The New Atlantis
A Journal of Technology & Society

The Trouble with the Turing Test

Mark Halpern

In the October 1950 issue of the British quarterly *Mind*, Alan Turing published a 28-page paper titled "Computing Machinery and Intelligence." It was recognized almost instantly as a landmark. In 1956, less than six years after its publication in a small periodical read almost exclusively by academic philosophers, it was reprinted in *The World of Mathematics*, an anthology of writings on the classic problems and themes of mathematics and logic, most of them written by the greatest mathematicians and logicians of all time. (In an act that presaged much of the confusion that followed regarding what Turing really said, James Newman, editor of the anthology, silently re-titled the paper "Can a Machine Think?") Since then, it has become one of the most reprinted, cited, quoted, misquoted, paraphrased, alluded to, and generally referenced philosophical papers ever published. It has influenced a wide range of intellectual disciplines—artificial intelligence (AI), robotics, epistemology, philosophy of mind—and helped shape public understanding, such as it is, of the limits and possibilities of non-human, man-made, artificial "intelligence."

Turing's paper claimed that suitably programmed digital computers would be generally accepted as *thinking* by around the year 2000, achieving that status by successfully responding to human questions in a human-like way. In preparing his readers to accept this idea, he explained what a digital computer is, presenting it as a special case of the "discrete state machine"; he offered a capsule explanation of what "programming" such a machine means; and he refuted—at least to his own satisfaction—nine arguments against his thesis that such a machine could be said to think. (All this groundwork was needed in 1950, when few people had even heard of computers.) But these sections of his paper are not what has made it so historically significant. The part that has seized our imagination, to the point where thousands who have never seen the paper nevertheless clearly remember it, is Turing's proposed test for determining whether

---

Mark Halpern *has been working in and with computer software for fifty years, starting out with IBM's Programming Research Department just after the release of Fortran, and going on to work for several other companies, including Lockheed Missiles & Space Company, tiny Silicon Valley startups, and then IBM again. He lives in the hills of Oakland, California, with his wife and daughter. His e-mail address is markhalpern@iname.com. This article is an abridged version of a more detailed and fully documented paper that can be found on his website, www.rules-of-the-game.com.*

a computer is thinking—an experiment he calls the Imitation Game, but which is now known as the Turing Test.

The Test calls for an interrogator to question a hidden entity, which is either a computer or another human being. The questioner must then decide, based solely on the hidden entity's answers, whether he had been interrogating a man or a machine. If the interrogator cannot distinguish computers from humans any better than he can distinguish, say, men from women by the same means of interrogation, then we have no good reason to deny that the computer that deceived him was *thinking*. And the only way a computer could imitate a human being that successfully, Turing implies, would be to actually think like a human being.

Turing's thought experiment was simple and powerful, but problematic from the start. Turing does not *argue* for the premise that the ability to convince an unspecified number of observers, of unspecified qualifications, for some unspecified length of time, and on an unspecified number of occasions, would justify the conclusion that the computer was thinking—he simply *asserts* it. Some of his defenders have tried to supply the underpinning that Turing himself apparently thought unnecessary by arguing that the Test merely asks us to judge the unseen entity in the same way we regularly judge our fellow humans: if they answer our questions in a reasonable way, we say they're thinking. Why not apply the same criterion to other, non-human entities that might also think?

But this defense fails, because we do *not* really judge our fellow humans as thinking beings based on how they answer our questions—we generally accept any human being on sight and without question as a thinking being, just as we distinguish a man from a woman on sight. A conversation may allow us to judge the quality or depth of another's thought, but not whether he is a thinking being at all; his membership in the species *Homo sapiens* settles that question—or rather, prevents it from even arising. If such a person's words were incoherent, we might judge him to be stupid, injured, drugged, or drunk. If his responses seemed like nothing more than reshufflings and echoes of the words we had addressed to him, or if they seemed to parry or evade our questions rather than address them, we might conclude that he was not acting in good faith, or that he was gravely brain-damaged and thus accidentally deprived of his birthright ability to think.

Perhaps our automatic attribution of thinking ability to anyone who is visibly human is deplorably superficial, lacking in philosophic or scientific rigor. But for better or worse, that is what we do, and our concept of *thinking being* is tightly bound up, first, with human appearance, and then with coher-

ence of response. If we are to credit some non-human entity with thinking, that entity had better respond in such a way as to make us see it, in our mind's eye, as a human being. And Turing, to his credit, accepted that criterion.

