its having a mind.

The philosopher John Searle is one of the key players in the dispute surrounding Turing’s claim, and we shall soon consider his arguments in detail. He has called Turing’s claim the “strong AI thesis,” the claim that “the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states.” What is the “right program”? How do we decide if a computer is “appropriately” programmed? Oddly, Searle makes no explicit comment on this in his original paper. In a later essay in Scientific American, however, he characterizes the approach of the strong AI community as follows:

[They believe that by designing the right programs with the right inputs and outputs, they are literally creating minds. They believe furthermore that they have a scientific test for determining success or failure: the Turing Test devised by Alan M. Turing, the founding father of artificial intelligence.]

Clearly, what Searle means by the “right program” is the program that will enable the computer to pass the Turing Test. It is an open question whether strong AI really does represent a commitment of most or many researchers in AI. Many may, in the end, be committed only to “weak AI,” which is simply the view that the study of the mind can be advanced by developing and studying computer models of various men-

CHAPTER 4. THE TURING TEST

tal processes. Although weak AI is of considerable methodological interest in cognitive science, it is not of much philosophical interest. Strong AI, on the other hand, incorporates a substantive philosophical doctrine. Because Searle's work has been so widely discussed, the "strong AI" label has gained wide usage.

The first thing to do is to consider why the strong AI thesis has any plausibility at all.

The Case for Strong AI

First, there is the matter of simplicity. The Turing Test is not complicated and the strong AI thesis does not depend upon any obscure metaphysical principles. At least, it doesn't at first glance seem to depend upon any such principles. By comparison, Cartesian dualism is much more complex, requiring the division of the world into two very different categories of being that nevertheless manage to interact causally. Simplicity is a strong point in favor of strong AI among scientists, who may be somewhat prone to distrust elaborate philosophical systems. This attitude is based on the scientists' recognition that in the past these philosophical systems seemed to block the advance of science. Cartesian dualism itself held that nothing is more evident and accessible to the mind than itself. Therefore, the notion of a "subconscious mind," as put forward by Freud and others, seemed to be a contradiction in terms. Dualism was an obstacle to the very possibility of a serious scientific investigation into the matter.

More important even than its simplicity is the fact that the strong AI thesis proposes a partially objective method for determining the presence of intelligence. Even though the interrogators make subjective judgments about the systems with which they are interacting, the system for dealing with these judgments in the aggregate is quite objective. There is no vagueness in the rules of the test and no difficulty in interpreting the outcome. The results of the procedure are publicly observable. Scientific progress in many fields has depended upon the development of such objective and public methods, so it seems appropriate to prefer them in considering questions of artificial intelligence, too.

In fact, the strong AI thesis proposes an operational definition of the mind: A mind is whatever set of functional capabilities will enable a system to behave in ways characteristic of systems already known to have minds, and those capabilities are detected by the Turing Test. This definition gains its power precisely by sidestepping any questions as to how the system achieves these capabilities. Mind is a logical function of what a system can do, regardless of how it does it, on this account. For this reason, the strong AI thesis is sometimes said to offer a "black box" theory of the mind.

"Black box" is an old engineering term used to refer to a part of a device that does some specifiable job, even though the engineer hasn't yet figured out just how she will get that part to work. In our context, the term "black box" means that we don't care about the details of the process that enables the system to pass the Turing Test. We don't care about its material composition either. Whether it is made of wet gray matter or tiny integrated circuits, it is all the same to this theory. All we care about is that it can pass for a human, under certain clearly definable and publicly observable conditions. If there are two machines that both pass the Turing Test, we don't care if they are structurally or materially similar or dissimilar; they are both intelligent systems to which all mental terms truly apply.

Another aspect of the strong AI thesis that appears to count in its favor is the apparent impartiality of the Turing Test. If Player B were an extraterrestrial instead of a computer or a human being, its possibly exotic appearance, physical composition, or biological structure would not count against it in our attempt to determine whether it has a mind. There are strong intuitive reasons for thinking that such factors shouldn't count. It would be "human chauvinism" to insist that all intelligent beings must resemble us in some physically specifiable ways. Even to insist that intelligent beings must be alive is apparently an instance of "biochauvinism." Contemporary biological science has opted for a functional definition of "alive" anyway; life is whatever performs all or most "life functions," such as assimilation, excretion, and reproduction. But what do these processes have to do with intelligence, mind, and consciousness? It seems entirely arbitrary to hold to these life functions as necessary conditions for something that is quite different from them:
mind. The fact that minds have so far only been associated with living things ought not to blind us to the possibility, at least, that they could be associated with nonliving things, or things whose status as living or nonliving cannot be decided.

The strong AI thesis purports to use the Turing Test to filter out what is irrelevant to mind and to isolate what is essential to it. In Turing’s own words, “The new problem [i.e., the Turing Test] has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man.” So the plausibility of the strong AI thesis depends upon the extent to which it does indeed capture what is essential to intelligence.

In the Turing Test, the participants exchange typed texts via some simple electronic apparatus. The fact that they are typed—on paper or on-screen—is simply a condition that eliminates the possibility of handwriting cues. The fact that they are texts is all-important. What the Turing Test in fact tests is the ability to generate “appropriate” output texts in response to input texts. If we were speaking strictly of humans, we would say that the ability to generate the appropriate texts demonstrates understanding of the input texts. We refer to these abilities together as “literacy.” In philosophy, it is often referred to as “linguistic competence.” Now, “understanding” is clearly a mental term; it refers to one of the things that we routinely accomplish by means of intelligence. If the ability to generate appropriate output really does demonstrate understanding, then the strong AI thesis must be correct.

Turing was not the first to single out linguistic competence as intimately associated with mind. Over three hundred years ago, Descartes claimed that the one thing that no automaton would never be able to do is use and understand language. The ability to use language seems to indicate intelligence in a way that nothing else does. This is the central intuition behind the strong AI thesis. In the next chapter, we shall examine this intuition in more detail. It is a strong intuition, however. Speaking and understanding in one’s native language do not feel very mechanical. When one is just learning a new language, using it may feel mechanical and rule-bound for a while, but mastery of that language consists precisely of shedding that feeling.

Another thing to note about the strong AI thesis is that it is behavioral. This means that a system’s mental capacities are defined (or at least identified) by its behavior. As I have already explained, behaviorism has had a powerful influence upon contemporary philosophy, due to its insistence that the only proper subject matter for scientific psychology is behavior, because only behavior is publicly observable and measurable. I have reviewed some of the difficulties of behaviorism, but that should not be taken to mean that strong AI was doomed from the start. The strong AI thesis is a variant of behaviorism that takes into account many of those difficulties. It is different in a number of respects from its parent doctrines. The difference lies in the restricted range of behavior that the strong AI thesis recognizes as relevant to the mind: textual output. Behaviorism in general, philosophical or psychological, makes no such restriction. Using the computer scientist’s shorthand “I/O” to stand for “input/output,” an accurate name for the philosophy of mind based upon the strong AI thesis would be “textual I/O behaviorism.” This name, however, does not appear in the literature. Instead, the label “Turing machine functionalism” is used. Searle himself remarks that,

[J]n much of AI there is a residual behaviorism or operationalism. Since appropriately programmed computers can have input-output patterns similar to those of human beings, we are tempted to postulate mental states in the computer similar to human mental states.  

Finally, in considering the case for the strong AI thesis, it is necessary to consider what sort of thesis it is. It is not a thesis about the capabilities or limitations of any contemporary or future computing machines. It is not the claim that the mind is a computer, nor that the brain is a computer, whatever such claims might mean. It is instead a thesis about the logically sufficient conditions for attributing intelligence and other mental properties to something. So it is a logical thesis, not a technological one. This will have to be kept in mind as we think about criticisms.

---

5. Turing, “Computing Machinery and Intelligence,” p. 5.

The Jukebox Argument

What is the proper way to go about criticizing something like the strong AI thesis? Some may find it counterintuitive without being able to say just why. One way to proceed is to search for counterexamples. In this case, a counterexample would be a case of something that could pass the Turing Test, but which anyone (especially a defender of the strong AI thesis) would agree lacks mental properties. What this really amounts to is testing the thesis against the set of concepts that we already have in place. These concepts structure our intuitions about things, so if we can find a counterexample we have likely found that the strong AI thesis is inconsistent with something that we already believe. The next thing to do is to isolate precisely the intuition that is the source of the conflict and weigh it against the new thesis and the arguments that go with it.

You might not immediately trust this appeal to “intuitions” at the heart of an allegedly objective investigation. Perhaps this method of doing things can be clarified by means of an interesting and widely accepted model of beliefs and belief change. The philosopher W. V. O. Quine has characterized the system of beliefs of any person as a kind of web, with each belief depending upon others. The critical process that I have been describing may be thought of as the attempt to add to, or modify, a bit of the web of belief. The search for counterexamples is a kind of “damage check.” If we find that the strong AI thesis logically forces us to discard some other beliefs that are very well secured in the web, then we have good reason to be suspicious of it. Much philosophical activity consists of such conceptual damage control.

The first counterexample that I want to present was devised by the philosopher Ned Block of MIT. First, recall a detail of the Turing Test:

---


---

It is timed. Consider that there are only a finite number of sentences that can be exchanged by the Turing Test players in a half-hour. That number is unthinkably large, to be sure, but it is finite. Imagine that each possible sentence, then, is stored on a little tape in the memory of a giant jukebox. Also in its memory is a set of instructions of the form, “If the input sentence is number 548,833,100,355, then output sentence number 1,336,396,448.” This set of instructions must also be very large. And that is the only kind of capability it has. It accepts an input sentence, looks it up in a long list, and plays back an output sentence. It acts, in effect, like a giant jukebox of “canned” sentences.

This machine would really be impossibly large; it is a pure thought experiment and should not for a moment be taken to be a practical approach to the problem. But thought experiments have their uses. We need to consider whether this sort of jukebox machine could, in principle, pass the Turing Test. One objection might be that there is more to passing the Turing Test than simply responding appropriately to the last sentence. In a real conversation, earlier sentences matter too. We often refer to things said much earlier; a machine that couldn’t do so would be easily exposed. This is a reasonable objection, but the imaginary sentence-playing jukebox can be fixed to meet it. We will still have the same library of canned sentences, but we will alter the rules a bit. They will be of the form, “If the sequence of sentences so far has been 100,245, followed by 506, followed by 1,245,747,699,012… and so on, then output sentence 3.” There should be a rule of that sort for every possible sequence of sentences producible in a half-hour Turing Test.

This modification increases the necessary size of the jukebox by many orders of magnitude, but that doesn’t matter. In our thought experiment, we are allowing ourselves as many universes as we need to build it. It should be clear that this sort of machine could and would pass the Turing Test because, in essence, every possible Turing Test conversation has been foreseen by its builders and the replies have been built in. The point, however, is that there is nothing in the least intelligent about this machine. It works on exactly the same principles as the jukebox in your local pizza parlor, a machine to which no one would be inclined to attribute intelligence. It's just a lot bigger.
It is important to consider just what it is that is missing in this Olympian jukebox in virtue of which we deny that it is intelligent. An obvious point is that this machine does not analyze the sentences presented to it, nor does it compose the output sentences. Instead, it simply matches the input to an entry in a list, retrieves the output sentence, and plays it back. It doesn’t process anything substantive in scale. When people use and understand language, on the other hand, they process utterances in a way that involves identifying the components of sentences and generating new sentences from substantive elements (such as words). This sort of processing is an important part of what we mean by “understanding.” The giant jukebox may be very good at playing back the right sentences, but it doesn’t understand anything.

If it is logically possible that there could be a machine that could pass the Turing Test without understanding anything, then simply passing the Turing Test cannot be a logically sufficient condition for understanding, which is an important part of intelligence. The strong AI thesis is therefore undermined by this counterexample. That is, Ned Block believes that if we must choose between the strong AI thesis and the denial of intelligence to the jukebox—and the thought experiment is designed to force precisely that choice upon us—then we must choose the latter. The attribution of intelligence to the jukebox is an unacceptably high intuitive price to pay for the strong AI thesis.

Some have complained that the jukebox scenario is simply too fantastic to mean anything. This machine would vastly overflow the bounds of the known universe; what good is a conclusion based upon such preposterous premises? I think one could reply that it is no more preposterous than Einstein’s thought experiments in which he asked us to imagine riding on the end of a beam of light and consider what the universe would look like from there. The logical point remains the same. Still, it is worth digressing to describe an actual computer program that operates in a manner not unlike that of our hypothetical jukebox.

The program, which was first mentioned in Chapter 1, is called ELIZA, and it is quite famous in AI circles. It was written in the 1960s by Joseph Weizenbaum, a computer scientist at MIT. (We shall meet Weizenbaum again in a later chapter.) ELIZA was created as an experiment in “natural language processing,” which is simply the attempt to get computers to deal with natural languages, like English, instead of special computer languages. Weizenbaum’s idea was to write a program that would mimic the responses of a nondirective therapist, the sort of psychotherapist who does not say much but simply echoes back what the patient has said. A session with a nondirective therapist might look like this (therapist’s remarks are in capital letters):

“I feel depressed.”
“TELL ME WHY YOU FEEL DEPRESSED.”
“I think I can’t succeed at anything.”
“YOU THINK YOU CAN’T SUCCEED AT ANYTHING.”
“Yes, that’s right.”
“PLEASE TELL ME MORE ABOUT WHY YOU FEEL DEPRESSED.”

And so on. Weizenbaum was able to write a program that was quite good at this sort of interaction. Versions of his program are now widely available on microcomputers. It is not a very large program. It takes the input sentence, searches it for cue words, tries to change verbs from first person to second person, and perhaps adds a tag such as “Please tell me more about . . .” If it can’t “parse” the input sentence, nor find an interesting cue word, it will play back one of its canned sentences, such as “What else do you want to tell me about this?”

ELIZA was a great success in that it does a creditable job of imitating the nondirective therapist. To Weizenbaum’s dismay, some people actually began to take the program seriously as a surrogate therapist. We shall return in a later chapter to the implications of that enthusiasm, but for now we may simply note that ELIZA in fact passes a restricted form of the Turing Test, as long as the human participant “plays along.” That is, it is quite easy to expose ELIZA’s limitations if you begin to engage in even very simple wordplay or unusual sentence structure. Still, the program is astonishingly good
at faking understanding, if you consider that it consists mainly of a modify-and-playback subroutine and a modest library of canned sentences. But nobody supposes for a moment that ELIZA is an intelligent program. And again, we base that judgment not so much on what ELIZA does or fails to do as on our knowledge of how ELIZA works. That shows that our judgment of ELIZA’s intelligence is intuitively informed by our knowledge of what is going on in the “black box.” The strong AI thesis, however, says that we ought not to care what goes on in there. The jukebox/ELIZA counterexample challenges just that stipulation.

It may be argued that the literate jukebox does exhibit a great deal of intelligence, but that that intelligence is embodied in a way that is different from what we are used to. On this line of reasoning, while the simple “look-up” routines involved may be entirely unsophisticated, the intelligence of the system is built into the enormous lists of conditional instructions, which cover every Turing Test contingency. Unlike the simple pizza-parlor jukebox, this set of instructions is vast and intricate.

