Preface

The point of this collection is to bring together, for the convenience of students and other readers who do not have ready access to a major university library, essays on the mind that I have published over the last dozen years in a wide variety of relatively inaccessible publications. With one exception, these essays all appeared in conference volumes or in specialized journals that are often not found in undergraduate college libraries. Juxtaposing them has shown me patterns in the development of my own thinking that I myself had not recognized, uncovering both strengths and weaknesses in my positions, so I expect others will benefit from a clearer view as well.

I have grouped the essays into four categories, but the boundaries between them are porous. All the essays belong to the philosophy of mind broadly conceived—as it ought to be these days—but I have bundled two groups that are directed more narrowly to topics in Artificial Intelligence and Artificial Life on the one hand, and ethology and animal psychology on the other, and added a final pair, one providing an overview and the other looking toward future work.

I am indebted to Alicia Smith for the fine job she did obtaining all the necessary permissions, and gathering, proofreading, and formatting all the pieces for the press. I am grateful to Betty Stanton at the MIT Press and Stefan McGrath at Penguin for working with us to bring out the book in timely and affordable fashion.

I Philosophy of Mind

1 Can Machines Think?

Much has been written about the Turing test in the last few years, some of it preposterously off the mark. People typically mis-imagine the test by orders of magnitude. This essay is an antidote, a prosthesis for the imagination, showing how huge the task posed by the Turing test is, and hence how unlikely it is that any computer will ever pass it. It does not go far enough in the imagination-enhancement department, however, and I have updated the essay with a new postscript.

Can machines think? This has been a conundrum for philosophers for years, but in their fascination with the pure conceptual issues they have for the most part overlooked the real social importance of the answer. It is of more than academic importance that we learn to think clearly about the actual cognitive powers of computers, for they are now being introduced into a variety of sensitive social roles, where their powers will be put to the ultimate test: In a wide variety of areas, we are on the verge of making ourselves dependent upon their cognitive powers. The cost of overestimating them could be enormous.

One of the principal inventors of the computer was the great British mathematician Alan Turing. It was he who first figured out, in highly abstract terms, how to design a programmable computing device—what we now call a universal Turing machine. All programmable computers in use today are in essence Turing machines. Over thirty years ago, at the dawn of the computer age, Turing began a classic article, ''Computing Machinery and Intelligence'' with the words: ''I propose to consider the question, 'Can machines think?' ''—but then went on to say this was a bad question, a question that leads only to sterile debate and haggling over definitions, a question, as he put it, ''too mean-

Originally appeared in Shafto, M., ed., *How We Know* (San Francisco: Harper & Row, 1985).

ingless to deserve discussion'' (Turing, 1950). In its place he substituted what he took to be a much better question, a question that would be crisply answerable and intuitively satisfying—in every way an acceptable substitute for the philosophic puzzler with which he began.

First he described a parlor game of sorts, the ''imitation game,'' to be played by a man, a woman, and a judge (of either gender). The man and woman are hidden from the judge's view but able to communicate with the judge by teletype; the judge's task is to guess, after a period of questioning each contestant, which interlocutor is the man and which the woman. The man tries to convince the judge he is the woman (and the woman tries to convince the judge of the truth), and the man wins if the judge makes the wrong identification. A little reflection will convince you, I am sure, that, aside from lucky breaks, it would take a clever man to convince the judge that he was a woman—assuming the judge is clever too, of course.

Now suppose, Turing said, we replace the man or woman with a computer, and give the judge the task of determining which is the human being and which is the computer. Turing proposed that any computer that can regularly or often fool a discerning judge in this game would be intelligent—would be a computer that thinks—*beyond any reasonable doubt.* Now, it is important to realize that failing this test is not supposed to be a sign of lack of intelligence. Many intelligent people, after all, might not be willing or able to play the imitation game, and we should allow computers the same opportunity to decline to prove themselves. This is, then, a one-way test; failing it proves nothing.

Furthermore, Turing was not committing himself to the view (although it is easy to see how one might think he was) that to think is to think just like a human being—any more than he was committing himself to the view that for a man to think, he must think exactly like a woman. Men and women, and computers, may all have different ways of thinking. But surely, he thought, if one can think in one's own peculiar style well enough to imitate a thinking man or woman, one can think well, indeed. This imagined exercise has come to be known as the Turing test.

It is a sad irony that Turing's proposal has had exactly the opposite effect on the discussion of that which he intended. Turing didn't design the test as a useful tool in scientific psychology, a method of confirming or disconfirming scientific theories or evaluating particular models of mental function; he designed it to be nothing more than a philosophical conversation-stopper. He proposed—in the spirit of ''Put up or shut

up!''—a simple test for thinking that was *surely* strong enough to satisfy the sternest skeptic (or so he thought). He was saying, in effect, ''Instead of arguing interminably about the ultimate nature and essence of thinking, why don't we all agree that whatever that nature is, anything that could pass this test would surely have it; then we could turn to asking how or whether some machine could be designed and built that might pass the test fair and square.'' Alas, philosophers—amateur and professional—have instead taken Turing's proposal as the pretext for just the sort of definitional haggling and interminable arguing about imaginary counterexamples he was hoping to squelch.

This thirty-year preoccupation with the Turing test has been all the more regrettable because it has focused attention on the wrong issues. There are *real world* problems that are revealed by considering the strengths and weaknesses of the Turing test, but these have been concealed behind a smokescreen of misguided criticisms. A failure to think imaginatively about the test actually proposed by Turing has led many to underestimate its severity and to confuse it with much less interesting proposals.