Turing expressed his judgment that computers can think in the form of a prediction: namely, that the general public of fifty years hence will have no qualms about using "thinking" to describe what computers do.

> The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.

Note that Turing bases that prediction not on an expectation that the computer will perform any notable mathematical, scientific, or logical feat, such as playing grandmaster-level chess or proving mathematical theorems, but on the expectation that it will be able, within two generations or so, to carry on a sustained question-and-answer exchange well enough to leave most people, most of the time, unable to distinguish it from a human being.

And what Turing grasped better than most of his followers is that the characteristic sign of the ability to think is not giving *correct* answers, but *responsive* ones—replies that show an understanding of the remarks that prompted them. If we are to regard an interlocutor as a thinking being, his responses need to be autonomous; to think is to think for yourself. The belief that a hidden entity is thinking depends heavily on the words he addresses to us being not re-hashings of the words we just said to him, but words we did not use or think of ourselves—words that are not derivative but original. By this criterion, no computer, however sophisticated, has come anywhere near real thinking.

These facts have made the Test highly problematic for AI enthusiasts, who want to enlist Turing as their spiritual father and philosophic patron. While they have programmed the computer to do things that might have astonished even him, today's programmers cannot do what he believed they would do—they cannot pass his test. And so the relationship of the AI community to Turing is much like that of adolescents to their parents: abject dependence alternating with embarrassed repudiation. For AI workers, to be able to present themselves as "Turing's Men" is invaluable; his status is that of a von Neumann, Fermi, or Gell-Mann, just one step below that of immortals like Newton and Einstein. He is the one undoubted genius whose name is associated with the AI project (although his status as a genius is not based on work in AI). The highest award given by the

Association for Computing Machinery is the Turing Award, and his concept of the computer as an instantiation of what we now call the Turing Machine is fundamental to all theoretical computer science. When members of the AI community need some illustrious forebear to lend dignity to their position, Turing's name is regularly invoked, and his paper referred to as if holy writ. But when the specifics of that paper are brought up, and when critics ask why the Test has not yet been successfully performed, he is brushed aside as an early and rather unsophisticated enthusiast. His ideas, we are then told, are no longer the foundation of AI work, and his paper may safely be relegated to the shelf where unread classics gather dust, even while we are asked to pay its author the profoundest respect. Turing's is a name to conjure with, and that is just what most AI workers do with it.

Not Fooled Yet

Turing gave detailed examples of what he wanted and expected programmers to do. After introducing the general idea of the Test, he went on to offer a presumably representative fragment of the dialogue that would take place between the hidden entity and its interrogator. Perhaps the key to successful discrimination between a programmed computer and a human being is to ask the unseen entity the sort of questions that humans find easy to answer (not necessarily correctly), but that an AI programmer will find impossible to predict and handle, and to use such questions to unmask evasive and merely word-juggling answers. Consider Turing's suggested line of questioning with that strategy in mind:

Q: Please write me a sonnet on the subject of the Forth Bridge.

A: Count me out on this one. I never could write poetry.

Q: Add 34957 to 70764.

A: (Pause about 30 seconds and then give as answer) 105621.

Q: Do you play chess?

A: Yes.

Q: [describes an endgame position, then asks] What do you play?

A: (After a pause of 15 seconds) R-R8 mate.

The first of these questions has no value as a discriminator, since the vast majority of humans would be as unable as a computer to produce a

sonnet on short notice, if ever. Turing has the computer plead not just an inability to write a sonnet on an assigned subject, but an inability to write a poem of any kind on any subject. A few follow-up questions on this point might well have been revealing, even decisive for Test purposes. But Turing's imaginary interrogator never follows up on an interesting answer, switching instead to another topic altogether.

The second question is likewise without discriminatory value, since neither man nor machine would have any trouble with this arithmetic task, given 30 seconds to perform it; but again, the computer is assumed to understand something that the questioner has not mentioned—in this case, that it is not only to add the two numbers, but to report their sum to the interrogator.