It is certainly true that there is much intelligence embodied in the literate jukebox. But I would argue that it is not the right sort of embodiment, from the standpoint of AI. By analogy, there is a great deal of intelligence embodied in a set of encyclopedias, but nobody supposes that the volumes themselves are therefore intelligent (or possess any other mental attributes). The intelligence embodied in the jukebox is of this sort. It is static, an achievement of intelligence, but it is not itself intelligent. It doesn’t do anything intelligent; it merely acts mechanically upon some very intelligently arranged and linked lists.

To sum up, Ned Block’s argument works by forcing us to make an intuitive link between an intelligent performance and the mechanism by which that performance is produced. The next criticism of the strong AI thesis that we shall consider also forces us to look beneath the surface of Turing Test–passing behavior but is different from Block’s in that it makes use of our intuitive understanding of understanding.

John Searle’s Chinese Room Argument

John Searle coined the term “strong AI thesis” in a paper entitled “Minds, Brains, and Programs,” which appeared in an interdisciplinary journal called Behavioral and Brain Sciences. In this hotly disputed paper, Searle devises a thought experiment that, like Block’s, is both ingenious and somewhat fantastic. As we shall see, however, Searle’s argument has some additional and far-reaching implications. They are so far-reaching, indeed, that this argument has become the focal point of dispute in the philosophy of AI.

The thought experiment itself is fairly simple to present. Imagine that you are locked in a small room, seated at a desk. On the right wall is a slot through which sheets of paper are occasionally slipped to you. On these sheets of paper are various marks, straight and curved lines arranged in complex patterns. To you, these patterns are quite meaningless. On the table in front of you is a large and complex manual, written in English. As pages with markings are fed in through the slot, you look through the manual for instructions. The instructions tell you to take a fresh sheet of paper and write marks on it, somewhat similar to the ones on the sheet that was passed to you. The precise configuration of marks that you are instructed to make depends upon what marks were on the sheets that have already come in through the slot.

It’s a long and tedious process, looking up instructions in this gigantic manual, but eventually you finish and pass the new page out through a slot in the left-hand wall. Once again, none of these marks mean anything to you. You do not know that the marks on the page being passed to you are in fact questions, written in Chinese, since you are totally unfamiliar with the Chinese language. You also do not know that the marks that the manual instructs you to write on the new sheet of paper are answers to those questions. In fact, they are perfectly appropriate answers, so much so that they constitute passing the Turing Test.

---

8. John Searle, “Minds, Brains, and Programs,” Behavioral and Brain Sciences, 3 (1980), pp. 417-24; the paper is already a classic and has been reprinted in several other books.
in Chinese. For you, however, it's just so much busywork. You do not understand any of it except the instructions in the manual.

In this thought experiment, you are part of a system that passes the Turing Test and therefore, according to the strong AI thesis, understands Chinese. Furthermore, in this scenario the manual of instructions represents the alleged computer program that would enable a computer to pass the Turing Test. You are simply acting as the CPU, the central processing unit, of the computer. Any computer program, after all, could be translated into a set of English instructions for a human to follow. But keep in mind that these instructions are not a dictionary of any sort. There is nothing in the manual that says "This symbol means 'banana'." There are only instructions for locating and writing marks on paper. The central idea is that you are put in this room to do exactly what a computer would do if programmed to pass the Turing Test; you are executing a formal program. According to Searle, it doesn't matter if the program is a relatively simple one, such as ELIZA, or more sophisticated, such as the program BORIS mentioned in Chapter 1. Since you can apparently execute this program without understanding any Chinese, it must be a mistake to suppose that executing such a program is sufficient to produce understanding.

Let's review the logic of the argument. According to the strong AI thesis, anything that passes the Turing Test can be truly said to have mental states and properties, of which understanding is one. That claim entails that if a programmed computer can pass the Turing Test, then it genuinely understands what is being "said" to it and what its responses mean. Searle is arguing that, if this is the case, then anything that does just what the computer does ought to understand its input and output. But a human being placed in just this position would understand nothing, so the strong AI thesis must be false.

Searle's argument gets its force from the fact that we are all directly aware of what it is to understand language, even though we may not be able to say just what that understanding amounts to. So he does not have to appeal to a theory of understanding so much as to our intuitive understanding of understanding. We could do just what the computer does and still understand nothing, so there is no reason to suppose that the computer understands anything either, Turing Test results notwithstanding.

It is important to appreciate the difference between the Chinese Room argument and the jukebox argument. Although they both employ the logical strategy of offering counterexamples to the strong AI thesis, they are quite different in their consequences. In the jukebox argument, Block appeals to our intuitive sense that jukeboxes, no matter how large, are machines devoid of mental properties. But the argument says nothing at all about whether some other kind of Turing Test-passing machine, such as a digital computer, might be a better candidate for mental attributes. In short, the jukebox argument simply points out the logical inadequacy of the Turing Test as a sufficient condition of any mental states.

The Chinese Room argument reaches deeper. There are no canned sentences here. We shall grant that the manual provides instructions for performing a full syntactic analysis of the input sentences, breaking them down into clauses and words. And we further grant that it puts the output sentences together by means of some complex generative grammatical theory. So we are granting that these instructions allow the person in the room (or a machine programmed according to these instructions) to do some of the very kinds of things that a jukebox doesn't do. We can even grant that the "appropriately programmed" computer performs the same formal operations as the human brain (if it should turn out to be correct to describe brain processes in this way). Even granting all this, Searle argues, the system does not understand Chinese. To do that, it must be able to do more than formal operations on syntactically defined tokens.

The basic nature of digital computers as token-manipulators was explained in an earlier chapter. There, it was pointed out that what digital computers do is manipulate discrete tokens in configurations that are (or can be made to be) meaningful to people and according to rules that allow those configurations to occur in useful sequences. For the moment, let's
imagine that it is really a computer in the little room, instead of a person with a manual. The computer is able to scan the input characters and translate them into its own internal system of tokens, according to some code. Its program—the "manual"—then puts it through a long and complex series of configurations at blinding speed, and at some point composes a reply, as opposed to simply playing back.

Searle's point is that while this sort of thing may well be called "natural language processing" and may very well enable some computer to pass the Turing Test, it is not sufficient for understanding. More generally, no system has mental states solely in virtue of its capabilities as a token-manipulator, according to Searle, no matter how sophisticated the processing may be. Understanding depends upon something else entirely. The something else that understanding depends upon is, according to Searle, the "causal powers" of the system. What this might mean is a complex topic to which we shall return in a later chapter. For now, it is sufficient to note that Searle is arguing that even if it turns out to be appropriate to describe the brain as performing formal operations in the information-processing sense, it must also be doing more than that. That other part of what the brain does, he claims, is inextricably involved in its ability to produce understanding and other intentional states.

Searle is not, by the way, claiming that artificial intelligence is impossible. He is claiming that nothing, be it a computer or a brain, has mental states solely in virtue of its properties as a token-manipulator or information processor. Like Block, he is insisting that it really does matter what is happening in the "black box," but not in the sense in which some strong AI people might agree with him: They might insist that the "right" sort of program be executed there. Searle is making the point that if what is happening in there is simply the automated manipulation of tokens—no matter what the algorithm—then there is no possibility that understanding or any other mental event is occurring. It is nevertheless possible in principle on Searle's view for a man-made machine to have the right "causal powers," and therefore to enjoy mental states. I mention this because Searle has occasionally been read as claiming that only biological systems—living things—can have mental states, but this is not what he is claiming. He writes,

For any artefact that we might build which had mental states equivalent to human mental states, the implementation of a computer program would not by itself be sufficient. Rather the artefact would have to have powers equivalent to the powers of the human brain.

He is not, therefore, a biochauvinist or vitalist, despite the complaints of some of his critics.

This brief survey by no means exhausts the literature of arguments that have been proposed to discredit the Turing Test and the strong AI thesis. I have chosen these two because they are especially clear and vivid and strike at the heart of the matter. Needless to say, the community of Turing Test enthusiasts has not given up in the face of these arguments. Before bringing this chapter to a conclusion, it will be useful to consider some replies to the arguments.

Some Replies to Block and Searle

One thing that Block's and Searle's arguments have in common is that they are both fantastic. There really is not room in the known universe for the machine that Block describes, and it would take years for the person in Searle's Chinese room (given the likely complexity of a Turing Test-passing program) to compute a single response. We need to consider whether these facts are relevant to the philosophical claims that these scenarios underwrite. Or is it possible that some relevant facts are being masked by these fantastic scenarios?

One of the more eloquent writers on these subjects, and an ardent supporter of the Turing Test, is Douglas R. Hofstadter. In an imaginative dialogue written for Scientific American, he has one of his characters say,

Anybody who thinks that somehow a program could be rigged up just to pull sentences out of storage like records in a jukebox, and that this program could pass the Turing Test, has not thought very hard about it. The funny part about it is that it is just this kind of unrealizable

program that some enemies of artificial intelligence cite when arguing against the concept of the Turing Test.\textsuperscript{11}

The reference to Block's argument is clear, even though it would be irresponsible to characterize Block as an "enemy of artificial intelligence." Hofstadter apparently feels that Block is guilty of a kind of misdirection in this argument. By getting us to think of his machine as "just" a sentence-retrieval apparatus, he gets us to overlook the complexity and sophistication of the retrieval. We are inclined, according to Hofstadter, to underestimate the intelligence of such a device.

Hofstadter is quite right about this, if we consider a jukebox that has to make a \textit{decision} about which canned sentence to play back, based upon its analysis of the input sentence. Recall, however, that we amended the thought experiment somewhat to disallow this. Instead of matching single input sentences to single output sentences, the second version of jukebox matches \textit{lists} of sentences—of which the current input sentence is simply the last—against internally stored lists, and then looks up and plays back the output sentence. So all of the machine's internal rules are of the form "Add input sentence to list, find resultant list in memory, and play back linked sentence." It's just (1) update list, (2) find list in memory, and (3) output sentence. Since there is nothing particularly sophisticated about this mode of operation, it follows that the sheer scale of the imaginary machine does not mask a genuinely intelligent manner of operation.

Furthermore, Hofstadter's very reply betrays a turning away from the strong AI thesis. Remember, it is a "black box" theory, and we are not supposed to care about what enables the device to pass the Turing Test. By suggesting that Block has misdirected us in the thought experiment to overlook the (allegedly) necessarily intelligent manner of operation of the jukebox, Hofstadter is in fact basing his reply to Block upon a sneak peek into the black box. He is in effect saying, "But look, there really must be intelligent things going on in there." In defending the Turing Test against a critic, Hofstadter reveals that something other than Turing Test—passing is at the heart of his own understanding of intelligence.

There is another "misdirection" argument that Searle himself considers in his original paper. If we take seriously the idea that the instruction manual in the little room contains the equivalent of a computer program for passing the Turing Test in Chinese, then we must recognize that the person in the room is only \textit{part} of the overall system, and not necessarily the most important part. It is arguable that the Chinese-understanding of the system resides in the manual itself and that if we consider the system as a whole—and not just the person in the room—it is appropriate to attribute understanding to it.

Searle in fact considers this reply in his article; he dubs it the "Systems Reply." To answer it, he suggests an amendment of the conditions of the thought experiment. This time, let the person in the room have the entire instruction manual \textit{memorized}, so that none of the "natural language processing" takes place outside of that person's own mind. Even under these circumstances, the person would still only be manipulating meaningless marks according to complex memorized rules; she still would not understand anything in virtue of following those rules.

As I mentioned at the beginning of this section, the instruction manual for passing the Turing Test (in any language) would be incredibly large and complex. Producing a single output sentence might involve carrying out millions or billions of instructions. The idea of memorizing this manual is indeed fantastic—fantastic in terms of scale, but it is not clear that this is philosophically objectionable.

The literature of comments for and against Searle's Chinese Room argument is extensive. Some of them will serve as focal points of subsequent chapters of this book. Before moving on to those chapters, however, I want to consider one argument in more detail. It anticipates some problems that will be addressed in later chapters. Working through this argument here will help to provide a context for those problems later.

The Robot Reply

I have already pointed out that the strong AI thesis could be dubbed "textual I/O behaviorism." We have also taken note of the fact that the isolation of a restricted form of linguistic competence as peculiarly important to deciding about mental states has a long history. It is appropriate to wonder if this isn't taking too narrow a view.

I once was involved in a long discussion about the adequacy of the Turing Test and the soundness of Searle's argument. The discussion took place on an electronic network service (called Usenet) rather than in person. For those who are unfamiliar with this sort of thing, one "posts" a message, which is relayed by telephone to other network sites where it can be read by anyone who is interested. These people can then post their replies, which are in turn circulated around, and the discussion is propagated. At one point, I was accused of inconsistency. How could I question the adequacy of the Turing Test when the very electronic medium that we were using closely resembled it? Wasn't I attributing all manner of mental states to my interlocutors simply in virtue of exchanging typed texts with them?

Indeed I was, but only because a number of other presumptions were already firmly in place. It was (and is) part of my general knowledge that there are no programs that come close to passing the Turing Test, much less engaging in philosophical discourse. The very idea of a test presupposes some plausible doubt in the matter. Where there was no doubt, I could not be said to be testing anyone. The same reply goes to those who claim that we are always Turing Testing each other in our everyday interactions. This is false, unless we employ a much looser and vaguer sense of the Turing Test than what Turing himself had in mind.

If the Turing Test is not penetrating enough to determine whether a system understands Chinese or anything else, what would be? What is involved in human understanding beyond the manipulation of texts? Although we shall look into these matters in more detail in the next chapter, a few points are salient.

---

12. I thank my Usenet interlocutors, Jim Baler, Paul Torek, and Michael Ellis, for their insight and candor. I hope that I may someday meet them!

---

THE ROBOT REPLY

A great deal of what we call understanding is grounded in experience. We don't just understand language; we understand what is said in language, and we do so at least partly in virtue of having the right kinds of experiences. That we are the sorts of beings who can do this is surely not incidental to our understanding of language. Linguistic meaning is grounded in nonlinguistic experience. This is how language acquires a semantics.

Searle writes,

The fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. The computer, to repeat, has a syntax but no semantics.13

We have already seen that a token in a computer's memory is only a "symbol" in virtue of its having a meaning assigned to it by a person. Tokens are just objects; symbols are objects that have content, meaning, or reference. They are about something. This "aboutness" is called intentionality by philosophers. The numbers that appear in the display window of your pocket calculator refer to your bank balance only because you are using them for that purpose. Their intentionality is derived from yours. They refer because you mean them to. But what is it for you to mean or refer in that way? Your intentionality is apparently not derived. Your mental states do not refer to things simply in virtue of being used for that purpose by someone else. Their intentionality is somehow intrinsic to you. How is nonderived or intrinsic intentionality possible?

This is one of the central questions in contemporary philosophy of mind. There is much dispute not only about how intrinsic intentionality is possible but whether it exists at all and, if it does, what sort of thing it is. I cannot give it a full treatment here, but what I said above seems relevant. Understanding (an intentional state) a language presupposes some mode of mapping between a system's linguistic and nonlinguistic processes. For this even to be possible it must be capable of having nonlinguistic processes. The Turing Test itself says nothing of whether

the system is even capable of any mode of nonlinguistic processing. I use the word "processing" instead of "experience" here so as not to complicate the matter any further by forcing the question of which processes are experiences. The difficulty of that question will be discussed in the next chapter.