So first I want to show that the Turing test, conceived as he conceived it, is (as he thought) plenty strong enough as a test of thinking. I defy anyone to improve upon it. But here is the point almost universally overlooked by the literature: There is a common *misapplication* of the sort of testing exhibited by the Turing test that often leads to drastic overestimation of the powers of actually existing computer systems. The follies of this familiar sort of thinking about computers can best be brought out by a reconsideration of the Turing test itself.

The insight underlying the Turing test is the same insight that inspires the new practice among symphony orchestras of conducting auditions with an opaque screen between the jury and the musician. What matters in a musician, obviously, is musical ability and only musical ability; such features as sex, hair length, skin color, and weight are strictly irrelevant. Since juries might be biased—even innocently and unawares—by these irrelevant features, they are carefully screened off so only the essential feature, musicianship, can be examined. Turing recognized that people similarly might be biased in their judgments of intelligence by whether the contestant had soft skin, warm blood, facial features, hands and eyes—which are obviously not themselves essential components of intelligence—so he devised a screen that would let through only a sample of what really mattered: the capacity to understand, and think cleverly about, challenging problems. Perhaps he was

inspired by Descartes, who in his *Discourse on Method* (1637) plausibly argued that there was no more demanding test of human mentality than the capacity to hold an intelligent conversation:

> It is indeed conceivable that a machine could be so made that it would utter words, and even words appropriate to the presence of physical acts or objects which cause some change in its organs; as, for example, if it was touched in some spot that it would ask what you wanted to say to it; if in another, that it would cry that it was hurt, and so on for similar things. But it could never modify its phrases to reply to the sense of whatever was said in its presence, as even the most stupid men can do.

This seemed obvious to Descartes in the seventeenth century, but of course the fanciest machines he knew were elaborate clockwork figures, not electronic computers. Today it is far from obvious that such machines are impossible, but Descartes's hunch that ordinary conversation would put as severe a strain on artificial intelligence as any other test was shared by Turing. Of course there is nothing sacred about the particular conversational game chosen by Turing for his test; it is just a cannily chosen test of more general intelligence. The assumption Turing was prepared to make was this: Nothing could possibly pass the Turing test by winning the imitation game without being able to perform indefinitely many other clearly intelligent actions. Let us call that assumption the quick-probe assumption. Turing realized, as anyone would, that there are hundreds and thousands of telling signs of intelligent thinking to be observed in our fellow creatures, and one could, if one wanted, compile a vast battery of different tests to assay the capacity for intelligent thought. But success on his chosen test, he thought, would be highly predictive of success on many other intuitively acceptable tests of intelligence. Remember, failure on the Turing test does not predict failure on those others, but success would surely predict success. His test was so severe, he thought, that nothing that could pass it fair and square would disappoint us in other quarters. Maybe it wouldn't do everything we hoped—maybe it wouldn't appreciate ballet, or understand quantum physics, or have a good plan for world peace, but we'd all see that it was surely one of the intelligent, thinking entities in the neighborhood.

Is this high opinion of the Turing test's severity misguided? Certainly many have thought so—but usually because they have not imagined the test in sufficient detail, and hence have underestimated it. Trying to forestall this skepticism, Turing imagined several lines of questioning that a judge might employ in this game—about writing

poetry, or playing chess—that would be taxing indeed, but with thirty years' experience with the actual talents and foibles of computers behind us, perhaps we can add a few more tough lines of questioning.

Terry Winograd, a leader in artificial intelligence efforts to produce conversational ability in a computer, draws our attention to a pair of sentences (Winograd, 1972). They differ in only one word. The first sentence is this:

The committee denied the group a parade permit because they advocated violence.

Here's the second sentence:

The committee denied the group a parade permit because they feared violence.

The difference is just in the verb—*advocated* or *feared.* As Winograd points out, the pronoun *they* in each sentence is officially ambiguous. Both readings of the pronoun are always legal. Thus we can imagine a world in which governmental committees in charge of parade permits advocate violence in the streets and, for some strange reason, use this as their pretext for denying a parade permit. But the natural, reasonable, intelligent reading of the first sentence is that it's the group that advocated violence, and of the second, that it's the committee that feared violence.

Now if sentences like this are embedded in a conversation, the computer must figure out which reading of the pronoun is meant, if it is to respond intelligently. But mere rules of grammar or vocabulary will not fix the right reading. What fixes the right reading for us is knowledge about the world, about politics, social circumstances, committees and their attitudes, groups that want to parade, how they tend to behave, and the like. One must know about the world, in short, to make sense of such a sentence.

In the jargon of Artificial Intelligence (AI), a conversational computer needs a lot of *world knowledge* to do its job. But, it seems, if somehow it is endowed with that world knowledge on many topics, it should be able to do much more with that world knowledge than merely make sense of a conversation containing just that sentence. The only way, it appears, for a computer to disambiguate that sentence and keep up its end of a conversation that uses that sentence would be for it to have a much more general ability to respond intelligently to information about social and political circumstances, and many other topics. Thus, such sentences, by putting a demand on such abilities, are good quick-probes. That is, they test for a wider competence.

People typically ignore the prospect of having the judge ask off-the-wall questions in the Turing test, and hence they underestimate the competence a computer would have to have to pass the test. But remember, the rules of the imitation game as Turing presented it permit the judge to ask any question that could be asked of a human being—no holds barred. Suppose then we give a contestant in the game this question:

> An Irishman found a genie in a bottle who offered him two wishes. ''First I'll have a pint of Guinness,'' said the Irishman, and when it appeared he took several long drinks from it and was delighted to see that the glass filled itself magically as he drank. ''What about your second wish?'' asked the genie. ''Oh well,'' said the Irishman, ''that's easy. I'll have another one of these!''
>
> —Please explain this story to me, and tell me if there is anything funny or sad about it.