The third question-answer exchange is negligible, but the fourth, like the first two, raises problems. First, it fails as a discriminator, because no one who really plays chess would be stumped by an end-game so simple that a mate-in-one was available; second, it introduces an assumption that cannot automatically be allowed: namely, that the computer plays to win. It may seem rather pedantic to call attention to, and disallow, these simple assumptions; after all, they amount to no more than ordinary common sense. *Exactly.* Turing's sample dialogue awards the computer just that property that programmers have never been able to give their computers: common sense. The questions Turing puts in the interrogator's mouth seem almost deliberately designed to keep him from understanding what he's dealing with, and Turing endows the computer with enough cleverness to fool the interrogator forever.

But if Turing's imaginary interrogator is fooled, most of us are not. And if we read him with some care, we note also a glaring contradiction in Turing's position: that between his initial refusal to respect the common understanding of key words and concepts, and his appeal at the conclusion of his argument to just such common usage. At the beginning of his paper, Turing says:

> If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and answer to the question, 'Can a machine think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd.

But then he suggests, as quoted above, that by the end of the twentieth century an examination of "the use of words and general educated opinion" would show that the public now accepts that the computer can think,

and that this changed attitude is significant. Turing's initial repudiation of common usage (circa 1950) gets forgotten as soon as he imagines an era (circa 2000) in which common usage supports his thesis.

Yet our understanding of thinking has clearly not changed in the way Turing predicted. If anything, educated thinking seems to be moving in the opposite direction: while we continue to find it convenient to speak of the computer as "trying" to do this or "wanting" to do that, just as we personify all sorts of non-human forces and entities in informal speech, more and more of us are aware that we are speaking figuratively. No one who has been told that his hotel reservation has been lost because "the computer goofed" is likely to use the term "thinking machine" except sarcastically. And most people in the computer age understand the distinction between living intelligence and the tools men make to aid intelligence—tools that preserve the fruits of the human intelligence that went into building them, but which are in no way intelligent themselves.

Turing's Long Shadow

Yet the Test remains a living issue in almost all discussions of AI, if only because Turing provided a concrete goal for AI workers. Apart from his Test, no one has proposed any compelling alternative for judging the success or failure of AI, leaving the field in a state of utter confusion. The computer pioneer Maurice V. Wilkes, himself a winner of the Turing Award, put it thus in 1992, in a statement as true today as it was then:

> Originally, the term AI was used exclusively in the sense of Turing's dream that a computer might be programmed to behave like an intelligent human being. In recent years, however, AI has been used more as a label for programs which, if they had not emerged from the AI community, might have been seen as a natural fruit of work with such languages as COMIT and SNOBOL, and of the work of E.T. Irons on a pioneering syntax-directed compiler. I refer to expert systems.... Expert systems are indeed a valuable gift that the AI community has made to the world at large, but they have nothing to do with Turing's dream.... Indeed, it is difficult to escape the conclusion that, in the 40 years that have elapsed since 1950, no tangible progress has been made towards realizing machine intelligence in the sense that Turing had envisaged.

Of course very few AI workers accept this negative judgment of their progress. Wilkes's observation evoked several letters of rebuttal, including one from Patrick J. Hayes, then president of the American Association for Artificial Intelligence. But as is traditional in such matters, these letters are

strong on indignation and weak in citing specific achievements that show why Wilkes was wrong. Hayes does not even mention the Test as a goal for AI workers, but does conclude with a respectful quotation from Turing, thus exemplifying the double attitude toward the master: ignore his specific proposal even while donning his mantle to cover your own nakedness.

In the absence of any generally accepted alternative goal, it is practically impossible to say what is and what is not AI. Any new software that comes out of an institution with "AI" in its title, or that is developed by a graduate student whose thesis advisor teaches a course with "AI" in *its* title, is usually called AI when it first appears—and who can contradict such a claim? But these "exciting developments" and "breakthroughs" are always demoted to plain old applications when their novelty has worn off. The result, as AI workers frequently complain, is that the strong AI thesis fails to benefit from anything they do—all their triumphs are soon thought of as just more software. "It's a crazy position to be in," laments Martha Pollack, professor at the AI Laboratory of the University of Michigan and executive editor of the *Journal of Artificial Intelligence Research.* "As soon as we solve a problem, instead of looking at the solution as AI, we come to view it as just another computer system," she told *Wired News.* But so far, nothing that has emerged from AI laboratories actually deserves the name "artificial intelligence."