A natural move at this point is to equip our Turing Test-passing computer with some sensory and motor peripherals. Let it perceive the world and interact with it, and let its interactions form and update the data structures that it uses to generate its clever Turing Test-passing texts. In short, let it be a robot and not just a computer in a box in a room. Searle considers this possibility, which he calls the "Robot Reply":

Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather it would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating, drinking—anything you like. The robot would, for example, have a television camera attached to it that enabled it to see, it would have arms and legs that enabled it to act, and all of this would be controlled by its computer brain. Such a robot would . . . have genuine understanding and other mental states.14

He considers this possibility as a counterargument to his own position, but goes on to reject it. The robot, according to Searle, is in no better position to understand Chinese than its room-bound predecessor.

Suppose that instead of the computer inside the robot, you put me inside the room and you give me again, as in the original Chinese case, more Chinese symbols with more instructions in English for matching Chinese symbols to Chinese symbols and feeding back Chinese symbols to the outside. Suppose unknown to me, some of the Chinese symbols that come to me come from a television camera attached to the robot, and other Chinese symbols that I am giving out serve to make the motors inside the robot move the robot's legs or

arms. It is important to emphasize that all I am doing is manipulating formal symbols: I know none of these other facts.15

There are some odd things about Searle's rebuttal of the Robot Reply. First, he depicts the tokens coming in from the sensors and those going out to the limbs as more "Chinese symbols," just like the Chinese symbols involved in the original scenario. This is certainly a mistake, and it appears to overlook something crucial. When you hear a voice speaking English you are first of all hearing sound, just as you are hearing sound when you listen to surf crashing on the beach. The spoken English is message-bearing sound; it consists of symbols. The surf noise is just sound. One of the first things that your brain must do is separate the message-bearing sound from all the other sounds, the background noise.16 It sorts the auditory input into language and nonlanguage. How things go from there to constitute understanding is still unknown, but I cannot believe that that first step is insignificant. Searle's remark misses it altogether.

The second odd thing about Searle's rebuttal is that it misses the point that even if sensorimotor capabilities are not sufficient for understanding, they are surely at least necessary. Since the Turing Test does not test for these capabilities, the Robot Reply is actually another argument against the adequacy of the Turing Test.

Psychologist Steven Harnad has developed this line of argument, as in the following passage:

Who is to say that the Turing Test, whether conducted in Chinese or in any other language, could be successfully passed without operations that draw on our sensory, motor, and other higher cognitive capacities as well? Where does the capacity to comprehend Chinese begin and the rest of our mental competence leave off? Searle has made the implicit assumption here—one that he shares with his


15. Ibid.

16. Of course, another possibility is that you are listening intently to the surf sounds and someone's talking is the background noise. The ability to shift background and foreground is yet another subtle aspect of our cognitive repertoire that we take for granted.
opponents in AI—that there could exist a self-sufficient “module” that was able to pass his purely verbal Turing Test without simultaneously being able to do everything else we can do, i.e., without also being able to pass the Total Turing Test.\textsuperscript{17}

The Total Turing Test tests the full range of capabilities, not just the ability to generate appropriate texts. Its use suggests a form of behaviorism, except it does not require reducing the meaning of statements about mental states to statements about behavior. Harnad dubs his theory “robotic functionalism.” An important part of the theory is that the sort of processing that could plausibly explain mental states is not properly “symbolic” processing at all. We shall return to this idea in Chapter 6.

This is by no means the end of the responses to Searle or to strong AI. To some extent, the next two chapters are continuations of the responses to Searle. The Chinese Room is never, it seems, quite closed. Instead, it represents a core problem, and it is to Searle’s credit that he has zeroed in on what is most philosophically problematic in AI. A deeper and more careful consideration of the strengths and weaknesses of some points of his argument will have to wait until we have clarified some other issues pertaining to mind and language. In particular, we need a more precise vocabulary for talking about intelligence and other mental properties. To that task we turn in the next chapter.

\textsuperscript{17} Steven Harnad, “Minds, Machines and Searle,” \textit{Journal of Experimental and Theoretical Artificial Intelligence}, 1, no. 1 (1989), 8–9. (ms.)


Chapter 5
The Nature of Intelligence

When artificial intelligence is discussed by laypersons, it is usually taken for granted that intelligence itself is relatively well understood. It is often assumed that intelligence is a unitary concept, embodying some single coherent meaning, even though it may not be completely understood. In fact, intelligence is anything but a unitary concept. The word “intelligence” is quite vague, encompassing a wide range of human and nonhuman capabilities, talents, and performances. Many workers in AI are aware of this vagueness and, in their research, single out one or more precise aspects of intelligence to attempt to model. This approach has its own difficulties, as we shall see later. But this point is often overlooked in philosophical discussions of artificial intelligence. The purpose of this chapter is to disambiguate the word “intelligence” and to discuss the possibilities of, and obstacles to, artificial realizations of the various aspects of intelligence.

In 1972 Hubert Dreyfus, a philosopher at the University of California at Berkeley, wrote a book called \textit{What Computers Can’t Do}.\textsuperscript{1} This book enraged many workers in AI, including philosophers working in the field. It has become something of a classic in the literature, if only because so many people felt compelled to respond to Dreyfus’s arguments. I shall adopt a less polemical tone in this chapter, but I shall cover some of the same ground that Dreyfus covered and make explicit reference to his ideas. Dreyfus was at great pains to sort out the many aspects
it doesn’t “know” it is creating you in the process, but there you are, emerging from its frantic activity almost magically.

This process of creating a self at one level out of the relatively mindless and uncomprehending activities amalgamated at another level is vividly illustrated in the next selection by John Searle, though he firmly resists that vision of what he is showing.

D.C.D.

What psychological and philosophical significance should we attach to recent efforts at computer simulations of human cognitive capacities? In answering this question, I find it useful to distinguish what I will call “strong” AI from “weak” or “cautious” AI (artificial intelligence). According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states. In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations.

I have no objection to the claims of weak AI, at least as far as this article is concerned. My discussion here will be directed at the claims I have defined as those of strong AI, specifically the claim that the appropriately programmed computer literally has cognitive states and that the
programs thereby explain human cognition. When I hereafter refer to AI, I have in mind the strong version, as expressed by these two claims.

I will consider the work of Roger Schank and his colleagues at Yale (Schank and Abelson 1977), because I am more familiar with it than I am with any other similar claims, and because it provides a very clear example of the sort of work I wish to examine. But nothing that follows depends upon the details of Schank’s programs. The same arguments would apply to Winograd’s SHRDLU (Winograd 1973), Weizenbaum’s ELIZA (Weizenbaum 1965), and indeed any Turing machine simulation of human mental phenomena. See “Further Reading” for Searle’s references.

Very briefly, and leaving out the various details, one can describe Schank’s program as follows: The aim of the program is to simulate the human ability to understand stories. It is characteristic of human beings’ story-understanding capacity that they can answer questions about the story even though the information that they give was never explicitly stated in the story. Thus, for example, suppose you are given the following story: “A man went into a restaurant and ordered a hamburger. When the hamburger arrived it was burned to a crisp, and the man stormed out of the restaurant angry, without paying for the hamburger or leaving a tip.” Now, if you are asked “Did the man eat the hamburger?” you will presumably answer, “No, he did not.” Similarly, if you are given the following story: “A man went into a restaurant and ordered a hamburger; when the hamburger came he was very pleased with it; and as he left the restaurant he gave the waitress a large tip before paying his bill;” and you are asked the question, “Did the man eat the hamburger?” you will presumably answer, “Yes, he ate the hamburger.” Now Schank’s machines can similarly answer questions about restaurants in this fashion. To do this, they have a “representation” of the sort of information that human beings have about restaurants, which enables them to answer such questions as those above, given these sorts of stories. When the machine is given the story and then asked the question, the machine will print out answers of the sort that we would expect human beings to give if told similar stories. Partisans of strong AI claim that in this question and answer sequence the machine is not only simulating a human ability but also (1) that the machine can literally be said to understand the story and provide the answers to questions, and (2) that what the machine and its program do explain the human ability to understand the story and answer questions about it.

Both claims seem to me to be totally unsupported by Schank’s work, as I will attempt to show in what follows. I am not, of course, saying that Schank himself is committed to these claims.

One way to test any theory of the mind is to ask oneself what it would be like if my mind actually worked on the principles that the theory says all minds work on. Let us apply this test to the Schank program with the following Gedankenexperiment. Suppose that I’m locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I’m not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that “formal” means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch a “script,” they call the second batch a “story,” and they call the third batch “questions.” Furthermore, they call the symbols I give them back in response to the third batch “answers to the questions,” and the set of rules in English that they gave me, they call the “program.” Now just to complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view—that is, from the point of view of somebody outside the room in which I am locked—my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don’t speak a word of Chinese. Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker. From the external point of view—from the point of view of someone reading my “answers”—the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninter-
pretted formal symbols. As far as the Chinese is concerned, I simply
behave like a computer: I perform computational operations on formally
specified elements. For the purposes of the Chinese, I am simply an
instantiation of the computer program.

Now the claims made by strong AI are that the programmed com-
puter understands the stories and that the program in some sense ex-
plains human understanding. But we are now in a position to examine
these claims in light of our thought experiment.

1. As regards the first claim, it seems to me quite obvious in the
example that I do not understand a word of the Chinese stories. I have
inputs and outputs that are indistinguishable from those of the native
Chinese speaker, and I can have any formal program you like, but I still
understand nothing. For the same reasons, Schank’s computer under-
stands nothing of any stories, whether in Chinese, English, or whatever,
since in the Chinese case the computer is me, and in cases where the
computer is not me, the computer has nothing more than I have in the
case where I understand nothing.

2. As regards the second claim, that the program explains human
understanding, we can see that the computer and its program do not
provide sufficient conditions of understanding since the computer and
the program are functioning, and there is no understanding. But does it
even provide a necessary condition or a significant contribution to un-
derstanding? One of the claims made by the supporters of strong AI is that
when I understand a story in English, what I am doing is exactly the same
—or perhaps more of the same—as what I was doing in manipulating the
Chinese symbols. It is simply more formal symbol manipulation that
distinguishes the case in English, where I do understand, from the case
in Chinese, where I don’t. I have not demonstrated that this claim is false,
but it would certainly appear an incredible claim in the example. Such
plausibility as the claim has derives from the supposition that we can
construct a program that will have the same inputs and outputs as native
speakers, and in addition we assume that speakers have some level of
description where they are also instantiations of a program. On the basis
of these two assumptions we assume that even if Schank’s program isn’t
the whole story about understanding, it may be part of the story. Well,
I suppose that is an empirical possibility, but not the slightest reason has
so far been given to believe that it is true, since what is suggested—
though certainly not demonstrated—by the example is that the computer
program is simply irrelevant to my understanding of the story. In the
Chinese case I have everything that artificial intelligence can put into me
by way of a program, and I understand nothing; in the English case I

understand everything, and there is so far no reason at all to suppose that
my understanding has anything to do with computer programs, that is,
with computational operations on purely formally specified elements. As
long as the program is defined in terms of computational operations on
purely formally defined elements, what the example suggests is that these
by themselves have no interesting connection with understanding. They
are certainly not sufficient conditions, and not the slightest reason has
been given to suppose that they are necessary conditions or even that
they make a significant contribution to understanding. Notice that the
force of the argument is not simply that different machines can have the
same input and output while operating on different formal principles—
that is not the point at all. Rather, whatever purely formal principles you
put into the computer, they will not be sufficient for understanding, since
a human will be able to follow the formal principles without understand-
ing anything. No reason whatever has been offered to suppose that such
principles are necessary or even contributory, since no reason has been
given to suppose that when I understand English I am operating with any
formal program at all.

Well, then, what is it that I have in the case of the English sentences
that I do not have in the case of the Chinese sentences? The obvious
answer is that I know what the former mean, while I haven’t the faintest
idea what the latter mean. But in what does this consist and why couldn’t
we give it to a machine, whatever it is? I will return to this question later,
but first I want to continue with the example.

I have had the occasions to present this example to several workers
in artificial intelligence, and, interestingly, they do not seem to agree on
what the proper reply to it is. I get a surprising variety of replies, and in
what follows I will consider the most common of these (specified along
with their geographic origins).

But first I want to block some common misunderstandings about
“understanding”: In many of these discussions one finds a lot of fancy
footwork about the word “understanding.” My critics point out that there
are many different degrees of understanding; that “understanding” is not
a simple two-place predicate; that there are even different kinds and
levels of understanding, and often the law of excluded middle doesn’t
even apply in a straightforward way to statements of the form “x under-
stands y”; that in many cases it is a matter for decision and not a simple
matter of fact whether x understands y; and so on. To all of these points
I want to say: of course, of course. But they have nothing to do with the
points at issue. There are clear cases in which “understanding” literally
applies and clear cases in which it does not apply; and these two sorts of
cases are all I need for this argument. I understand stories in English; to a lesser degree I can understand stories in French; to a still lesser degree, stories in German; and in Chinese, not at all. My car and my adding machine, on the other hand, understand nothing; they are not in that line of business. We often attribute "understanding" and other cognitive predicates by metaphor and analogy to cars, adding machines, and other artifacts, but nothing is proved by such attributions. We say, "The door knows when to open because of its photoelectric cell," "The adding machine knows how (understands how, is able) to do addition and subtraction but not division," and "The thermostat perceives changes in the temperature." The reason we make these attributions is quite interesting, and it has to do with the fact that in artifacts we extend our own intentionality; our tools are extensions of our purposes, and so we find it natural to make metaphorical attributions of intentionality to them; but I take it no philosophical ice is cut by such examples. The sense in which an automatic door "understands instructions" from its photoelectric cell is not at all the sense in which I understand English. If the sense in which Schank's programmed computers understand stories is supposed to be the metaphorical sense in which the door understands, and not the sense in which I understand English, the issue would not be worth discussing. But Newell and Simon (1963) write that the kind of cognition they claim for computers is exactly the same as for human beings. I like the straightforwardness of this claim, and it is the sort of claim I will be considering. I will argue that in the literal sense the programmed computer understands what the car and the adding machine understand, namely, exactly nothing. The computer understanding is not just (like my understanding of German) partial or incomplete; it is zero.

Now to the replies:

1. **The Systems Reply (Berkeley).** "While it is true that the individual person who is locked in the room does not understand the story, the fact is that he is merely part of a whole system, and the system does understand the story. The person has a large ledger in front of him in which are written the rules, he has a lot of scratch paper and pencils for doing calculations, he has 'data banks' of sets of Chinese symbols. Now, understanding is not being ascribed to the mere individual; rather it is being ascribed to this whole system of which he is a part."

---

*Also, "understanding" implies both the possession of mental (intentional) states and the truth (validity, success) of these states. For the purposes of this discussion we are concerned only with the possession of the states.

†Intentionality is by definition that feature of certain mental states by which they are directed at or about objects and states of affairs in the world. Thus, beliefs, desires, and intentions are intentional states; undirected forms of anxiety and depression are not.