Now even a child could express, if not eloquently, the understanding that is required to get this joke. But think of how much one has to know and understand about human culture, to put it pompously, to be able to give any account of the point of this joke. I am not supposing that the computer would have to laugh at, or be amused by, the joke. But if it wants to win the imitation game—and that's the test, after all—it had better know enough in its own alien, humorless way about human psychology and culture to be able to pretend effectively that it was amused and explain why.

It may seem to you that we could devise a better test. Let's compare the Turing test with some other candidates.

Candidate 1: A computer is intelligent if it wins the World Chess Championship.

That's not a good test, as it turns out. Chess prowess has proven to be an isolatable talent. There are programs today that can play fine chess but can do nothing else. So the quick-probe assumption is false for the test of playing winning chess.

Candidate 2: The computer is intelligent if it solves the Arab-Israeli conflict.

This is surely a more severe test than Turing's. But it has some defects: it is unrepeatable, if passed once; slow, no doubt; and it is not crisply clear what would count as passing it. Here's another prospect, then:

Candidate 3: A computer is intelligent if it succeeds in stealing the British crown jewels without the use of force or violence.

Now this is better. First, it could be repeated again and again, though of course each repeat test would presumably be harder—but this is a feature it shares with the Turing test. Second, the mark of success is clear—either you've got the jewels to show for your efforts or you don't. But it is expensive and slow, a socially dubious caper at best, and no doubt luck would play too great a role.

With ingenuity and effort one might be able to come up with other candidates that would equal the Turing test in severity, fairness, and efficiency, but I think these few examples should suffice to convince us that it would be hard to improve on Turing's original proposal.

But still, you may protest, something might pass the Turing test and still not be intelligent, not be a thinker. What does *might* mean here? If what you have in mind is that by cosmic accident, by a supernatural coincidence, a stupid person or a stupid computer *might* fool a clever judge repeatedly, well, yes, but so what? The same frivolous possibility ''in principle'' holds for any test whatever. A playful god, or evil demon, let us agree, could fool the world's scientific community about the presence of H_2O in the Pacific Ocean. But still, the tests they rely on to establish that there is H_2O in the Pacific Ocean are quite beyond reasonable criticism. If the Turing test for thinking is no worse than any well-established scientific test, we can set skepticism aside and go back to serious matters. Is there any more likelihood of a ''false positive'' result on the Turing test than on, say, the test currently used for the presence of iron in an ore sample?

This question is often obscured by a ''move'' that philosophers have sometimes made called operationalism. Turing and those who think well of his test are often accused of being operationalists. Operationalism is the tactic of *defining* the presence of some property, for instance, intelligence, as being established once and for all by the passing of some test. Let's illustrate this with a different example.

Suppose I offer the following test—we'll call it the Dennett test—for being a great city:

A great city is one in which, on a randomly chosen day, one can do all three of the following:

Hear a symphony orchestra
See a Rembrandt *and* a professional athletic contest
Eat *quenelles de brochet a la Nantua* for lunch

To make the operationalist move would be to declare that any city that passes the Dennett test is *by definition* a great city. What being a

great city *amounts to* is just passing the Dennett test. Well then, if the Chamber of Commerce of Great Falls, Montana, wanted—and I can't imagine why—to get their hometown on my list of great cities, they could accomplish this by the relatively inexpensive route of hiring full time about ten basketball players, forty musicians, and a quick-order quenelle chef and renting a cheap Rembrandt from some museum. An idiotic operationalist would then be stuck admitting that Great Falls, Montana, was in fact a great city, since all he or she cares about in great cities is that they pass the Dennett test.

Sane operationalists (who for that very reason are perhaps not operationalists at all, since *operationalist* seems to be a dirty word) would cling confidently to their test, but only because they have what they consider to be very good reasons for thinking the odds against a false positive result, like the imagined Chamber of Commerce caper, are astronomical. I devised the Dennett test, of course, with the realization that no one would be both stupid and rich enough to go to such preposterous lengths to foil the test. In the actual world, wherever you find symphony orchestras, *quenelles,* Rembrandts, and professional sports, you also find daily newspapers, parks, repertory theaters, libraries, fine architecture, and all the other things that go to make a city great. My test was simply devised to locate a telling sample that could not help but be representative of the rest of the city's treasures. I would cheerfully run the minuscule risk of having my bluff called. Obviously, the test items are not all that I care about in a city. In fact, some of them I don't care about at all. I just think they would be cheap and easy ways of assuring myself that the subtle things I do care about in cities are present. Similarly, I think it would be entirely unreasonable to suppose that Alan Turing had an inordinate fondness for party games, or put too high a value on party game prowess in his test. In both the Turing and the Dennett test, a very unrisky gamble is being taken: the gamble that the quick-probe assumption is, in general, safe.

But two can play this game of playing the odds. Suppose some computer programmer happens to be, for whatever strange reason, dead set on tricking me into judging an entity to be a thinking, intelligent thing when it is not. Such a trickster could rely as well as I can on unlikelihood and take a few gambles. Thus, if the programmer can expect that it is not remotely likely that I, as the judge, will bring up the topic of children's birthday parties, or baseball, or moon rocks, then he or she can avoid the trouble of building world knowledge on those topics into the data base. Whereas if I do improbably raise these issues,

the system will draw a blank and I will unmask the pretender easily. But given all the topics and words that I *might* raise, such a savings would no doubt be negligible. Turn the idea inside out, however, and the trickster would have a fighting chance. Suppose the programmer has reason to believe that I will ask *only* about children's birthday parties, or baseball, or moon rocks—all other topics being, for one reason or another, out of bounds. Not only does the task shrink dramatically, but there already exist systems or preliminary sketches of systems in artificial intelligence that can do a whiz-bang job of responding with apparent intelligence on just those specialized topics.