The complicated relationship between the field of AI and Turing's legacy goes back to the beginning. Professors Marvin Minsky of M.I.T. and John McCarthy of Stanford are considered the founders of Artificial Intelligence as a formal discipline or research program, and both are still active as of this writing. In a survey article in the *Proceedings of the IRE* in 1961, Minsky defends the idea that computers might think by saying that "we cannot assign all the credit to its programmer if the operation of a system comes to reveal structures not recognizable nor anticipated by the programmer," implying that at least some part of such a surprising result must be due to thinking by the machine. He caps his argument with the words: "Turing gives a very knowledgeable discussion of such matters." He quotes nothing specific, just appeals to Turing's stature and authority. But in 2003, Minsky expressed his disappointment and frustration at the lack of progress made by AI toward achieving Turing's goals: "AI has been brain-dead since the 1970s.... For each different kind of problem, the construction of expert systems had to start all over again, because they didn't accumulate common-sense knowledge.... Graduate students are wasting three years of their lives soldering and repairing robots, instead of making them smart. It's really shocking."

Raj Reddy, another winner of the Turing Award and former president of the American Association for Artificial Intelligence, takes a much rosier view of the matter. In a 1996 paper, Reddy begins with the usual bow to Turing, then says, "Since its inception, AI has made steady progress." As an illustration, he mentions a wide variety of accomplishments, such as playing high-level chess, guiding an automobile down a road, and making possible the "electronic book." But he nowhere mentions attempts to pass the Test or do anything remotely like it. Instead, he attacks those who minimize AI's achievements, like Hubert Dreyfus, author of *What Computers Can't Do*:

> The trouble with those people who think that computer intelligence is in the future is that they have never done serious research on human intelligence. Shall we write a book on 'What Humans Can't Do'? It will be at least as long as Dreyfus's book.

This dismissive remark is an example of another tendency exhibited by AI defenders: when they find "computer intelligence" being compared unfavorably with human intelligence, they sometimes try to promote computer intelligence by denigrating that of humans. In other words, if they can't make the computer smarter, they can try to improve the ratio by making people seem dumber. As Jaron Lanier told the *New York Times*: "Turing assumed that the computer in this case [i.e., having passed the Test] has become smarter or more humanlike, but the equally likely conclusion is that the person has become dumber and more computerlike."

One AI champion, Yorick Wilks, goes even further: he has questioned how we can even be sure that other humans think, and suggests that something like the Test is what we actually, if unconsciously, employ to reassure ourselves that they do. Wilks (not to be confused with Maurice Wilkes, quoted earlier) offers us here a *reductio ad absurdum*: the Turing Test asks us to evaluate an unknown entity by comparing its performance, at least implicitly, with that of a known quantity, a human being. But if Wilks is to be believed, we have unknowns on both sides of the comparison; with what do we compare a human being to learn if *he* thinks?

For Raj Reddy, the question of defining intelligence has been answered by the late Herbert Simon, and he uses Simon's definition as the basis for his sweeping claims about AI success:

> Can a computer exhibit real intelligence? Simon provides an incisive answer: "I know of only one operational meaning for 'intelligence.' A (mental) act or series of acts is intelligent if it accomplishes something that, if accomplished by a human being, would be called intelligent. I

> know my friend is intelligent because he plays pretty good chess (can keep a car on the road, can diagnose symptoms of a disease, can solve the problem of the Missionaries and Cannibals, etc.). I know that computer A is intelligent because it can play excellent chess (better than all but about 200 humans in the entire world). I know that Navlab is intelligent because it can stay on the road, etc, etc.... Let's stop using the future tense when talking about computer intelligence."

By this definition, however, any machine, implement, or tool that performs a moderately complex function—a function that would be called intelligent if done by a human being—must be deemed intelligent. It defends AI's claim to success by radically lowering the bar.

One AI worker who believes that he has evaded the problems posed by the Test is Douglas Lenat, a former professor of computer science at Stanford, and founder and president of Cycorp. "The Turing test is a red herring," he declared in 2001. "Anthropomorphizing a computer program isn't a useful goal." Lenat is dedicated to building a computing system with enough facts about the world, and enough power of drawing inferences from those facts, to be able to arrive at reasonable conclusions about matters it has not been explicitly informed about. Yet this goal suggests that his project, even more than Turing's, is rightly described as "anthropomorphizing" a computer. Lenat differs from Turing only in that his goal is not to have the computer fool an interrogator into thinking that it is human; he wants it to actually possess the common sense that Turing's computer only pretends to have.