---

**Minds, Brains, and Programs**

My response to the systems theory is quite simple: Let the individual internalize all of these elements of the system. He memorizes the rules in the ledger and the data banks of Chinese symbols, and he does all the calculations in his head. The individual then incorporates the entire system. There isn’t anything at all to the system that he does not encompass. We can even get rid of the room and suppose he works outdoors. All the same, he understands nothing of the Chinese, and a fortiori neither does the system, because there isn’t anything in the system that isn’t in him. If he doesn’t understand, then there is no way the system could understand because the system is just a part of him.

Actually I feel somewhat embarrassed to give even this answer to the systems theory because the theory seems to me so implausible to start with. The idea is that while a person doesn’t understand Chinese, somehow the conjunction of that person and bits of paper might understand Chinese. It is not easy for me to imagine how someone who was not in the grip of an ideology would find the idea at all plausible. Still, I think many people who are committed to the ideology of strong AI will in the end be inclined to say something very much like this; so let us pursue it a bit further. According to one version of this view, while the man in the internalized systems example doesn’t understand Chinese in the sense that a native Chinese speaker does (because, for example, he doesn’t know that the story refers to restaurants and hamburgers, etc.), still the man as a formal symbol manipulation system really does understand Chinese. The subsystem of the man that is the formal symbol manipulation system for Chinese should not be confused with the subsystem for English.

So there are really two subsystems in the man: one understands English, the other Chinese, and "it's just that the two systems have little to do with each other." But, I want to reply, not only do they have little to do with each other, they are not even remotely alike. The subsystem that understands English (assuming we allow ourselves to talk in this jargon of "subsystems" for a moment) knows that the stories are about restaurants and eating hamburgers, he knows that he is being asked questions about restaurants and that he is answering questions as best he can by making various inferences from the content of the story, and so on. But the Chinese system knows none of this. Whereas the English subsystem knows that "hamburgers" refers to hamburgers, the Chinese subsystem knows only that "squiggle squiggle" is followed by "squiggle squiggle." All he knows is that various formal symbols are being introduced at one end and manipulated according to rules written in English, and other symbols are going out at the other end. The whole point of the original example was to argue that such symbol manipulation by itself couldn’t be sufficient for understanding Chinese in any literal sense be...
cause the man could write "squiggle squiggle" after "squiggle squiggle" without understanding anything in Chinese. And it doesn't meet that argument to postulate subsystems within the man because the subsystems are no better off than the man was in the first place; they still don't have anything even remotely like what the English-speaking man (or subsystem) has. Indeed, in the case as described, the Chinese subsystem is simply a part of the English subsystem, a part that engages in meaningless symbol manipulation according to rules in English.

Let us ask ourselves what is supposed to motivate the systems reply in the first place; that is, what independent grounds are there supposed to be for saying that the agent must have a subsystem within him that literally understands stories in Chinese? As far as I can tell the only grounds are that in the example I have the same input and output as native Chinese speakers and a program that goes from one to the other. But the whole point of the examples has been to try to show that that couldn't be sufficient for understanding, in the sense in which I understand stories in English, because a person, and hence the set of systems that go to make up a person, could have the right combination of input, output, and program and still not understand anything in the relevant literal sense in which I understand English. The only motivation for saying there must be a subsystem in me that understands Chinese is that I have a program and I can pass the Turing test; I can fool native Chinese speakers. But precisely one of the points at issue is the adequacy of the Turing test. The example shows that there could be two "systems," both of which pass the Turing test, but only one of which understands; and it is no argument against this point to say that since they both pass the Turing test they must both understand, since this claim fails to meet the argument that the system in me that understands English has a great deal more than the system that merely processes Chinese. In short, the systems reply simply begs the question by insisting without argument that the system must understand Chinese.

Furthermore, the systems reply would appear to lead to consequences that are independently absurd. If we are to conclude that there must be cognition in me on the grounds that I have a certain sort of input and output and a program in between, then it looks like all sorts of noncognitive subsystems are going to turn out to be cognitive. For example, there is a level of description at which my stomach does information processing, and it instantiates any number of computer programs, but I take it we do not want to say that it has any understanding (cf. Pylyshyn 1980). But if we accept the systems reply, then it is hard to see how we avoid saying that stomach, heart, liver, and so on are all understanding subsystems, since there is no principled way to distinguish the motivation for saying the Chinese subsystem understands from saying that the stomach understands. It is, by the way, not an answer to this point to say that the Chinese system has information as input and output and the stomach has food and food products as input and output, since from the point of view of the agent, from my point of view, there is no information in either the food or the Chinese—the Chinese is just so many meaningless squiggles. The information in the Chinese case is solely in the eyes of the programmers and the interpreters, and there is nothing to prevent them from treating the input and output of my digestive organs as information if they so desire.

This last point bears on some independent problems in strong AI, and it is worth digressing for a moment to explain it. If strong AI is to be a branch of psychology, then it must be able to distinguish those systems that are genuinely mental from those that are not. It must be able to distinguish the principles on which the mind works from those on which nonmental systems work; otherwise it will offer us no explanations of what is specifically mental about the mental. And the mental-nonmental distinction cannot be just in the eye of the beholder but it must be intrinsic to the systems; otherwise it would be up to any beholder to treat people as nonmental and, for example, hurricanes as mental if he likes. But quite often in the AI literature the distinction is blurred in ways that would in the long run prove disastrous to the claim that AI is a cognitive inquiry. McCarthy, for example, writes, "Machines as simple as thermostats can be said to have beliefs, and having beliefs seems to be a characteristic of most machines capable of problem solving performance" (McCarthy 1979). Anyone who thinks strong AI has a chance as a theory of the mind ought to ponder the implications of that remark. We are asked to accept it as a discovery of strong AI that the hunk of metal on the wall that we use to regulate the temperature has beliefs in exactly the same sense that we, our spouses, and our children have beliefs, and furthermore that "most" of the other machines in the room—telephone, tape recorder, adding machine, electric light switch—also have beliefs in this literal sense. It is not the aim of this article to argue against McCarthy's point, so I will simply assert the following without argument. The study of the mind starts with such facts as that humans have beliefs, while thermostats, telephones, and adding machines don't. If you get a theory that denies this point you have produced a counterexample to the theory and the theory is false. One gets the impression that people in AI who write this sort of thing think they can get away with it because they don't really take it seriously, and they don't think anyone else will either. I propose, for a moment at least, to take it seriously. Think hard for one minute about what would be necessary to establish that that hunk of metal
on the wall over there had real beliefs, beliefs with direction of fit, propositional content, and conditions of satisfaction; beliefs that had the possibility of being strong beliefs or weak beliefs; nervous, anxious, or secure beliefs; dogmatic, rational, or superstitious beliefs; blind faiths or hesitant cogitations; any kind of beliefs. The thermostat is not a candidate. Neither is stomach, liver, adding machine, or telephone. However, since we are taking the idea seriously, notice that its truth would be fatal to strong AI's claim to be a science of the mind. For now the mind is everywhere. What we wanted to know is what distinguishes the mind from thermostats and livers. And if McCarthy were right, strong AI wouldn't have a hope of telling us that.

2. The Robot Reply (Yale). "Suppose we wrote a different kind of program from Schank's program. Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating, drinking—anything you like. The robot would, for example, have a television camera attached to it that enabled it to see, it would have arms and legs that enabled it to 'act,' and all of this would be controlled by its computer 'brain.' Such a robot would, unlike Schank's computer, have genuine understanding and other mental states."

The first thing to notice about the robot reply is that it tacitly conceives that cognition is not solely a matter of formal symbol manipulation, since this reply adds a set of causal relations with the outside world (cf. Fodor 1980). But the answer to the robot reply is that the addition of such "perceptual" and "motor" capacities adds nothing by way of understanding, in particular, or intentionality, in general, to Schank's original program. To see this, notice that the same thought experiment applies to the robot case. Suppose that instead of the computer inside the robot, you put me inside the room and, as in the original Chinese case, you give me more Chinese symbols with more instructions in English for matching Chinese symbols to Chinese symbols and feeding back Chinese symbols to the outside. Suppose, unknown to me, some of the Chinese symbols that come to me come from a television camera attached to the robot and other Chinese symbols that I am giving out serve to make the motors inside the robot move the robot's legs or arms. It is important to emphasize that all I am doing is manipulating formal symbols: I know none of these facts. I am receiving "information" from the robot's "perceptual" apparatus, and I am giving out "instructions" to its motor apparatus without knowing either of these facts. I am the robot's homunculus, but

Minds, Brains, and Programs

unlike the traditional homunculus, I don't know what's going on. I don't understand anything except the rules for symbol manipulation. Now in this case I want to say that the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program. And furthermore, by instantiating the program I have no intentional states of the relevant type. All I do is follow formal instructions about manipulating formal symbols.

3. The Brain Simulator Reply (Berkeley and M.I.T.). "Suppose we design a program that doesn't represent information that we have about the world, such as the information in Schank's scripts, but simulates the actual sequence of neuron firings at the synapses of the brain of a native Chinese speaker when he understands stories in Chinese and gives answers to them. The machine takes in Chinese stories and questions about them as input, it simulates the formal structure of actual Chinese brains in processing these stories, and it gives out Chinese answers as outputs. We can even imagine that the machine operates, not with a single serial program, but with a whole set of programs operating in parallel, in the manner that actual human brains presumably operate when they process natural language. Now surely in such a case we would have to say that the machine understood the stories; and if we refuse to say that, wouldn't we also have to deny that native Chinese speakers understood the stories? At the level of the synapses, what would or could be different about the program of the computer and the program of the Chinese brain?"

Before countering this reply I want to digress to note that it is an odd reply for any partisan of artificial intelligence (or functionalism, etc.) to make: I thought the whole idea of strong AI is that we don't need to know how the brain works to know how the mind works. The basic hypothesis, or so I had supposed, was that there is a level of mental operations consisting of computational processes over formal elements that constitute the essence of the mental and can be realized in all sorts of different brain processes, in the same way that any computer program can be realized in different computer hardware; On the assumptions of strong AI, the mind is to the brain as the program is to the hardware, and thus we can understand the mind without doing neurophysiology. If we had to know how the brain worked to do AI, we wouldn't bother with AI. However, even getting this close to the operation of the brain is still not sufficient to produce understanding. To see this, imagine that instead of a monolingual man in a room shuffling symbols we have the man operate an elaborate set of water pipes with valves connecting them. When the man receives the Chinese symbols, he looks up in the program, written in English, which valves he has to turn on and off. Each water connection..."
corresponds to a synapse in the Chinese brain, and the whole system is rigged up so that after doing all the right firings, that is after turning on all the right facets, the Chinese answers pop out at the output end of the series of pipes.

Now where is the understanding in this system? It takes Chinese as input, it simulates the formal structure of the synapses of the Chinese brain, and it gives Chinese as output. But the man certainly doesn’t understand Chinese, and neither do the water pipes, and if we are tempted to adopt what I think is the absurd view that somehow the conjunction of man and water pipes understands, remember that in principle the man can internalize the formal structure of the water pipes and do all the “neuron firings” in his imagination. The problem with the brain simulator is that it is simulating the wrong things about the brain. As long as it simulates only the formal structure of the sequence of neuron firings at the synapses, it won’t have simulated what matters about the brain, namely its causal properties, its ability to produce intentional states. And that the formal properties are not sufficient for the causal properties is shown by the water pipe example: we can have all the formal properties carved off from the relevant neurobiological causal properties.

4. The Combination Reply (Berkeley and Stanford). “While each of the previous three replies might not be completely convincing by itself as a refutation of the Chinese room counterexample, if you take all three together they are collectively much more convincing and even decisive. Imagine a robot with a brain-shaped computer lodged in its cranial cavity, imagine the computer programmed with all the synapses of a human brain, imagine the whole behavior of the robot is indistinguishable from human behavior, and now think of the whole thing as a unified system and not just as a computer with inputs and outputs. Surely in such a case we would have to ascribe intentionality to the system.”

I entirely agree that in such a case we would find it rational and indeed irresistible to accept the hypothesis that the robot had intentionality, as long as we knew nothing more about it. Indeed, besides appearance and behavior, the other elements of the combination are really irrelevant. If we could build a robot whose behavior was indistinguishable over a large range from human behavior, we would attribute intentionality to it, pending some reason not to. We wouldn’t need to know in advance that its computer brain was a formal analogue of the human brain.

But I really don’t see that this is any help to the claims of strong AI, and here’s why: According to strong AI, instantiating a formal program with the right input and output is a sufficient condition of, indeed is constitutive of, intentionality. As Newell (1979) puts it, the essence of the mental is the operation of a physical symbol system. But the attributions of intentionality that we make to the robot in this example have nothing to do with formal programs. They are simply based on the assumption that if the robot looks and behaves sufficiently like us, then we would suppose, until proven otherwise, that it must have mental states like ours that cause and are expressed by its behavior and it must have an inner mechanism capable of producing such mental states. If we knew independently how to account for its behavior without such assumptions we would not attribute intentionality to it, especially if we knew it had a formal program. And this is precisely the point of my earlier reply to objection II.

Suppose we knew that the robot’s behavior was entirely accounted for by the fact that a man inside it was receiving uninterpreted formal symbols from the robot’s sensory receptors and sending out uninterpreted formal symbols to its motor mechanisms, and the man was doing this symbol manipulation in accordance with a bunch of rules. Furthermore, suppose the man knows none of these facts about the robot, all he knows is which operations to perform on which meaningless symbols. In such a case we would regard the robot as an ingenious mechanical dummy. The hypothesis that the dummy has a mind would now be unwarranted and unnecessary, for there is now no longer any reason to ascribe intentionality to the robot or to the system of which it is a part (except of course for the man’s intentionality in manipulating the symbols). The formal symbol manipulations go on, the input and output are correctly matched, but the only real locus of intentionality is the man, and he doesn’t know any of the relevant intentional states; he doesn’t, for example, see what comes into the robot’s eyes, he doesn’t intend to move the robot’s arm, and he doesn’t understand any of the remarks made to or by the robot. Nor, for the reasons stated earlier, does the system of which man and robot are a part.

To see this point, contrast this case with cases in which we find it completely natural to ascribe intentionality to members of certain other primate species such as apes and monkeys and to domestic animals such as dogs. The reasons we find it natural are, roughly, two: We can’t make sense of the animal’s behavior without the ascription of intentionality, and we can see that the beasts are made of similar stuff to ourselves—that is an eye, that a nose, this is its skin, and so on. Given the coherence of the animal’s behavior and the assumption of the same causal stuff underlying it, we assume both that the animal must have mental states underlying its behavior, and that the mental states must be produced by mechanisms made out of the stuff that is like our stuff. We would certainly make similar assumptions about the robot unless we had some reason not to,
but as soon as we knew that the behavior was the result of a formal program, and that the actual causal properties of the physical substance were irrelevant we would abandon the assumption of intentionality.

There are two other responses to my example that come up frequently (and so are worth discussing) but really miss the point.

5. The Other Minds Reply (Yale). "How do you know that other people understand Chinese or anything else? Only by their behavior. Now the computer can pass the behavioral tests as well as they can (in principle), so if you are going to attribute cognition to other people you must in principle also attribute it to computers."