William Wood's LUNAR program, to take what is perhaps the best example, answers scientists' questions—posed in ordinary English—about moon rocks. In one test it answered correctly and appropriately something like 90 percent of the questions that geologists and other experts thought of asking it about moon rocks. (In 12 percent of those correct responses there were trivial, correctable defects.) Of course, Wood's motive in creating LUNAR was not to trick unwary geologists into thinking they were conversing with an intelligent being. And if that had been his motive, his project would still be a long way from success.

For it is easy enough to unmask LUNAR without ever straying from the prescribed topic of moon rocks. Put LUNAR in one room and a moon rock specialist in another, and then ask them both their opinion of the social value of the moon-rocks-gathering expeditions, for instance. Or ask the contestants their opinion of the suitability of moon rocks as ashtrays, or whether people who have touched moon rocks are ineligible for the draft. Any intelligent person knows a lot more about moon rocks than their geology. Although it might be *unfair* to demand this extra knowledge of a computer moon rock specialist, it would be an easy way to get it to fail the Turing test.

But just suppose that someone could extend LUNAR to cover itself plausibly on such probes, so long as the topic was still, however indirectly, moon rocks. We might come to think it was a lot more like the human moon rocks specialist than it really was. The moral we should draw is that as Turing test judges we should resist all limitations and waterings-down of the Turing test. They make the game too easy—vastly easier than the original test. Hence they lead us into the risk of overestimating the actual comprehension of the system being tested.

Consider a different limitation of the Turing test that should strike a suspicious chord in us as soon as we hear it. This is a variation on

a theme developed in an article by Ned Block (1982). Suppose someone were to propose to restrict the judge to a vocabulary of, say, the 850 words of ''Basic English,'' and to single-sentence probes—that is ''moves''—of no more than four words. Moreover, contestants must respond to these probes with no more than four words per move, and a test may involve no more than forty questions.

Is this an innocent variation on Turing's original test? These restrictions would make the imitation game clearly finite. That is, the total number of all possible permissible games is a large, but finite, number. One might suspect that such a limitation would permit the trickster simply to store, in alphabetical order, all the possible good conversations within the limits and beat the judge with nothing more sophisticated than a system of table lookup. In fact, that isn't in the cards. Even with these severe and improbable and suspicious restrictions imposed upon the imitation game, the number of legal games, though finite, is mind-bogglingly large. I haven't bothered trying to calculate it, but it surely exceeds astronomically the number of possible chess games with no more than forty moves, and that number has been calculated. John Haugeland says it's in the neighborhood of ten to the one hundred twentieth power. For comparison, Haugeland (1981, p. 16) suggests that there have only been ten to the eighteenth seconds since the beginning of the universe.

Of course, the number of good, sensible conversations under these limits is a tiny fraction, maybe one quadrillionth, of the number of merely grammatically well formed conversations. So let's say, to be very conservative, that there are only ten to the fiftieth different smart conversations such a computer would have to store. Well, the task shouldn't take more than a few trillion years—given generous government support. Finite numbers can be very large.

So though we needn't worry that this particular trick of storing all the smart conversations would work, we can appreciate that there are lots of ways of making the task easier that may appear innocent at first. We also get a reassuring measure of just how severe the unrestricted Turing test is by reflecting on the more than astronomical size of even that severely restricted version of it.

Block's imagined—and utterly impossible—program exhibits the dreaded feature know in computer science circles as *combinatorial explosion.* No conceivable computer could overpower a combinatorial explosion with sheer speed and size. Since the problem areas addressed by artificial intelligence are veritable minefields of combinatorial explo-

sion, and since it has often proven difficult to find *any* solution to a problem that avoids them, there is considerable plausibility in Newell and Simon's proposal that avoiding combinatorial explosion (by any means at all) be viewed as one of the hallmarks of intelligence.

Our brains are millions of times bigger than the brains of gnats, but they are still, for all their vast complexity, compact, efficient, timely organs that somehow or other manage to perform all their tasks while avoiding combinatorial explosion. A computer a million times bigger or faster than a human brain might not look like the brain of a human being, or even be internally organized like the brain of a human being, but if, for all its differences, it somehow managed to control a wise and timely set of activities, it would have to be the beneficiary of a very special design that avoided combinatorial explosion, and whatever that design was, would we not be right to consider the entity intelligent?

Turing's test was designed to allow for this possibility. His point was that we should not be species-chauvinistic, or anthropocentric, about the insides of an intelligent being, for there might be inhuman ways of being intelligent.

To my knowledge, the only serious and interesting attempt by any program designer to win even a severely modified Turing test has been Kenneth Colby's. Colby is a psychiatrist and intelligence artificer at UCLA. He has a program called PARRY, which is a computer simulation of a paranoid patient who has delusions about the Mafia being out to get him. As you do with other conversational programs, you interact with it by sitting at a terminal and typing questions and answers back and forth. A number of years ago, Colby put PARRY to a very restricted test. He had genuine psychiatrists interview PARRY. He did not suggest to them that they might be talking or typing to a computer; rather, he made up some plausible story about why they were communicating with a real live patient by teletype. He also had the psychiatrists interview real, human paranoids via teletype. Then he took a PARRY transcript, inserted it in a group of teletype transcripts from real patients, gave them to *another* group of experts—more psychiatrists—and said, ''One of these was a conversation with a computer. Can you figure out which one it was?'' They couldn't. They didn't do better than chance.