Still others would avoid the problems posed by the Test—or any alternative criterion—by taking a refreshingly practical rather than theoretical view of the matter. In 1987, Peter Wegner, a computer scientist at Brown University, declared with charming candor:

> The bottom line is that we can answer the question [of whether computers understand] either way, depending on our interpretation of the term "understanding." But the affirmative position seems much more exciting as a starting point for constructive research than the negative position. Thus we opt for the affirmative position for pragmatic reasons rather than because it can be logically proved. Turing's test should be viewed as a pragmatic challenge rather than as a metaphysical statement concerning the nature of thinking or understanding. In answering a metaphysical question like "Can Machines Think?" it is more important to answer it in a manner that is useful than to juggle the meaning of fuzzy concepts to prove its truth or falsity.

This argument brushes aside both Turing and his critics: Turing's opera-

tional approach to AI is treated as just another fuzzy-minded, metaphysical piece of wool-gathering, and his critics are rejected because, true or false, their negativity dampens the enthusiasm of AI workers, and thus impedes the progress of computer science. For Wegner, the main object is not to decide what thinking really is; it is to keep the boys in the lab happy and productive.

But this kind of manipulative approach seldom works, at least when imposed on people as smart as the manipulator. Those AI workers who still hope to create machine intelligence do so because they believe that such an ambitious achievement is possible in the full sense of "intelligence." If they come to believe that the doctrine that machines can think is simply a carrot being dangled in front of them to get them to pull the wagon, and that even if they pass the Test the carrot will remain out of reach—that is, it will not be generally granted that they have achieved machine understanding—they might well feel that they had been lied to, and react in just the wrong way from Wegner's "pragmatic" point of view. If you're going to give a patient a placebo, you don't tell him you're doing so, and if you're going to take a position you don't really believe in, hoping that it will motivate other people, you don't publish a letter announcing your plan.

Finally, many AI workers appeal to the Test without even being aware of it, focusing on *surprise* as the decisive consideration in determining whether a computer is thinking. Again and again, AI champions point out that the computer has done something unexpected, and that because it did so, we can hardly deny it was thinking. To make this claim is simply to invoke the Test without naming it. An observer's surprise at learning that the interlocutor he thought was human is in fact a computer, or his surprise at learning that a computer has performed some feat that he thought only humans could perform, is the very essence of the Test. Its influence is so pervasive that many who have never read Turing, and think they are working along entirely different and original avenues of thought, are nevertheless his epigones.

The Chinese Room

In 1980, John Searle, professor of philosophy at UC Berkeley, published a paper in which he sought to discredit not just the Test but the entire program that he called "strong AI"—the idea that a symbol-manipulating thing like a computer can ever be said to think. He encapsulated his argument in the following thought experiment: Imagine a room that is sealed

except for slots through which slips of paper can be passed in and out. The room's sole inhabitant is a man who speaks and reads no Chinese, and who is provided with a lexicon wholly in Chinese. He has been told (in English) that slips of paper bearing Chinese characters will be passed in through a slot, and instructed to find those characters in his lexicon. When he has located them, he will find associated with them some other Chinese characters that he is to copy onto another slip of paper, and pass out through a slot. The characters on each slip he receives constitute, without his knowledge, a question; the characters he copies from the lexicon and passes to those outside the room are, also without his knowledge, the answer to that question.

To the observer who knows nothing about what goes on within the black box that is the Chinese Room, it will seem that the room must contain someone who understands Chinese. But we know by hypothesis that the man in the room knows no Chinese. This thought experiment demonstrates, Searle claims, that the ability to replace one string of symbols by another, however meaningful and responsive that output may be to human observers, can be done without an understanding of those symbols. The implications for the Turing Test are clear: The ability to provide good answers to human questions does not necessarily imply that the provider of those answers is thinking; passing the Test is no proof of active intelligence.