This objection really is only worth a short reply. The problem in this discussion is not about how I know that other people have cognitive states, but rather what it is that I am attributing to them when I attribute cognitive states to them. The thrust of the argument is that it couldn't be just computational processes and their output because the computational processes and their output can exist without the cognitive state. It is no answer to this argument to feign anesthesia. In "cognitive sciences" one presupposes the reality and knowability of the mental in the same way that in physical sciences one has to presuppose the reality and knowability of physical objects.

6. The Many Mansions Reply (Berkeley). "Your whole argument presupposes that AI is only about analog and digital computers. But that just happens to be the present state of technology. Whatever these causal processes are that you say are essential for intentionality (assuming you are right), eventually we will be able to build devices that have these causal processes, and that will be artificial intelligence. So your arguments are in no way directed at the ability of artificial intelligence to produce and explain cognition."

I really have no objection to this reply save to say that it in effect trivializes the project of strong AI by redefining it as whatever artificially produces and explains cognition. The interest of the original claim made on behalf of artificial intelligence is that it was a precise, well defined thesis: mental processes are computational processes over formally defined elements. I have been concerned to challenge that thesis. If the claim is redefined so that it is no longer that thesis, my objections no longer apply because there is no longer a testable hypothesis for them to apply to.

Let us now return to the question I promised I would try to answer: Granted that in my original example I understand the English and I do not understand the Chinese, and granted therefore that the machine doesn't understand either English or Chinese, still there must be something about me that makes it the case that I understand English and a corresponding something lacking in me that makes it the case that I fail to understand Chinese. Now why couldn't we give those somethings, whatever they are, to a machine?

I see no reason in principle why we couldn't give a machine the capacity to understand English or Chinese, since in an important sense our bodies with our brains are precisely such machines. But I do see very strong arguments for saying that we could not give such a thing to a machine where the operation of the machine is defined solely in terms of computational processes over formally defined elements; that is, where the operation of the machine is defined as an instantiation of a computer program. It is not because I am the instantiation of a computer program that I am able to understand English and have other forms of intentionality (I am, I suppose, the instantiation of any number of computer programs), but as far as we know it is because I am a certain sort of organism with a certain biological (i.e., chemical and physical) structure, and this structure, under certain conditions, is causally capable of producing perception, action, understanding, learning, and other intentional phenomena. And part of the point of the present argument is that only something that had those causal powers could have that intentionality. Perhaps other physical and chemical processes could produce exactly these effects; perhaps, for example, Martians also have intentionality but their brains are made of different stuff. That is an empirical question, rather like the question whether photosynthesis can be done by something with a chemistry different from that of chlorophyll.

But the main point of the present argument is that no purely formal model will ever be sufficient by itself for intentionality because the formal properties are not by themselves constitutive of intentionality, and they have by themselves no causal powers except the power, when instantiated, to produce the next stage of the formalism when the machine is running. And any other causal properties that particular realizations of the formal model have, are irrelevant to the formal model because we can always put the same formal model in a different realization where those causal properties are obviously absent. Even if, by some miracle, Chinese speakers exactly realize Schank's program, we can put the same program in English speakers, water pipes, or computers, none of which understand Chinese, the program notwithstanding.

What matters about brain operations is not the formal shadow cast by the sequence of synapses but rather the actual properties of the sequences. All the arguments for the strong version of artificial intelligence that I have seen insist on drawing an outline around the shadows cast...
by cognition and then claiming that the shadows are the real thing.

By way of concluding I want to try to state some of the general philosophical points implicit in the argument. For clarity I will try to do it in a question-and-answer fashion, and I begin with that old chestnut of a question:

“Could a machine think?”

The answer is, obviously, yes. We are precisely such machines.

“Yes, but could an artifact, a man-made machine, think?”

Assuming it is possible to produce artificially a machine with a nervous system, neurons with axons and dendrites, and all the rest of it, sufficiently like ours, again the answer to the question seems to be obviously, yes. If you can exactly duplicate the causes, you could duplicate the effects. And indeed it might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those that human beings use. It is, as I said, an empirical question.

“OK, but could a digital computer think?”

If by “digital computer” we mean anything at all that has a level of description where it can correctly be described as the instantiation of a computer program, then again the answer is, of course, yes, since we are the instantiations of any number of computer programs, and we can think.

“But could something think, understand, and so on solely in virtue of being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?”

This I think is the right question to ask, though it is usually confused with one or more of the earlier questions, and the answer to it is no.

“Why not?”

Because the formal symbol manipulations by themselves don’t have any intentionality: they are quite meaningless; they aren’t even symbol manipulations, since the symbols don’t symbolize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output.

The aim of the Chinese room example was to try to show this by showing that as soon as we put something into the system that really does have intentionality (a man), and we program him with the formal program, you can see that the formal program carries no additional intentionality. It adds nothing, for example, to a man’s ability to understand Chinese.

Minds, Brains, and Programs

Precisely that feature of AI that seemed so appealing—the distinction between the program and the realization—proves fatal to the claim that simulation could be duplication. The distinction between the program and its realization in the hardware seems to be parallel to the distinction between the level of mental operations and the level of brain operations. And if we could describe the level of mental operations as a formal program, then it seems we could describe what was essential about the mind without doing either introspective psychology or neuropsychology of the brain. But the equation “mind is to brain as program is to hardware” breaks down at several points, among them the following three:

First, the distinction between program and realization has the consequence that the same program could have all sorts of crazy realizations that had no form of intentionality. Weizenbaum (1976, Ch. 2), for example, shows in detail how to construct a computer using a roll of toilet paper and a pile of small stones. Similarly, the Chinese story understanding program can be programmed into a sequence of water pipes, a set of wind machines, or a monolingual English speaker, none of which thereby acquires an understanding of Chinese. Stones, toilet paper, wind, and water pipes are the wrong kind of stuff to have intentionality in the first place—only something that has the same causal powers as brains can have intentionality—and though the English speaker has the right kind of stuff for intentionality you can easily see that he doesn’t get any extra intentionality by memorizing the program, since memorizing it won’t teach him Chinese.

Second, the program is purely formal, but the intentional states are not in that way formal. They are defined in terms of their content, not their form. The belief that it is raining, for example, is not defined as a certain formal shape, but as a certain mental content with conditions of satisfaction, a direction of fit (see Searle 1979), and the like. Indeed the belief as such hasn’t even got a formal shape in this syntactic sense, since one and the same belief can be given an indefinite number of different syntactic expressions in different linguistic systems.

Third, as I mentioned before, mental states and events are literally a product of the operation of the brain, but the program is not in that way a product of the computer.

“Well if programs are in no way constitutive of mental processes, why have so many people believed the converse? That at least needs some explanation.”

I don’t really know the answer to that one. The idea that computer simulations could be the real thing ought to have seemed suspicious in the first place because the computer isn’t confined to simulating mental
operations, by any means. No one supposes that computer simulations of a five-alarm fire will burn the neighborhood down or that a computer simulation of a rainstorm will leave us all drenched. Why on earth would anyone suppose that a computer simulation of understanding actually understood anything? It is sometimes said that it would be frightfully hard to get computers to feel pain or fall in love, but love and pain are neither harder nor easier than cognition or anything else. For simulation, all you need is the right input and output and a program in the middle that transforms the former into the latter. That is all the computer has for anything it does. To confuse simulation with duplication is the same mistake, whether it is pain, love, cognition, fires, or rainstorms.

Still, there are several reasons why AI must have seemed—and to many people perhaps still does seem—in some way to reproduce and thereby explain mental phenomena, and I believe we will not succeed in removing these illusions until we have fully exposed the reasons that give rise to them.

First, and perhaps most important, is a confusion about the notion of “information processing”: many people in cognitive science believe that the human brain, with its mind, does something called “information processing,” and analogously the computer with its program does information processing; but fires and rainstorms, on the other hand, don’t do information processing at all. Thus, though the computer can simulate the formal features of any process whatever, it stands in a special relation to the mind and brain because when the computer is properly programmed, ideally with the same program as the brain, the information processing is identical in the two cases, and this information processing is really the essence of the mental. But the trouble with this argument is that it rests on an ambiguity in the notion of “information.” In the sense in which people “process information” when they reflect, say, on problems in arithmetic or when they read and answer questions about stories, the programmed computer does not do “information processing.” Rather, what it does is manipulate formal symbols. The fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. The computer, to repeat, has a syntax but no semantics. Thus, if you type into the computer “2 plus 2 equals?” it will type out “4.” But it has no idea that “4” means 4 or that it means anything at all. And the point is not that it lacks some second-order information about the interpretation of its first-order symbols, but rather that its first-order symbols don’t have any interpretations as far as the computer is concerned. All the computer has is more symbols. The introduction of the notion of

“information processing” therefore produces a dilemma: either we construct the notion of “information processing” in such a way that it implies intentionality as part of the process or we don’t. If the former, then the programmed computer does not do information processing, it only manipulates formal symbols. If the latter, then, though the computer does information processing, it is only doing so in the sense in which adding machines, typewriters, stomachs, thermostats, rainstorms, and hurricanes do information processing; namely, they have a level of description at which we can describe them as taking information in at one end, transforming it, and producing information as output. But in this case it is up to outside observers to interpret the input and output as information in the ordinary sense. And no similarity is established between the computer and the brain in terms of any similarity of information processing.

Second, in much of AI there is a residual behaviorism or operationalism. Since appropriately programmed computers can have input-output patterns similar to those of human beings, we are tempted to postulate mental states in the computer similar to human mental states. But once we see that it is both conceptually and empirically possible for a system to have human capacities in some realm without having any intentionality at all, we should be able to overcome this impulse. My desk adding machine has calculating capacities, but no intentionality, and in this paper I have tried to show that a system could have input and output capabilities that duplicated those of a native Chinese speaker and still not understand Chinese, regardless of how it was programmed. The Turing test is typical of the tradition in being unashamedly behavioristic and operationalistic, and I believe that if AI workers totally repudiated behaviorism and operationalism much of the confusion between simulation and duplication would be eliminated.

Third, this residual operationalism is joined to a residual form of dualism. Indeed strong AI only makes sense given the dualistic assumption that, where the mind is concerned, the brain doesn’t matter. In strong AI (and in functionalism, as well) what matters are programs, and programs are independent of their realization in machines; indeed, as far as AI is concerned, the same program could be realized by an electronic machine, a Cartesian mental substance, or a Hegelian world spirit. The single most surprising discovery that I have made in discussing these issues is that many AI workers are quite shocked by my idea that actual human mental phenomena might be dependent on actual physical-chemical properties of actual human brains. But if you think about it a minute you can see that I should not have been surprised; for unless you accept some form of dualism the strong AI project hasn’t got a chance. The
project is to reproduce and explain the mental by designing programs, but unless the mind is not only conceptually but empirically independent of the brain you couldn’t carry out the project, for the program is completely independent of any realization. Unless you believe that the mind is separable from the brain both conceptually and empirically—dualism in a strong form—you cannot hope to reproduce the mental by writing and running programs since programs must be independent of brains or any other particular forms of instantiation. If mental operations consist in computational operations on formal symbols, then it follows that they have no interesting connection with the brain; the only connection would be that the brain just happens to be one of the indefinitely many types of machines capable of instantiating the program. This form of dualism is not the traditional Cartesian variety that claims there are two sorts of substances, but it is Cartesian in the sense that it insists that what is specifically mental about the mind has no intrinsic connection with the actual properties of the brain. This underlying dualism is masked from us by the fact that AI literature contains frequent fulminations against “dualism”; what the authors seem to be unaware of is that their position presupposes a strong version of dualism.

“Could a machine think?” My own view is that only a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as brains. And that is the main reason strong AI has had little to tell us about thinking, since it has nothing to tell us about machines. By its own definition, it is about programs, and programs are not machines. Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena. No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle because of a deep and abiding dualism: the mind they suppose is a matter of formal processes and is independent of quite specific material causes in the way that milk and sugar are not.

In defense of this dualism the hope is often expressed that the brain is a digital computer (early computers, by the way, were often called “electronic brains”). But that is no help. Of course the brain is a digital computer. Since everything is a digital computer, brains are too. The point is that the brain’s causal capacity to produce intentionality cannot consist in its instantiating a computer program, since for any program you like it is possible for something to instantiate that program and still not have any mental states. Whatever it is that the brain does to produce intentionality, it cannot consist in instantiating a program since no program, by itself, is sufficient for intentionality.*

Reflections

This article originally appeared together with twenty-eight responses from assorted people. Many of the responses contained excellent commentary, but reprinting them would have overloaded this book, and in any case some were a little too technical. One of the nice things about Searle’s article is that it is pretty much understandable by someone without special training in AI, neurology, philosophy, or other disciplines that have a bearing on it.

Our position is quite opposed to Searle’s, but we find in Searle an eloquent opponent. Rather than attempt to give a thorough rebuttal to his points, we will concentrate on a few of the issues he raises, leaving our answers to his other points implicit, in the rest of this book.

Searle’s paper is based on his ingenious “Chinese room thought experiment,” in which the reader is urged to identify with a human being executing by hand the sequence of steps that a very clever AI program would allegedly go through as it read stories in Chinese and answered questions about them in Chinese in a manner sufficiently human-seeming as to be able to pass the Turing test. We think Searle has committed a serious and fundamental misrepresentation by giving the impression that it makes any sense to think that a human being could do this. By buying this image, the reader is unwittingly sucked into an impossibly unrealistic concept of the relation between intelligence and symbol manipulation.

The illusion that Searle hopes to induce in readers (naturally he doesn’t think of it as an illusion!) depends on his managing to make readers overlook a tremendous difference in complexity between two systems at different conceptual levels. Once he has done that, the rest is a piece of cake. At the outset, the reader is invited to identify with Searle

---

*I am indebted to a rather large number of people for discussion of these matters and for their patient attempts to overcome my ignorance of artificial intelligence. I would especially like to thank Ned Block, Hubert Dreyfus, John Haugeland, Roger Schank, Robert Wilensky, and Terry Winograd.
as he hand-simulates an existing AI program that can, in a limited way, answer questions of a limited sort, in a few limited domains. Now, for a person to hand-simulate this, or any currently existing AI program—that is, to step through it at the level of detail that the computer does—would involve days, if not weeks or months, of arduous, horrendous boredom. But instead of pointing this out, Searle—as deft at distracting the reader’s attention as a practiced magician—switches the reader’s image to a hypothetical program that passes the Turing test! He has jumped up many levels of competency without so much as a passing mention. The reader is again invited to put himself or herself in the shoes of the person carrying out the step-by-step simulation, and to “feel the lack of understanding” of Chinese. This is the crux of Searle’s argument.

Our response to this (and, as we shall show later, Searle’s response as well, in a way) is basically the “Systems Reply” that it is a mistake to try to impute the understanding to the (incidentally) animate simulator; rather it belongs to the system as a whole, which includes what Searle casually characterizes as “a few slips of paper.” This offhand comment, we feel, reveals how Searle’s image has blinded him to the realities of the situation. A thinking computer is as repugnant to John Searle as non-Euclidean geometry was to its unwitting discoverer, Gerolamo Saccheri, who thoroughly disowned his own creation. The time—the late 1700s—was not quite ripe for people to accept the conceptual expansion caused by alternate geometries. About fifty years later, however, non-Euclidean geometry was rediscovered and slowly accepted.