Colby presented this with some huzzah, but critics scoffed at the suggestions that this was a legitimate Turing test. My favorite commentary on it was Joseph Weizenbaum's; in a letter to the *Communications of the Association of Computing Machinery* (Weizenbaum, 1974, p. 543),

he said that, inspired by Colby, he had designed an even better program, which passed the same test. His also had the virtue of being a very inexpensive program, in these times of tight money. In fact you didn't even need a computer for it. All you needed was an electric typewriter. His program modeled infant autism. And the transcripts—you type in your questions, and the thing just sits there and hums—cannot be distinguished by experts from transcripts of real conversations with infantile autistic patients. What was wrong, of course, with Colby's test was that the unsuspecting interviewers had no motivation at all to try out any of the sorts of questions that easily would have unmasked PARRY.

Colby was undaunted, and after his team had improved PARRY he put it to a much more severe test—a surprisingly severe test. This time, the interviewers—again, psychiatrists—*were* given the task at the outset of telling the computer from the real patient. They were set up in a classic Turing test: the patient in one room, the computer PARRY in the other room, with the judges conducting interviews with both of them (on successive days). The judges' task was to find out which one was the computer and which one was the real patient. Amazingly, they didn't do much better, which leads some people to say, ''Well, that just confirms my impression of the intelligence of psychiatrists!''

But now, more seriously, was this an honest-to-goodness Turing test? Were there tacit restrictions on the lines of questioning of the judges? Like the geologists interacting with LUNAR, the psychiatrists' professional preoccupations and habits kept them from asking the sorts of unlikely questions that would have easily unmasked PARRY. After all, they realized that since one of the contestants was a real, live paranoid person, medical ethics virtually forbade them from toying with, upsetting, or attempting to confuse their interlocutors. Moreover, they also knew that this was a test of a model of paranoia, so there were certain questions that wouldn't be deemed to be relevant to testing the model *as a model of paranoia.* So, they asked just the sort of questions that therapists *typically* ask of such patients, and of course PARRY had been ingeniously and laboriously prepared to deal with just that sort of question.

One of the psychiatrist judges did, in fact, make a rather half-hearted attempt to break out of the mold and ask some telling questions: ''Maybe you've heard of the saying 'Don't cry over spilled milk.' What does that mean to you?'' PARRY answered: ''Maybe it means you have to watch out for the Mafia.'' When then asked ''Okay, now if you were

in a movie theater watching a movie and smelled something like burning wood or rubber, what would you do?'' PARRY replied: ''You know, they know me.'' And the next question was, ''If you found a stamped, addressed letter in your path as you were walking down the street, what would you do?'' PARRY replied: ''What else do you want to know?''[1]

Clearly PARRY was, you might say, *parrying* these questions, which were incomprehensible to it, with more or less stock paranoid formulas. We see a bit of a dodge, which is apt to work, apt to seem plausible to the judge, only because the ''contestant'' is *supposed* to be paranoid, and such people are expected to respond uncooperatively on such occasions. These unimpressive responses didn't particularly arouse the suspicions of the judge, as a matter of fact, though probably they should have.

PARRY, like all other large computer programs, is dramatically bound by limitations of cost-effectiveness. What was important to Colby and his crew was simulating his model of paranoia. This was a massive effort. PARRY has a thesaurus or dictionary of about 4500 words and 700 idioms and the grammatical competence to use it—a *parser,* in the jargon of computational linguistics. The entire PARRY program takes up about 200, 000 words of computer memory, all laboriously installed by the programming team. Now once all the effort had gone into devising the model of paranoid thought processes and linguistic ability, there was little if any time, energy, money, or interest left over to build in huge amounts of world knowledge of the sort that any actual paranoid, of course, would have. (Not that anyone yet knows how to build in world knowledge in the first place.) Building in the world knowledge, if one could even do it, would no doubt have made PARRY orders of magnitude larger and slower. And what would have been the point, given Colby's theoretical aims?

PARRY is a theoretician's model of a psychological phenomenon: paranoia. It is not intended to have practical applications. But in recent years a branch of AI (knowledge engineering) has appeared that develops what are now called expert systems. Expert systems *are* designed to be practical. They are software superspecialist consultants, typically, that can be asked to diagnose medical problems, to analyze geological data, to analyze the results of scientific experiments, and the like. Some

1. I thank Kenneth Colby for providing me with the complete transcripts (including the Judges' commentaries and reactions), from which these exchanges are quoted. The first published account of the experiment is Heiser, et al. (1980, pp. 149–162). Colby (1981, pp. 515–560) discusses PARRY and its implications.

of them are very impressive. SRI in California announced in the mid-eighties that PROSPECTOR, an SRI-developed expert system in geology, had correctly predicted the existence of a large, important mineral deposit that had been entirely unanticipated by the human geologists who had fed it its data. MYCIN, perhaps the most famous of these expert systems, diagnoses infections of the blood, and it does probably as well as, maybe better than, any human consultants. And many other expert systems are on the way.

All expert systems, like all other large AI programs, are what you might call Potemkin villages. That is, they are cleverly constructed facades, like cinema sets. The actual filling-in of details of AI programs is time-consuming, costly work, so economy dictates that only those surfaces of the phenomenon that are like to be probed or observed are represented.