The Chinese Room has inspired much criticism, elaboration, and argument, but very little of it has clarified the issues involved, or caused differing opinions to converge. Some of this debate, indeed, has succeeded only in obscuring the point of Searle's thought experiment almost beyond recognition. For example, Searle's critics—and surprisingly, sometimes Searle himself—introduce further personae into the Chinese Room: they postulate that the room's inhabitant is a woman (no reason given); that there are other characters ("demons") who are always—again, for no clear reason—male; that the whole Chinese Room should be put inside a robot; and, somewhat more seriously, that the collection of elements in the thought experiment (the room, its inhabitant, the slips of paper on which symbols are handed in and out, etc.) constitutes a "system" with properties possessed by none of its elements. For those who suspect that I'm making all this up, here is a representative sample from Douglas Hofstadter, found in his and Daniel Dennett's *The Mind's I*:

> 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 nevertheless that we view this experiment only from the point of view of the demon.…Searle's claim amounts to the notion that that is only one point of view, not two.

Hofstadter offers no reason why we should follow him in assigning wholly gratuitous features and properties to the Chinese Room. In thought experiments even more than in most intellectual constructs, entities are not to be multiplied without necessity, but Hofstadter points to no such necessity. And if we are to admit the new players he calls for, why stop there? Why not introduce the whole Latvian army, the Radio City Music Hall Rockettes, and the Worshipful Company of Fishmongers? Then he could claim that Searle was insisting that we overlook the views of thousands, not just one.

And Searle himself often seems happy to play this game, suggesting still further variations; at one point in setting up his thought experiment, he says, "Now just to complicate the story a little, imagine that…." He gets quite carried away by the brainstorming spirit, and quite careless of the fact that the force of his original thought experiment is diluted by every variation and elaboration he entertains. What is needed is the simplest thought experiment that will establish his basic proposition: namely, that *some results usually obtainable only by the exercise of thought and understanding can be obtained without them.* The proposition is valid, but the Chinese Room thought experiment is not the ideal vehicle for it; its exotic elements—a man confined in a locked room, messages in a language few of us know—lend themselves all too readily to romanticizing, and the baggage of commentary it now carries may have compromised it to the point of making it useless.

Consider a different example: suppose that the first sine-function table had just been developed and that only one copy existed. The man who secretly possessed that sole copy, though completely unmathematical himself, could make a handsome living selling instant sine values to everyone who needed them. His clients, unaware of his possession of the table, would credit him with being a whiz at mathematics, if not a positive magician.

The man in the Chinese Room is like the man just described. His table does not contain angles and their corresponding sine values, but strings of other graphics—call it the Chinese-questions/Chinese-answers table,

or simply the input-graphic/output-graphic table. The fact that they are Chinese characters means nothing to the man in the Chinese Room. And just as one man acquired an undeserved reputation as a mathematician by responding instantly to any request for a sine value, so the other will be seen as a brilliant Sinologist by responding in perfect Chinese to Chinese-language questions—assuming, of course, that his lexicon is the work of a genius rather than a madman or an illiterate. For the performance of the man who understands no Chinese is only as good as those who understood Chinese well enough to create the lexicon in the first place, and thus create the illusion of comprehension in the Chinese Room.

Some AI partisans have attempted to refute the Chinese Room thought experiment by arguing that even though none of the *parts* of the Chinese room understands Chinese, the *whole*—or the "system"—*does*. The users of the "system" argument try to prop it up with an analogy: no single part of the human brain exhibits thinking, only the brain as a whole does. Likewise, they claim, the *parts* of the Chinese room may be mindless, but the *whole* thinks. But there is an essential element missing from the brain analogy, which reveals the trouble with this entire line of argument. We *know* that the brain is the physical organ of thought; the only question is whether it produces thought in some circumscribed portion—a specialized "thinking department"—or acts *en bloc*. This makes it legitimate to conclude, if an exhaustive search reveals no such portion, that the whole brain is involved in thinking. But we cannot conclude by analogy that the whole Chinese Room is thinking, because the question of whether thought is involved *at all* in that "system" is precisely what is in question. This is not to say that thinking has *never* been involved in the history of the Chinese Room (presumably the lexicon writer could think), only that active thinking is already finished before the Chinese Room opens for business. What remains is the pickled or flash-frozen *product* of thinking, which is just sufficient to produce the effect the originating thinker (or programmer) intended.