Perhaps the same will happen with “artificial intentionality”—if it is ever created. If there ever came to be a program that could pass the Turing test, it seems that Searle, instead of marveling at the power and depth of that program, would just keep on insisting that it lacked some marvelous “causal powers of the brain” (whatever they are). To point out the vacuity of that notion, Zenon Pylyshyn, in his reply to Searle, wondered if the following passage, quite reminiscent of Zuboff’s “Story of a Brain” (selection 12), would accurately characterize Searle’s viewpoint:

If more and more of the cells in your brain were to be replaced by integrated circuit chips, programmed in such a way as to keep the input-output function of each unit identical to that of the unit being replaced, you would in all likelihood just keep right on speaking exactly as you are doing now except that you would eventually stop meaning anything by it. What we outside observers might take to be words would become for you just certain noises that circuits caused you to make.

The weakness of Searle’s position is that he offers no clear way to tell when genuine meaning—or indeed the genuine “you”—has vanished from this system. He merely insists that some systems have intentionality by virtue of their “causal powers” and that some don’t. He vacillates about what those powers are due to. Sometimes it seems that the brain is composed of “the right stuff,” but other times it seems to be something else. It is whatever seems convenient at the moment—now it is the slippery essence that distinguishes “form” from “content,” now another essence that separates syntax from semantics, and so on.

To the Systems-Reply advocates, Searle offers the thought that the human being in the room (whom we shall from now on refer to as “Searle’s demon”) should simply memorize, or incorporate, all the material on the “few slips of paper.” As if a human being could, by any conceivable stretch of the imagination, do this. The program on those “few slips of paper” embodies the entire mind and character of something as complex in its ability to respond to written material as a human being is, by virtue of being able to pass the Turing test. Could any human being simply “swallow up” the entire description of another human being’s mind? We find it hard enough to memorize a written paragraph; but Searle envisions the demon as having absorbed what in all likelihood would amount to millions, if not billions, of pages densely covered with abstract symbols—and moreover having all of this information available, whenever needed, with no retrieval problems. This unlikely aspect of the scenario is all lightly described, and it is not part of Searle’s key argument to convince the reader that it makes sense. In fact, quite the contrary—a key part of his argument is in glossing over these questions of orders of magnitude, for otherwise a skeptical reader will realize that nearly all of the understanding must lie in the billions of symbols on paper, and practically none of it in the demon. The fact that the demon is animate is an irrelevant—indeed, misleading—side issue that Searle has mistaken for a very significant fact.

We can back up this argument by exhibiting Searle’s own espousal of the Systems Reply. To do so, we should first like to place Searle’s thought experiment in a broader context. In particular, we would like to show how Searle’s setup is just one of a large family of related thought experiments, several of which are the topics of other selections in this book. Each member of this family of thought experiments is defined by a particular choice of “knob settings” on a thought-experiment generator. Its purpose is to create—in your mind’s eye—various sorts of imaginary simulations of human mental activity. Each different thought experiment is an “intuition pump” (Dennett’s term) that magnifies one facet or other of the issue, tending to push the reader toward certain conclusions. We see approximately five knobs of interest, although it is possible that someone else could come up with more.
Knob 1. This knob controls the physical “stuff” out of which the simulation will be constructed. Its settings include: neurons and chemicals; water pipes and water; bits of paper and symbols on them; toilet paper and stones; data structures and procedures; and so on.

Knob 2. This knob controls the level of accuracy with which the simulation attempts to mimic the human brain. It can be set at an arbitrarily fine level of detail (particles inside atoms), at a coarser level such as that of cells and synapses, or even at the level that AI researchers and cognitive psychologists deal with: that of concepts and ideas, representations and processes.

Knob 3. This knob controls the physical size of the simulation. Our assumption is that microminiaturization would allow us to make a teeny-weeny network of water pipes or solid-state chips that would fit inside a thimble, and conversely that any chemical process could be blown up to the macroscopic scale.

Knob 4. This critical knob controls the size and nature of the demon who carries out the simulation. If it is a normal-sized human being, we shall call it a “Searle’s demon.” If it is a tiny elflike creature that can sit inside neurons or on particles, then we shall call it a “Haugeeland’s demon,” after John Haugeland, whose response to Searle featured this notion. The settings of this knob also determine whether the demon is animate or inanimate.

Knob 5. This knob controls the speed at which the demon works. It can be set to make the demon work blindingly fast (millions of operations per microsecond) or agonizingly slowly (maybe one operation every few seconds).

Now, by playing with various knob settings, we can come up with various thought experiments. One choice yields the situation described in section 26, “A Conversation with Einstein’s Brain.” Another choice yields Searle’s Chinese room experiment. In particular, that involves the following knob settings:

Knob 1: neurons and chemicals
Knob 2: neural-firing level
Knob 3: brain size
Knob 4: eensy-weensy demon
Knob 5: dazzlingly fast demon

What Haugeland wants us to envision is this: A real woman’s brain is, unfortunately, defective. It no longer is able to send neurotransmitters from one neuron to another. Luckily, however, this brain is inhabited by an incredibly tiny and incredibly speedy Haugeland’s demon, who intervenes every single time any neuron would have been about to release neurotransmitters into a neighboring neuron. This demon “tickles” the appropriate synapse of the next neuron in a way that is functionally indistinguishable, to that neuron, from the arrival of genuine neurotransmitters. And the H-demon is so swift that he can jump around from synapse to synapse in trillions of a second, never falling behind schedule. In this way the operation of the woman’s brain proceeds exactly as it would have, if she were healthy. Now, Haugeland asks Searle, does the woman still think—that is, does she possess intentionality—or, to recall the words of Professor Jefferson as cited by Turing, does she merely “artificially signal”? From which to look at the experiment. Let us add a little color to this drab experiment and say that the simulated Chinese speaker involved is a woman and that the demons (if animate) are always male. Now we have a choice between the demon’s-eye view and the system’s-eye view. Remember that by hypothesis, both the demon and the simulated woman are equally capable of articulating their views on whether or not they are understanding, and on what they are experiencing. Searle is insistent, nonetheless, that we see this experiment only from the point of view of the demon. He insists that no matter what the simulated woman claims (in Chinese, of course) about her understanding, we should disregard her claims, and pay attention to the demon inside, who is carrying out the symbol manipulation. Searle’s claim amounts to the notion that actually there is only one point of view, not two. If one accepts the way Searle describes the whole experiment, this claim has great intuitive appeal, since the demon is about our size, speaks our language, and works at about our speed—and it is very hard to identify with a “woman” whose answers come at the rate of one per century (with luck)—and in “meaningless squiggles and squiggles,” to boot.

But if we change some of the knob settings, we can also alter the ease with which we change point of view. In particular, Haugeland’s variation involves switching various knobs as follows:

Knob 1: paper and symbols
Knob 2: concepts and ideas
Knob 3: room size
Knob 4: human-sized demon
Knob 5: slow setting (one operation every few seconds)
You might expect Searle to urge us to listen to and identify with the demon, and to eschew the Systems Reply, which would be, of course, to listen to and identify with the woman. But in his response to Hauge land, Searle surprises us—he chooses to listen to her this time and to ignore the demon who is cursing us from his tiny vantage point, yelling up to us, “Fools! Don’t listen to her! She’s merely a puppet whose every action is caused by my tickling, and by the program embedded in these many neurons that I zip around among.” But Searle does not heed the H-demon’s warning cries. He says, “Her neurons still have the right causal powers; they just need some help from the demon.”

We can construct a mapping between Searle’s original setup and this modified setup. To the “few slips of paper” now correspond all the synapses in the woman’s brain. To the AI program written on these “few slips of paper” corresponds the entire configuration of the woman’s brain; this amounts to a gigantic prescription telling the demon when and how to know which synapses to tickle. To the act of writing “meaningless squiggles and squoggles of Chinese” on paper corresponds the act of tickling her synapses. Suppose we take the setup as is, except that we’ll vary the size and speed knobs. We’ll blow the woman’s brain up to the size of the Earth, so that the demon becomes an “us-sized” S-demon instead of a tiny H-demon. And let’s also have the S-demon act at speeds reasonable for humans, instead of zipping thousands of miles throughout this bulbous brain in mere microseconds. Now which level does Searle wish us to identify with? We won’t speculate, but it seems to us that if the Systems Reply was compelling in the previous case, it should still be so in this case.

It must be admitted that Searle’s thought experiment vividly raises the question of what understanding a language really is. We would like to digress for a moment on that topic. Consider the question: “What kind of ability to manipulate the written or spoken symbols of a language amounts to a true understanding of that language?” Parrots who parrot English do not understand English. The recorded voice of a woman announcing the exact time of day on the telephone time service is not the mouthpiece of a system that understands English. There is no mentality behind that voice—it has been skimmed off of its mental substrate, yet retains a human-seeming quality. Perhaps a child would wonder how anyone could have so boring a job, and could do it so reliably. This would amuse us. It would be another matter, of course, if her voice were being driven by a flexible AI program that could pass the Turing test!

Imagine you are teaching a class in China. Further, imagine that you are aware of formulating all your thoughts in English and then of applying last-minute transformation rules (in reality, they would be last-split-second rules) that convert the English thoughts into instructions for moving your mouth and vocal cords in strange, “meaningless” ways—and yet, all your pupils sit there and seem quite satisfied with your performance. When they raise their hands, they utter exotic sounds that, although they are completely meaningless to you, are equipped to deal with, as you quickly apply some inverse rules and recover the English meanings underlying them. . . . Would you feel you were actually speaking Chinese? Would you feel you had gained some insight into the Chinese mentality? Or—can you actually imagine this situation? Is it realistic? Could anyone actually speak a foreign language well using this method?

The standard line is “You must learn to think in Chinese.” But in what does this consist? Anyone who has experienced it will recognize this description: The sounds of the second language pretty soon become “unheard”—you hear right through them, rather than hearing them, as you see right through a window, rather than seeing the window. Of course, you can make yourself hear a familiar language as pure uninterpreted sound if you try very hard, just as you can look at a windowpane if you want; but you can’t have your cake and eat it too—you can’t hear the sounds both with and without their meanings. And so most of the time people hear mainly meaning. For those people who learn a language because of enchantment with its sounds, this is a bit disappointing—and yet mastery of those sounds, even if one no longer hears them naively, is a beautiful, exhilarating experience. (It would be an interesting thing to try to apply this same kind of analysis to the hearing of music, where the distinction between hearing bare sounds and hearing their “meanings” is far less well understood, yet seems very real.)

Learning a second language involves transcending one’s own native language. It involves mixing the new language right in with the medium in which thought takes place. Thoughts must be able to germinate as easily (or nearly as easily) in the new language as in one’s native language. The way in which a new language’s habits seep down level by level and finally get absorbed into neurons is a giant mystery still. But one thing for certain is that mastery of a language does not consist in getting your “English subsystem” to execute for you a program of rules that enable you to deal with a language as a set of meaningless sounds and marks. Somehow, the new language must fuse with your internal representational system—your repertoire of concepts, images, and so on—in the same intimate way as English is fused with it. To think precisely about this, one must develop a very clear notion of the concept of levels of implementation, a computer-science concept of great power.

Computer scientists are used to the idea that one system can “emu-
late" another system. In fact, it follows from a theorem proven in 1936 by Alan Turing that any general-purpose digital computer can take on the guise of any other general-purpose digital computer, and the only difference to the outside world will be one of speed. The verb "emulate" is reserved for simulations, by a computer, of another computer, while "simulate" refers to the modeling of other phenomena, such as hurricanes, population curves, national elections, or even computer users.

A major difference is that simulation is almost always approximate, depending on the nature of the model of the phenomenon in question, whereas emulation is in a deep sense exact. So exact is it that when, say, a Sigma-5 computer emulates a computer with different architecture—say a DEC PDP-10—the users of the machine will be unaware that they are not dealing with a genuine DEC. This embedding of one architecture in another gives rise to so-called "virtual machines"—in this case, a virtual DEC-10. Underneath every virtual machine there is always some other machine. It may be a machine of the same type, it may even be another virtual machine. In his book Structured Computer Organization, Andrew Tanenbaum uses this notion of virtual machines to explain how large computer systems can be seen as a stack of virtual machines implemented one on top of the other—the bottommost one being, of course, a real machine! But in any case, the levels are sealed off from each other in a watertight way, just as Searle's demon was prevented from talking to the Chinese speaker he was part of. (It is intriguing to imagine what kind of conversation would take place—assuming that there were an interpreter present, since Searle's demon knows no Chinese!)

Now in theory, it is possible to have any two such levels communicate with each other, but this has traditionally been considered bad style; level-mixing is forbidden. Nonetheless, it is probable that this forbidden fruit—this blurring of two implementational levels—is exactly what goes on when a human "system" learns a second language. The second language does not run on top of the first one as a kind of software parasite, but rather becomes equally fundamentally implanted in the hardware (or nearly so). Somehow, absorption of a second language involves bringing about deep changes in one's underlying "machine"—a vast and coherent set of changes in the ways that neurons fire, so sweeping a set of changes that it creates new ways for the higher-level entities—the symbols—to trigger one another.

To parallel this in a computer system, a higher-level program would have to have some way of creating changes inside the "demon" that is carrying its program out. This is utterly foreign to the present style in computer science of implementing one level above another in a strictly vertical, sealed-off fashion. The ability of a higher level to loop back and affect lower levels—its own underpinnings—is a kind of magic trick which we feel is very close to the core of consciousness. It will perhaps one day prove to be a key element in the push toward ever-greater flexibility in computer design, and of course in the approach toward artificial intelligence. In particular, a satisfactory answer to the question of what "understanding" really means will undoubtedly require a much sharper delineation of the ways in which different levels in a symbol-manipulating system can depend on and affect one another. All in all, these concepts have proven elusive, and a clear understanding of them is probably a good ways off yet.

In this rather confusing discussion of many levels, you may have started to wonder what in the world "level" really means. It is a most difficult question. As long as levels are sealed off from each other, like Searle's demon and the Chinese-speaking woman, it is fairly clear. When they begin to blur, beware! Searle may admit that there are two levels in his thought experiment, but he is reluctant to admit that there are two occupied points of view—two genuine beings that feel and "have experience." He is worried that once we admit that some computational systems might have experiences, that would be a Pandora's box and all of a sudden "mind would be everywhere"—in the churning of stomachs, livers, automobile engines, and so on.

Searle seems to believe that any system whatsoever can be ascribed beliefs and feelings and so on, if one looks hard enough for a way to describe the system as an instantiation of an AI program. Obviously, that would be a disturbing notion, leading the way to panpsychism. Indeed, Searle believes that the AI people have unwittingly committed themselves to a panpsychic vision of the world.

Searle's escape from his self-made trap is to maintain that all those "beliefs" and "feelings" that you will uncover in inanimate objects and so forth when you begin seeing mind everywhere are not genuine but "pseudo." They lack intentionality! They lack the causal powers of the brain! (Of course, Searle would caution others to beware of confusing these notions with the naïvely dualistic notion of "soul.")

Our escape is to deny that the trap exists at all. It is incorrect to see minds everywhere. We say: minds do not lurk in car engines or livers any more than brains lurk in car engines and livers.