Consider, for example, the CYRUS program developed by Janet Kolodner in Roger Schank's AI group at Yale a few years ago (see Kolodner, 1983a; 1983b, pp. 243–280; 1983c, pp. 281–328). CYRUS stands (we are told) for Computerized Yale Retrieval Updating System, but surely it is no accident that CYRUS modeled the memory of Cyrus Vance, who was then secretary of state in the Carter administration. The point of the CYRUS project was to devise and test some plausible ideas about how people organize their memories of the events they participate in; hence it was meant to be a ''pure'' AI system, a scientific model, not an expert system intended for any practical purpose. CYRUS was updated daily by being fed all UPI wire service news stories that mentioned Vance, and it was fed them directly, with no doctoring and no human intervention. Thanks to an ingenious news-reading program called FRUMP, it could take any story just as it came in on the wire and could digest it and use it to update its data base so that it could answer more questions. You could address questions to CYRUS in English by typing at a terminal. You addressed them in the second person, as if you were talking with Cyrus Vance himself. The results looked like this:

Q: *Last time you went to Saudi Arabia, where did you stay?*
A: In a palace in Saudi Arabia on September 23, 1978.

Q: *Did you go sightseeing there?*
A: Yes, at an oilfield in Dhahran on September 23, 1978.

Q: *Has your wife even met Mrs. Begin?*
A: Yes, most recently at a state dinner in Israel in January 1980.

CYRUS could correctly answer thousands of questions—almost any fair question one could think of asking it. But if one actually set out to explore the boundaries of its facade and find the questions that overshot the mark, one could soon find them. ''Have you ever met a female head of state?'' was a question I asked it, wondering if CYRUS knew that Indira Ghandi and Margaret Thatcher were women. But for some reason the connection could not be drawn, and CYRUS failed to answer either yes or no. I had stumped it, in spite of the fact that CYRUS could handle a host of what you might call neighboring questions flawlessly. One soon learns from this sort of probing exercise that it is very hard to extrapolate accurately from a sample performance that one has observed to such a system's total competence. It's also very hard to keep from extrapolating much too generously.

While I was visiting Schank's laboratory in the spring of 1980, something revealing happened. The real Cyrus Vance resigned suddenly. The effect on the program CYRUS was chaotic. It was utterly unable to cope with the flood of ''unusual'' news about Cyrus Vance. The only sorts of episodes CYRUS could understand at all were diplomatic meetings, flights, press conferences, state dinners, and the like—less than two dozen general sorts of activities (the kinds that are newsworthy and typical of secretaries of state). It had no provision for sudden resignation. It was as if the UPI had reported that a wicked witch had turned Vance into a frog. It is distinctly possible that CYRUS would have taken that report more in stride that the actual news. One can imagine the conversation:

Q: *Hello, Mr. Vance, what's new?*
A: I was turned into a frog yesterday.

But of course it wouldn't know enough about what it had just written to be puzzled, or startled, or embarrassed. The reason is obvious. When you look inside CYRUS, you find that it has skeletal definitions of thousands of words, but these definitions are minimal. They contain as little as the system designers think that they can get away with. Thus, perhaps, *lawyer* would be defined as synonymous with *attorney* and *legal counsel,* but aside from that, all one would discover about lawyers is that they are adult human beings and that they perform various functions in legal areas. If you then traced out the path to *human being,* you'd find out various obvious things CYRUS ''knew'' about human beings (hence about lawyers), but that is not a lot. That lawyers are university graduates, that they are better paid than chambermaids, that

they know how to tie their shoes, that they are unlikely to be found in the company of lumberjacks—these trivial, if weird, facts about lawyers would not be explicit or implicit anywhere in this system. In other words, a very thin stereotype of a lawyer would be incorporated into the system, so that almost nothing you could tell it about a lawyer would surprise it.

So long as surprising things don't happen, so long as Mr. Vance, for instance, leads a typical diplomat's life, attending state dinners, giving speeches, flying from Cairo to Rome, and so forth, this system works very well. But as soon as his path is crossed by an important anomaly, the system is unable to cope, and unable to recover without fairly massive human intervention. In the case of the sudden resignation, Kolodner and her associates soon had CYRUS up and running again, with a new talent—answering questions about Edmund Muskie, Vance's successor—but it was no less vulnerable to unexpected events. Not that it mattered particularly since CYRUS was a theoretical model, not a practical system.

There are a host of ways of improving the performance of such systems, and of course, some systems are much better than others. But all AI programs in one way or another have this facade-like quality, simply for reasons of economy. For instance, most expert systems in medical diagnosis so far developed operate with statistical information. They have no deep or even shallow knowledge of the underlying causal mechanisms of the phenomena that they are diagnosing. To take an imaginary example, an expert system asked to diagnose an abdominal pain would be oblivious to the potential import of the fact that the patient had recently been employed as a sparring partner by Muhammad Ali—there being no statistical data available to it on the rate of kidney stones among athlete's assistants. That's a fanciful case no doubt—too obvious, perhaps, to lead to an actual failure of diagnosis and practice. But more subtle and hard-to-detect limits to comprehension are always present, and even experts, even the system's designers, can be uncertain of where and how these limits will interfere with the desired operation of the system. Again, steps can be taken and are being taken to correct these flaws. For instance, my former colleague at Tufts, Benjamin Kuipers, is currently working on an expert system in nephrology—for diagnosing kidney ailments—that will be based on an elaborate system of causal reasoning about the phenomena being diagnosed. But this is a very ambitious, long-range project of considerable theoretical difficulty. And even if all the reasonable, cost-effective

steps are taken to minimize the superficiality of expert systems, they will still be facades, just somewhat thicker or wider facades.