It is worthwhile expanding on this a little. If you can see all the complexity of thought processes in a churning stomach, then what's to prevent you from reading the pattern of bubbles in a carbonated beverage as coding for the Chopin piano concerto in E minor? And don't the holes in pieces of Swiss cheese code for the entire history of the United States? Sure they do—in Chinese as well as in English. After all, all things
are written everywhere! Bach’s Brandenburg concerto no. 2 is coded for in the structure of Hamlet—and Hamlet was of course readable (if you’d only known the code) from the structure of the last piece of birthday cake you gobbled down.

The problem is, in all these cases, that of specifying the code without knowing in advance what you want to read. For otherwise, you could pull a description of anyone’s mental activity out of a baseball game or a blade of grass by an arbitrarily constructed a posteriori code. But this is not science.

Minds come in different grades of sophistication, surely, but minds worth calling minds exist only where sophisticated representational systems exist, and no describable mapping that remains constant in time will reveal a self-updating representational system in a car engine or a liver. Perhaps one could read mentality into a rumbling car engine in somewhat the way that people read extra meanings into the structures of the Great Pyramids or Stonehenge, the music of Bach, Shakespeare’s plays, and so on—namely, by fabricating far-fetched numerological mapping schemes that can be molded and flexed whenever needed to fit the desires of the interpreter. But we doubt that that is what Searle intends (we do grant that he intends).

Minds exist in brains and may come to exist in programmed machines. If and when such machines come about, their causal powers will derive not from the substances they are made of, but from their design and the programs that run in them. And the way we will know they have those causal powers is by talking to them and listening carefully to what they have to say.

D.R.H.

Once upon a time there was a dualist. He believed that mind and matter are separate substances, just how they interacted he did not pretend to know—this was one of the “mysteries” of life. But he was sure they were quite separate substances.

This dualist, unfortunately, led an unbearably painful life—not because of his philosophical beliefs, but for quite different reasons. And he had excellent empirical evidence that no respite was in sight for the rest of his life. He longed for nothing more than to die. But he was deterred from suicide by such reasons as: (1) he did not want to hurt other people by his death; (2) he was afraid suicide might be morally wrong; (3) he was afraid there might be an afterlife, and he did not want to risk the possibility of eternal punishment. So our poor dualist was quite desperate.

Then came the discovery of the miracle drug! Its effect on the taker was to annihilate the soul or mind entirely but to leave the body functioning exactly as before. Absolutely no observable change came over the taker; the body continued to act just as if it still had a soul. Not the closest friend or observer could possibly know that the taker had taken the drug, unless the taker informed him.

Do you believe that such a drug is impossible in principle? Assuming you believe it possible, would you take it? Would you regard it as immoral? Is it tantamount to suicide? Is there anything in Scriptures forbid-
Turing Test Considered Harmful

Patrick Hayes
Beckman Institute
University of Illinois
Urbana, Illinois 61801
U. S. A.

Kenneth Ford
Institute for Human & Machine Cognition
University of West Florida
Pensacola, Florida 32514
U. S. A.

Abstract
Passing the Turing Test is not a sensible goal for Artificial Intelligence. Adherence to Turing’s vision from 1950 is now actively harmful to our field. We review problems with Turing’s idea, and suggest that, ironically, the very cognitive science that he tried to create must reject his research goal.

1 Introduction
Alan Turing was one of the greatest scientists of this century. His paper [1950], “Computing Machinery and Intelligence” inspired the creation of our field, giving it a vision, a philosophical charter and its first great challenge, the Turing Test.

The Turing Test has been with AI since its inception, and has always partly defined the field. Some AI pioneers seriously adopted it as a long-range goal, and some long-standing research programs are still guided by it; for others it has come to provide more of a vision to define and motivate the whole field. For example, in his recent text, Ginsberg [1993] defines AI as “the enterprise of constructing a physical symbol system that can reliably pass the Turing Test.”

Passing the Turing Test is now often understood to mean something like “making an artificial intelligence,” without paying too much attention to the details. In this article, however, we will take Turing seriously. We do not think he was being merely metaphorical or speaking in some loose, inspirational way. He seems to have been suggesting the imitation game as a definite goal for a program of research. It was supposed to be a concrete and relatively well-defined goal and hence to avoid the philosophical quagmire that Turing (correctly) predicted would result from debates about whether a computer could properly be described as “intelligent.”

But taken this seriously, we will argue, it is no longer a useful idea. The Turing Test had a historical role in getting AI started, but it is now a burden to the field, damaging its public reputation and its own intellectual coherence. We must explicitly reject the Turing Test in order to find a more mature description of our goals; it is time to move it from the textbooks to the history books.

2 Head Games
As the reader probably knows, the Turing Test comprises an imitation game which involves a man, a woman, and a judge, all communicating (but unable to see one another) in a three-way conversation. (The sex of the judge is not specified, and we will use “he” for reasons purely of grammatical simplicity.) The immediate task of the judge is to decide which of the other two is the woman, and the task of each of the players is to persuade the judge that he or she is the woman and that the other is the man. Thus, the game is a test of the ability of a man to pretend to be a woman, and of a woman to resist being judged a man. To make the game more exact, Turing proposes to use an average score over many conversations and to limit the length of each conversation to, say, 10 minutes. Turing then simply says that we should try to make a machine which could successfully “take the place of a man” in this game 70% of the time.

Turing is usually understood to mean that the game should be played with the question of gender (e.g., being female) replaced by the question of species (e.g., being human), so that the judge is faced with the task of differentiating a human participant from a machine pretending to be human. We will call this the species test.

However, Turing does not mention any change to the rules of the imitation game, and there is no need to interpret him as meaning to do so. If we take him at his word, the test is rather clever. It has a woman and a machine each trying to convince the judge that they are a woman, and the judge’s task is still to decide which is the woman and which, therefore, is not. But this judge is not thinking about the differences between women and machines, but between women and men. The hypothesis that one of his subjects is not human is not even in his natural space of initial possibilities. This judge has exactly the same problem to solve as a judge in the original imitation game and could be expected to bring the same attitudes and skills to the problem. We will call this the gender test.2

There are some standard objections to Turing Tests in either version. For example, they seem likely to be extremely difficult: so difficult, indeed, that someone who declared one as his immediate research goal would now probably not be taken seriously. They ignore or sidestep many aspects of

1 Ruling out vision avoids such complexities as skill in cross-dressing, but in any case Turing thought that machine vision was likely to be very difficult and was irrelevant to the goal of his project.

current AI research, such as vision and robotics, and seem too closely bound up with natural language understanding to now be a beacon for the entire field. These are familiar objections, but there are deeper ones. The imitation game itself has some basic design flaws, which are inherited by any version of the Turing Test. Later, we will argue that AI should not be defined as an imitation of human abilities in any case.

3 On Not Detecting Anything

One of the first lessons learned by a graduate student in psychology is never to design an experiment to detect nothing. This is such a fundamental error that it has been given a title: confirming the null hypothesis. It is impossible either to completely define the experimental conditions (how hard should one look for the thing that might not be there?) or to come to a firm conclusion (what if one had looked harder, or differently?). The imitation game is precisely such a design, in which a difference between two behaviors is what isn’t being detected. Assume for a moment that one accepts the Turing Test as valid: if an artificial intelligence could reliably pass a given instantation of the test, it would have demonstrated either that its intelligence was genuine or that the judge was not clever enough to ask sufficiently telling questions. But this raises the problem of what exactly are the telling questions? Ironically then, the issue that the Turing Test was supposed to avoid remains in force: would it be an adequate criterion for intelligence?

The imitation game conditions say nothing about the judge, but the success of the game depends crucially on how clever, knowledgeable, and insightful the judge is. A clever judge will be looking out for subtle signs of femininity. For example, sociolinguistic studies by Robin Lakoff have shown that adult American women tend to use a wider range of color words than men do. A woman will typically distinguish crimson and scarlet, where a man will usually describe them simply as red, a word which many adult women regard as ambiguous. A good imitation-game judge would know this and be alert for this sign of womanhood; and therefore, a successful player must also be expected to know it, and use it. And of course this applies to any other detectable sexual differences in word usage. But how many such differences are there? The question will always be a matter for research. The imitation game does not have a stable endpoint.

The zero-sum competitive design of Turing’s game has more odd consequences. It would not be enough simply to exhibit female use of color words, for example. If a female player notices her opponent (whom she knows to be male) using words like “puce” or “magenta,” she might challenge him to engage in an explicit debate about color to test his knowledge and explicitly draw the judge’s attention to such attempts to mislead, and a male player would need to be able to deal with this. To be a successful player it would not be sufficient simply to have, and therefore exhibit the linguistic symptoms of, a feminine attitude to color; one would have to consciously know of those symptoms and use this knowledge in tactical planning. To be successful at the imitation game, one would have to be thinking all the time about techniques of female impersonation.

Note that such conscious strategic use of sociolinguistics is quite different from exhibiting a symptom of some underlying cognitive difference. Some writers have objected to such things as an implemented model of female use of color vocabulary on the grounds that any such model involving “knowledge representation” can be only understood as modeling conscious thought. The difference between someone who, quite unwillingly, uses a rich color vocabulary and someone who consciously uses knowledge of sociolinguistics to improve his or her imitation game performance provides a vivid illustration of the necessary distinction.

Another problem with null-effect experiments is that they cannot measure anything. The imitation game can test only for complete success. A man who failed to seem feminine, in say, 10% of what he said would almost always fail the imitation game: to pass, one has to be totally convincing almost all the time. This is a criticism of these tests not only as a guide to research—they provide no way to measure partial progress toward the goal—but also of the goal itself. Even in humans we recognize the possibility that an intellectual talent need not correlate with conversational skill or debating ability; but using any kind of imitation game as our research goal denies this simple insight and declares that we must strive to create a fully human-like collection of abilities, organized to succeed in winning an argument.

4 Turing Test Problems

All of these criticisms apply directly to the Turing Tests. This is what we would have to make our program able to do: not talk like a human because it thinks like a human, or even talk like a woman because it thinks like a woman, but rather to talk like a woman as a result of thinking about how best to talk like a woman. The gender test is not a test of making an artificial human, but of making a mechanical transvestite.

Our point here, to emphasize, is not a moral one; rather it is concerned with what the program would have to be thinking about in order to be successful in these artificial games. Human players would also be forced into these artificial frames of mind, which arise simply from the tactical pressures of the games themselves. For example, to succeed at the species test, a machine must not just pass as human, it must succeed in persuading the judge that its human opponent is a machine. To do this would require more than ordinary conversational abilities. The competitive nature of this test makes it essential that the machine give a human-like impression in every possible way and be alert for any way in which its opponent might seem mechanical. To pass this test, a machine would have to not just give a human-like impression, but also be an expert on making a good impression, be always aware of the impression it was giving, and be ready to defend itself against accusations of giving the wrong impression. It would have to take care not to exhibit any inhuman talents which it might have; it would have to always cleverly lie, cheat, and disguise. The winner of the Loebner competition, for example, sometimes deliberately “mystyped” a word, then backspaced to correct it at human typing speed. This strategy is clever, but surely such tricks should not be central to our subject. To pass the species test we must make not an artificial intelligence, but an artificial con artist.

HAYES AND FORD 973
The species test further reveals the poor experimental design of the imitation game in the difficulty of obtaining an unbiased judge. The general perception of what are essentially human talents keeps shifting. As AI progresses and more and more tasks previously considered to involve human abilities are performed by machines, a judge in the naive Turing Test will gain more and more subtle ways of detecting the behavior of nonhuman machines, just as a skilled doctor will become more adept at recognizing subtle symptoms. Three hundred years ago, when Pascal described his “calculating machine” (a machine roughly similar to an automobile odometer), European academics were astonished that a machine could perform arithmetic, an intellectual ability that only a few humans possessed. Even as late as the second world war, when Turing was working, the ability to perform complex mental calculations rapidly was considered evidence of intellectual talent, found useful throughout science and engineering, and given academic recognition. The ability to perform simultaneous translation may soon be reduced to the “merely mechanical.”

When Eliza first appeared, some people found its conversational abilities quite human-like. No machine until then could have reacted even in such a simple way to what had been said to it. But during the Loebner competition, many programs were instantly revealed as nonhuman precisely by the first hint of resemblance of their behavior to that of Eliza. Amazingly, the boundary shifts in both directions. Some judges in the Loebner competition rated a machine as a machine on the grounds that she produced extended, well-written paragraphs of informative text, which is now apparently considered to be an inhuman ability in parts of our culture. The Loebner competition illustrates very clearly how the imitation game inevitably slides from a concern with cognitive status to being a test of the ability of the human species to discriminate its members from mechanical imposters.

Turing Tests suffer from another flaw: it is not clear what exactly they can be failing to detect. While Turing was careful to not suggest the test as a definition of humanity or intelligence, the fact that it is often described that way is revealing. Let us say that a Turing Test is a test of “human conversational competence.” But what is that, exactly? The only answer is the ability to pass the corresponding Turing Test. The tests are circular: they define the qualities they are claiming to be evidence for. But whatever their quality, it cannot be characteristic of humanity, since many humans would fail a Turing Test. Since one of the players must be judged to be a machine, half the human population would fail the species test.

5 Inhuman Intelligence

These have all been criticisms of the design of Turing’s test considered as some kind of experiment. One might argue, however, that it should be regarded more as a spur to technological progress, much as the goal of getting a man on the moon was undertaken to spur the development of the US space program. However, we will argue that using this test to define our field, even loosely, now leads the field to disown and reject its own successes.

Notice first how parochial, one might even say arrogant, a perspective is assumed by the imitation game. Why should we take it as our goal to build something which is just like us? A dog would never win any imitation game; but there seems to be no doubt that dogs exhibit cognition, and a machine with the cognitive and communicative abilities of a dog would be an interesting challenge for AI (and might useful be incorporated, for example, into automobiles.)

Likewise, the species detection aspect of the Turing Test has served to focus much AI research on those facets of human behavior which are least susceptible to useful generalization precisely because they are not shared by other species (silicon-based or otherwise). As we develop a general science of cognition, it is the aspects of human thought which are not distinctively human that seem the most fundamental. As others have emphasized, cognitive science is a science of cognition, not particularly of human cognition; but we cannot expect to be able to understand human cognition without first having a firm grasp of the basic principles of cognition.

From a practical perspective, why would anyone want to build machines that could pass the Turing Test? As many have observed, there is no shortage of humans, and we already have well-proven ways of making more of them. Human cognition, even high-quality human cognition, is not in short supply. What extra functionality would such a machine provide, even if we could build it?

One answer is that if we could make a human intelligence then we could make a superhuman intelligence just by getting a better processor and extra memory. This vision—the HAL vision of AI—is cited by several AI pioneers (e.g., McCarthy, Feigenbaum, Minsky) and by many people who are worried about what these superintelligences might do. But if we abandon the Turing Test vision, the goal naturally shifts from making artificial superhumans which can replace us, to making superhumanly intelligent artifacts which we can use to amplify and support our own cognitive abilities, just as people use hydraulic power to amplify their muscular abilities. This is in fact occurring, of course, and has been clearly foreseen and articulated by others; our point here is only to emphasize how different this goal is from the one that Turing left us with. AI should play a central role in this exciting new technology, but to do so it must turn its back on Turing’s dream.

The area generally called “expert systems” has been hugely successful as a technology but is widely perceived, both in the academic community and in the commercial marketplace, as having somehow failed to achieve its goal. The systems produced are often described as “brittle,” for example, which is a way of saying that they perform only in their intended domain. That it only performs its intended task would hardly be considered a criticism of most machines, however; and we suggest that it seems a valid criticism here solely because of the lingering influence of the Turing Test measure of AI success. Specialized AI systems are sometimes criticized as being “idiot savants”; but if we abandon the goal of making artificial people, we can rejoice in making useful idiot savants.

1 The worries seem to arise from the idea that the superhuman intelligence would not just be smart, but would also have superhuman political ambition and be vulnerable to human moral temptation. This might indeed be worse, but not even the imitation game requires this.
The authors look forward happily to having several such idiots for lawn mowing, tax preparation, etc... We are not here to simply reiterate the widespread "hype" complaint that AI has promised too much and found itself unable to deliver. In this area, in fact, AI systems have delivered the goods very well, sometimes spectacularly well. But even this success is often somehow sickied over with the pale cast of Turing Test insufficiency.

Low fidelity simulations of human behavior are quite a different goal from systems which complement, surpass, and extend our cognitive attributes. The Turing Test does not admit of weaker, different, or even stronger forms of intelligence than those deemed human. This puts AI engineering in a rather ridiculous position. Our most useful computer applications (including AI programs) are often valuable exactly by virtue of their lack of humanity. A truly human-like program would be nearly useless.

One can detect a trend in the marketplace in which instead of selling "intelligence," even a limited version of it called "expertise," engineers are incorporating what might be called cognitive functionality into products whose overall behavior would often not be thought of as particularly intelligent. As AI progresses, we become able to make computers do more and more things, and some of these would be regarded as requiring intelligence—or at any rate cognitive ability of some kind—if a human did them. But this functionality is not made into a special category called "AI ability," or taken to the market as anything having to do with human beings. In fact, the AI is often quite invisible in the final product. There are cameras, copiers, televisions, automobiles, battery rechargers and laptop operating systems all with algorithms incorporated into them which use AI ideas and techniques, but they are not usually advertised as "intelligent" or "expert." They certainly could not pass a Turing Test and there is no particular reason to suppose that they represent an application of a part or component of something that might one day pass a Turing Test. The designers of these systems construe AI as an enabling technology, and reject the Turing Test as a criterion for success. The influence of the Turing Test vision is so pervasive, however, that such work is often not called artificial-intelligence just for this reason. This is a tragedy for AI. Our subject is fuelling technical revolutions and changing the world, but Turing's ghost orders us to disinherit these successes.

One is not going to get something which can pass the Turing Test by eventually assembling a collection of these techniques. It would be both far too good and far too bad. It would be lightning-fast and superhumanly accomplished in some ways, curiously inept in others. We could find ways of disguising its inhuman talents, of course; Turing considers this kind of problem explicitly, observing that a machine can always pretend to be worse at arithmetic than it really is. But if one's aim is to provide better machines for people to use, what a silly business to get involved in! It is like glueing a beak and feathers onto an airplane to make it look more like a bird.

6 On Computational Wings

This flight metaphor is quite precise and worth pursuing in more detail. Early attempts to make flying machines often did things like attaching a beak onto the front, or trying to make a wing which would flap like a bird's wing (This extraordinarily persistent idea is found in Leonardo's notebooks and in a textbook on airplane design published in 1911). It is easy for us to smile at such naivete, but one should realize that it made good sense at the time. What birds did was incredible, and nobody really knew how they did it. It always seemed to involve feathers and flapping. Maybe the beak was critical for stability. When one's ignorance was almost total, it made good sense to copy as much of the natural thing as one could, if only to find out what aspects were essential and which were not. A few hundred years ago the idea of artificial flight could have been defined—indeed, often was so characterized—as the idea of making a machine that could fly like a bird. Birds were the only available exemplars for flight then, just as humans were the only exemplars for cognition when Turing was writing. The Turing Test version of artificial flight is just that: make a machine which would be indistinguishable from a bird, if all you could see of it was how it flew. This bird making was the goal of artificial flight for centuries. Most early attempts to make gliders copied aspects of bird structure. As late as 1980, Lielenthal's pioneering experiments with man-carrying gliders used wings and tails clearly based on bird anatomy, and a US patent was issued at the turn of the century for a "flying suit" with wing-linkages covered with feathers.

But progress was actually made when this aim of imitating nature was abandoned. The technology of flight advanced rapidly once workers gave themselves clearly-defined functional goals, separated from any notion of imitating biology, and strove to achieve these goals by any means available. The Wright brothers clearly separated the problems of power-to-weight ratio, lift, lateral stability, pitch and yaw control and solved them one at a time, using such unnatural devices as box kites, launch catapults and vertical fin surfaces. The first successful flyers were very unlike birds, and did not fly like birds. Likewise, the new science of aerodynamics made rapid progress only once it had artifacts with which to perform controlled experiments. The idea of the airfoil (which is crucial to the performance of all birds except the hummingbirds) was not discovered in nature. The shape of real bird wings is far too complex and flexible to suggest the idea of the airfoil; but once it had been discovered by experiments with artificial flyers, and its basic role understood from theory, a gull's wing can easily be recognized as one. Birds are incredibly efficient and clever flyers, and aeronautical engineers still look to them for inspiration; but this productive interaction between technology and biology did not come about by the engineers taking as their goal the task of imitating nature. Indeed, it happened as a direct result of abandoning that naif notion and seeking instead to identify general principles of stable flight and create machines based on them. Similar things are happening now in cognitive science where computational ideas originating in AI are being successfully applied in cognitive psychology, linguistics, and neuroscience.

Artificial flight both transcends and lags far behind natural flight, just as AI machines both surpass and lag behind human intelligence. Airplanes fly at Mach 3, miles above the clouds; but we doubt if an airplane will ever be able to land on the...
branch of a tree, or scoop a swimming fish from the ocean. Machines can’t lay eggs, of course, but even if we restrict ourselves to matters of flying, birds have talents that will probably always escape technology. No aircraft will pass the Turing Test for flight.

Perhaps human conversation will always be beyond computer abilities in its complexity and subtlety. If so, we should not think that AI has failed, even if the aim of our science is to understand intelligence and of our technology to amplify and extend it. Neither of them should be trying to reproduce it. That is unnecessary for the science and insufficient for the technology.

Even if one’s primary goal is essentially psychological, to understand human intelligence, attempting to build a replica of a human is not a sensible approach. But there is no reason why AI, or more generally cognitive science, should define itself in terms of human intelligence or cognition. While this was a natural way to begin, just as flight pioneers began by trying to imitate bird wings, the science itself provides explanations for cognition which deny the uniqueness of any biologically defined categories. Its general insights and ideas apply equally well to electronic computers as to nervous systems, much as aerodynamics applies equally well to airflow over a metal wing as to one covered with feathers. This is not a new observation, but we have only recently begun to understand the extent to which it implies a rejection of imitation-game criteria for success in AI, and how pervasive the consequences of these criteria are.

7 Turing’s Ghost
Two venerable intellectual threads weave through history and converge on Alan Turing. To his great credit he grasped them and started knitting what has become the rich tapestry of motivations and ideas that comprise AI. One thread is the idea that machines might somehow process meanings, which runs through Hobbes, Pascal, Leibniz, Boole, Babbage, and many others. The other is the ancient ambition, which is probably older than civilization, to steal divine power by making something come alive. It is reflected, for example, in the Greek myth of Pygmalion, the Golem legend, and the Frankenstein story. Turing was perhaps the first person to be in a position to see how these ancient themes might be brought together. Whether or not he intended it, his insight that technology might, at long last, be able to reach a kind of divine power almost certainly played a key role in motivating early AI projects. Viewed in this historical context, Turing’s suggestion of an imitation game seems more understandable, but the same historical view suggests strongly that we must now distinguish legend from science. AI is the proud heir of Boole, Babbage, and Turing, but not of Mary Shelley.

We suspect that several subfields of AI have tended to reject their association with their parent precisely because they found it necessary to develop methodologies which are inconsistent with any kind of Turing Test. Vision, for example, is a perfectly well defined area of scientific investigation or technological ambition within which one can work without feeling obliged to also thereby accept a larger goal of creating a complete intelligent machine. Just as the Turing Tests allow for no degree of partial success, the research programs they define cannot be sensibly taken apart into subfields without an implicit agreement to conform to some kind of grand intellectual architecture, which is not a reasonable constraint to put on either a science or a technology. Turing’s legacy alienates maturing subfields with methodological inconsistencies. Abandoning the Turing test as an ultimate goal is almost a requirement for any rational research program which declares itself interested in any particular part of cognition or mental ability. Let us emphasize again, we do not deny the need for this abandonment. The harm is done when this is perceived as abandoning AI.

Allowing our field to be defined by a Turing Test also harms its reputation, in direct and subtle ways. Perhaps not surprisingly, many lay critics of AI assume the field to be defined by Turing’s goal, and by this light it does not seem to be doing very well. For just one example, Frederick Allen, writing in ‘The Atlantic’ (1994):

> Today traditional artificial intelligence, or AI, is a backwater at best, and the confidence with which it was once pursued seems unimaginable. Nobody has ever designed a program that can converse at all convincingly on a single subject, and the field has splintered into disparate parts. ... The grand vision has nearly vanished.

Such a pessimistic summary of a flourishing research area like ours may seem merely to reflect ignorance. But if one identifies AI with the goal of passing a Turing Test—the “grand vision” to which Allen refers—then he is perfectly correct. The mistake lies in allowing that identification.

Another much-cited attack on AI may arise in part from an insight into the Turing Test. As we have emphasized, in order to succeed at the imitation game even a human player would be obliged to think consciously, to an unnatural extent, about what effects his utterances might be having on a listener. It is quite natural to go from this insight to Searle’s idea that AI programs can be only a simulation of cognition, leading to his notorious distinction between strong and weak AI.

The Turing Test indeed challenges a computer to simulate a woman, rather than be one.

Finally, perhaps the most subtle kind of damage which the Turing Test has done to AI is by limiting its sights. Ironically, Turing’s daring vision may in fact be too restrictive. All versions of the Turing Test are based on a massively anthropocentric view of the nature of intelligence. Turing correctly insisted that his test was not meant to define intelligence. Nevertheless, in giving us this touchstone of success, he chose human intelligence—in fact, even more peculiarly, the arguing skill of the educated English middle class in playing a kind of party game—as our goal. But the very science which Turing directed us towards provides a perspective from which a much broader and more satisfying account of intelligence is emerging. The Turing Test focuses our attention on the most human details of behavior, rather than general computational principles of cognition.

8 What Is AI?
AI has always wondered how to define itself, and has engaged in a long-running territorial battle with other parts of computer science. Techniques are often developed in AI and later absorbed into mainstream computer science. Unlike most subdisciplines of computer science, AI seems to be defined
not by its methods but by the source of their inspiration.\footnote{Attempts to define AI in terms of its computational methods never work properly. For example, if AI is the study of search then successful learning processes automatically remove themselves from the discipline.}

So, which parts of computer science are part of AI? We suggest a rather radical answer to this question: all of them. AI is not a part of computer science in the way that compiler design, object-oriented programming or genetic algorithms are. AI is the business of using computation to make machines act more intelligently, or to somehow amplify human intelligence. It is not a particular collection of methods, or a programming style. Any technique can be used by a program to do something intelligent, or to display a cognitive ability.

Until perhaps a decade ago many computational methods were pioneered in AI or in close association with it. One of the first compilers was reported at a meeting on intelligent machines. Larry Tesler has suggested that AI be defined to be the part of computer science where things don't work properly yet; the edge of the ice, as it were. But this may have been simply a historical consequence of the fact that many of the creative pioneers of computer science had accepted Turing's dream, and were struggling to make computers act "intelligently." It isn't true any longer, and many of the most exciting new ideas in computation are now being developed in other parts of computer science which have quite different aims. But it would be foolish to regard these methods as somehow excluded from AI.

For example, there has been a long-standing intellectual struggle in machine translation between methods based on explicit semantic representations and those which apply statistical techniques to large lexical corpora. This is often described as a battle between AI methods and other, non-AI methods. While this may be an accurate account of the sociology of the two sides, it makes no scientific or technological sense. Our aim might be to model the skill of human translators, or to make an effective mechanical translator: either way, we should not have any ad hoc constraints on what computational methods to use. If it works, or seems plausible, try it. AI has difficult enough problems already, without also having its technical hands tied.

Consider again the analogy with flight. If cognitive psychology, psycholinguistics, etc. are like the study of natural flight in all its complexity, and AI is like aeronautical engineering, then computer science supplies the aerodynamic theory. The fundamental insight of cognitive science might be summarized by saying that computational science supplies, as it were, the dynamics of cognition. Just as Turing predicted almost half a century ago, the empirical sciences of natural cognition now share a computational vocabulary with the engineering discipline of AI.

This picture of our field defines it in a more useful and more mature way than Turing could give us. AI is the engineering of cognition based on the computational vision which runs through and informs all of cognitive science. We expect AI to produce cognitive artifacts: things that think, see, communicate, plan, play and argue in some way. Perhaps not in a human way, but somehow useful to humans. Exactly what counts as "cognitive" will shift and change, and be altered by the science itself, just as the meaning of words like "energy" has been changed by physics. But ultimately, this doesn't matter. Turing's ultimate aim, which we can happily share, was not to describe the difference between thinking people and unthinking machines, but to remove it.

9 Coda: The Human Condition

Colby [1975, 1981] has argued that we should consider variations of the gender test, where the judge is asked to make different kinds of distinction. For example, the judge might be asked to decide which of the interrogants was really a child, or really an Englishman; or, in a more familiar example, the judge might be a clinical psychologist trying to diagnose which of them is really paranoid. On this view, Turing's choice of sex as the topic of conversation had no particular significance. It may have been chosen simply because it seems impossible to give any a priori bounds on the subject matter of the resulting conversation.

However, Turing was a careful enough thinker that he would have suggested this diverse-topic interpretation of his game had this been what he had in mind, and we again propose to take him at face value. He seems to have chosen the topic of sexual identity deliberately. It is hard to avoid noticing that for Turing, the problem of how to convincingly display a sexual identity was more than just deliberately vague. It was a real problem at the very core of his emotional and social life. Turing was openly gay at a time when homosexuality was a crime in England and was widely regarded as unnatural and deviant. He was prosecuted for homosexuality, and avoided prison only by submitting to a six-month program of "rehabilitation" involving hormone injections which, among other things, caused his body to grow breasts. This bizarre and horrifying treatment is thought to have been part of the reason for his suicide in 1954.

We suspect that Turing chose this topic because he wanted the test to be about what it really means to be human. This is why he has set us up in this way. He tells us, quite clearly, to try to make a program which can do as well as a man at pretending to be a woman. If we really tried to do this, we might be forced into thinking very hard about what it really means to be not just a thinker, but a human being in a human society, with all its difficulties and complexities. If this was what Turing meant, then we need not reject it as our ultimate goal.

References