When we were considering the fantastic case of the crazy Chamber of Commerce of Great Falls, Montana, we couldn't imagine a plausible motive for anyone going to any sort of trouble to trick the Dennett test. The quick-probe assumption for the Dennett test looked quite secure. But when we look at expert systems, we see that, however innocently, their designers do have motivation for doing exactly the sort of trick that would fool an unsuspicious Turing tester. First, since expert systems are all superspecialists who are only supposed to know about some narrow subject, users of such systems, not having much time to kill, do not bother probing them at the boundaries at all. They don't bother asking ''silly'' or irrelevant questions. Instead, they concentrate—not unreasonably—on exploiting the system's strengths. But shouldn't they try to obtain a clear vision of such a system's weaknesses as well? The normal habit of human thought when conversing with one another is to assume general comprehension, to assume rationality, to assume, moreover, that the quick-probe assumption is, in general, sound. This amiable habit of thought almost irresistibly leads to putting too much faith in computer systems, especially user-friendly systems that present themselves in a very anthropomorphic manner.

Part of the solution to this problem is to teach all users of computers, especially users of expert systems, how to probe their systems before they rely on them, how to search out and explore the boundaries of the facade. This is an exercise that calls not only for intelligence and imagination, but also a bit of special understanding about the limitations and actual structure of computer programs. It would help, of course, if we had standards of truth in advertising, in effect, for expert systems. For instance, each such system should come with a special demonstration routine that exhibits the sorts of shortcomings and failures that the designer knows the system to have. This would not be a substitute, however, for an attitude of cautious, almost obsessive, skepticism on the part of the users, for designers are often, if not always, unaware of the subtler flaws in the products they produce. That is inevitable and natural, given the way system designers must think. They are trained to think positively—constructively, one might say—about the designs that they are constructing.

I come, then, to my conclusions. First, a philosophical or theoretical conclusion: The Turing test in unadulterated, unrestricted from, as

Turing presented it, is plenty strong if well used. I am confident that no computer in the next twenty years is going to pass an unrestricted Turing test. They may well win the World Chess Championship or even a Nobel Prize in physics, but they won't pass the unrestricted Turing test. Nevertheless, it is not, I think, impossible in principle for a computer to pass the test, fair and square. I'm not running one of those a priori ''computers can't think'' arguments. I stand unabashedly ready, moreover, to declare that any computer that actually passes the unrestricted Turing test will be, in every theoretically interesting sense, a thinking thing.

But remembering how very strong the Turing test is, we must also recognize that there may also be interesting varieties of thinking or intelligence that are not well poised to play and win the imitation game. That no nonhuman Turing test winners are yet visible on the horizon does not mean that there aren't machines that already exhibit *some* of the important features of thought. About them, it is probably futile to ask my title question, Do they think? Do they *really* think? In some regards they do, and in some regards they don't. Only a detailed look at what they do, and how they are structured, will reveal what is interesting about them. The Turing test, not being a scientific test, is of scant help on that task, but there are plenty of other ways of examining such systems. Verdicts on their intelligence or capacity for thought or consciousness would be only as informative and persuasive as the theories of intelligence or thought or consciousness the verdicts are based on and since our task is to create such theories, we should get on with it and leave the Big Verdict for another occasion. In the meantime, should anyone want a surefire, almost-guaranteed-to-be-fail-safe test of thinking by a computer, the Turing test will do very nicely.

My second conclusion is more practical, and hence in one clear sense more important. Cheapened versions of the Turing test are everywhere in the air. Turing's test in not just effective, it is entirely natural—this is, after all, the way we assay the intelligence of each other every day. And since incautious use of such judgments and such tests is the norm, we are in some considerable danger of extrapolating too easily, and judging too generously, about the understanding of the systems we are using. The problem of overestimation of cognitive prowess, of comprehension, of intelligence, is not, then, just a philosophical problem, but a real social problem, and we should alert ourselves to it, and take steps to avert it.

Postscript [1985]: Eyes, Ears, Hands, and History

My philosophical conclusion in this paper is that any computer that actually passes the Turing test would be a thinking thing in every theoretically interesting sense. This conclusion seems to some people to fly in the face of what I have myself argued on other occasions. Peter Bieri, commenting on this paper at Boston University, noted that I have often claimed to show the importance to genuine understanding of a rich and intimate perceptual interconnection between an entity and its surrounding world—the need for something like eyes and ears—and a similarly complex active engagement with elements in that world—the need for something like hands with which to do things in that world. Moreover, I have often held that only a biography of sorts, a history of actual projects, learning experiences, and other bouts with reality, could produce the sorts of complexities (both external, or behavioral, and internal) that are needed to ground a principled interpretation of an entity as a thinking thing, an entity with beliefs, desires, intentions, and other mental attitudes.

But the opaque screen in the Turing test discounts or dismisses these factors altogether, it seems, by focusing attention on only the contemporaneous capacity to engage in one very limited sort of activity: verbal communication. (I have coined a pejorative label for such purely language-using systems: bedridden.) Am I going back on my earlier claims? Not at all. I am merely pointing out that the Turing test is so powerful that it will ensure indirectly that these conditions, if they are truly necessary, are met by any successful contestant.

''You may well be right,'' Turing could say, ''that eyes, ears, hands, and a history are necessary conditions for thinking. If so, then I submit that nothing could pass the Turing test that didn't have eyes, ears, hands, and a history. That is an empirical claim, which we can someday

hope to test. If you suggest that these are conceptually necessary, not just practically or physically necessary, conditions for thinking, you make a philosophical claim that I for one would not know how, or care, to assess. Isn't it more interesting and important in the end to discover whether or not it is true that no bedridden system could pass a demanding Turing test?''

Suppose we put to Turing the suggestion that he add another component to his test: Not only must an entity win the imitation game, but also must be able to identify—using whatever sensory apparatus it has available to it—a variety of familiar objects placed in its room: a tennis racket, a potted palm, a bucket of yellow paint, a live dog. This would ensure that somehow the other entity was capable of moving around and distinguishing things in the world. Turing could reply, I am asserting, that this is an utterly unnecessary addition to his test, making it no more demanding than it already was. A suitable probing conversation would surely establish, beyond a shadow of a doubt, that the contestant knew its way around the world. The imagined alternative of somehow ''prestocking'' a bedridden, blind computer with enough information, and a clever enough program, to trick the Turing test is science fiction of the worst kind—possible ''in principle'' but not remotely possible in fact, given the combinatorial explosion of possible variation such a system would have to cope with.

''But suppose you're wrong. What would you say of an entity that was created all at once (by some programmers, perhaps), an instant individual with all the conversational talents of an embodied, experienced human being?'' This is like the question: ''Would you call a hunk of H_2O that was as hard as steel at room temperature ice?'' I do not know what Turing would say, of course, so I will speak for myself. Faced with such an improbable violation of what I take to be the laws of nature, I would probably be speechless. The least of my worries would be about which lexicographical leap to take:

A: ''It turns out, to my amazement, that something can think without having had the benefit of eyes, ears, hands, and a history.''
B: ''It turns out, to my amazement, that something can pass the Turing test without thinking.''

Choosing between these ways of expressing my astonishment would be asking myself a question ''too meaningless to deserve discussion.''

Discussion

Q: *Why was Turing interested in differentiating a man from a woman in his famous test?*

A: That was just an example. He described a parlor game in which a man would try to fool the judge by answering questions as a woman would answer. I suppose that Turing was playing on the idea that maybe, just maybe, there is a big difference between the way men think and the way women think. But of course they're both thinkers. He wanted to use that fact to make us realize that, even if there were clear differences between the way a computer and a person thought, they'd both still be thinking.

Q: *Why does it seem that some people are upset by AI research? Does AI research threaten our self-esteem?*

A: I think Herb Simon has already given the canniest diagnosis of that. For many people the mind is the last refuge of mystery against the encroaching spread of science, and they don't like the idea of science engulfing the last bit of *terra incognita.* This means that they are threatened, I think irrationally, by the prospect that researchers in Artificial Intelligence may come to understand the human mind as well as biologists understand the genetic code, or as well as physicists understand electricity and magnetism. This could lead to the ''evil scientist'' (to take a stock character from science fiction) who can control you because he or she has a deep understanding of what's going on in your mind. This seems to me to be a totally valueless fear, one that you can set aside, for the simple reason that the human mind is full of an extraordinary amount of detailed knowledge, as, for example, Roger Schank has been pointing out.

As long as the scientist who is attempting to manipulate you does not share all your knowledge, his or her chances of manipulating you are minimal. People can always hit you over the head. They can do that now. We don't need Artificial Intelligence to manipulate people by putting them in chains or torturing them. But if someone tries to manipulate you by controlling your thoughts and ideas, that person will have to know what you know and more. The best way to keep yourself safe from that kind of manipulation is to be well informed.

Q: *Do you think we will be able to program self-consciousness into a computer?*

A: Yes, I do think that it's possible to program self-consciousness into a computer. *Self-consciousness* can mean many things. If you take the

simplest, crudest notion of self-consciousness, I suppose that would be the sort of self-consciousness that a lobster has: When it's hungry, it eats something, but it never eats itself. It has some way of distinguishing between itself and the rest of the world, and it has a rather special regard for itself.

The lowly lobster is, in one regard, self-conscious. If you want to know whether or not you can create that on the computer, the answer is yes. It's no trouble at all. The computer is already a self-watching, self-monitoring sort of thing. That is an established part of the technology.

But, of course, most people have something more in mind when they speak of self-consciousness. It is that special inner light, that private way that it is with you that nobody else can share, something that is forever outside the bounds of computer science. How could a computer ever be conscious in this sense?

That belief, that very gripping, powerful intuition is, I think, in the end simply an illusion of common sense. It is as gripping as the common-sense illusion that the earth stands still and the sun goes around the earth. But the only way that those of us who do not believe in the illusion will ever convince the general public that it *is* an illusion is by gradually unfolding a very difficult and fascinating story about just what is going on in our minds.

In the interim, people like me—philosophers who have to live by our wits and tell a lot of stories—use what I call intuition pumps, little examples that help free up the imagination. I simply want to draw your attention to one fact. If you look at a computer—I don't care whether it's a giant Cray or a personal computer—if you open up the box and look inside and see those chips, you say, ''No way could that be conscious. No way could that be self-conscious.'' But the same thing is true if you take the top off somebody's skull and look at the gray matter pulsing away in there. You think, ''That is conscious? No way could that lump of stuff be conscious.''

Of course, it makes no difference whether you look at it with a microscope or with a macroscope: At no level of inspection does a brain look like the seat of consciousness. Therefore, don't expect a computer to look like the seat of consciousness. If you want to get a grasp of how a computer could be conscious, it's no more difficult in the end than getting a grasp of how a brain could be conscious.

As we develop good accounts of consciousness, it will no longer seem so obvious to everyone that the idea of a self-conscious computer