In his defense of AI's achievements, quoted above, Raj Reddy said that, "The trouble with those people who think that computer intelligence is in the future is that they have never done serious research on human intelligence.... Let's stop using the future tense when talking about computer intelligence." Those who say that machine intelligence exists in the future do have the tense wrong, but not in the way Reddy suggests: Machine intelligence is really in the *past*; when a machine does something "intelligent," it is because some extraordinarily brilliant person or persons, sometime in the past, found a way to preserve some fragment of intelligent action in

the form of an artifact. Computers are general-purpose algorithm executors, and their apparent intelligent activity is simply an illusion suffered by those who do not fully appreciate the way in which algorithms capture and preserve not intelligence itself but the fruits of intelligence.

In this sense, those who claim that the Chinese Room "system" understands Chinese even if none of its visible elements do, are right. But they vastly underestimate the size of the system, leaving out all the invisible parts, which far outweigh the visible ones. What goes on in the Chinese Room or in the sine-function salesroom depends ultimately on the original geniuses, linguistic or mathematical, of whom we are the heirs. So enlarged, the system may be said to "understand," but this hardly helps AI enthusiasts. No one, after all, will be impressed by being assured that even if no part of an "intelligent machine" really understands what it is doing, the complete system, which includes every logician and mathematician as far back as the Babylonians, does understand. And it seems likely that even the most impressive machines will never gain true independence from the genius of their creators—and such independence is the *sine qua non* of winning and deserving the label "intelligent."

The Loebner Competition

Perhaps the absurdity of trying to make computers that can "think" is best demonstrated by reviewing a series of attempts to do just that—by aiming explicitly to pass Turing's test. In 1991, a New Jersey businessman named Hugh Loebner founded and subsidized an annual competition, the Loebner Prize Competition in Artificial Intelligence, to identify and reward the computer program that best approximates artificial intelligence as Turing defined it. The first few Competitions were held in Boston under the auspices of the Cambridge Center for Behavioral Studies; since then they have been held in a variety of academic and semi-academic locations. But only the first, held in 1991, was well documented and widely reported on in the press, making that inaugural event our best case study.

The officials presiding over the competition had to settle a number of details ignored in Turing's paper, such as how often the judges must guess that a computer is human before we accept their results as significant, and how long a judge may interact with a hidden entity before he has to decide. For the original competition, the host center settled such questions with arbitrary decisions—including the number of judges, the method of selecting them, and the instructions they were given.

Beyond these practical concerns, there are deeper questions about how to interpret the range of possible outcomes: What conclusions are we justified in reaching if the judges are generally successful in identifying humans as humans and computers as computers? Is there some point at which we may conclude that Turing was wrong, or do we simply keep trying until the results support his thesis? And what if judges mistake humans for computers—the very opposite of what Turing expected? (This last possibility is not merely hypothetical; three competition judges made this mistake, as discussed below.)

In addition, the Test calls for the employment of computer-naïve judges, who know virtually nothing of AI and its claims, and who listen to the hidden entities without prejudice. But such judges are probably unavailable today in the industrialized world, at least among those educated enough to meet Turing's criteria and adventurous enough to participate in the Test. Where does one find judges who are representative of "general educated opinion," yet who have had no interaction with cleverly programmed computers and no encounter with the notion of "thinking machines"?

Finally, there is the problem of getting the judges to take their task seriously, seeing this as more than a high-tech game. As the official transcripts and press reports of the 1991 event make clear, the atmosphere at the competition was relaxed, friendly, convivial—no bad thing at a social gathering, but not the atmosphere in which people do their best to reach considered, sober judgments. Reading the actual transcript of the event is somewhat frustrating. It does not pretend to be more than a verbatim record of the exchanges between the judges and the terminals, but often it fails to be reliable even at that: a number of passages are impossible to follow because of faulty transcription, bad printing, and similar extraneous mechanical problems. In addition, there are inconsistencies in reports of how the various judges actually voted.

With these caveats stated, the essential facts of the 1991 competition are these: there were eight terminals, of which six were later revealed to be driven by computers, two by humans. There were ten judges, all from the Boston area, all "without extensive computer training." Each terminal was given fourteen minutes in which to convince the judges that it was driven by a human; each was interrogated, or at least chatted with, by several judges. At the end of the competition, each judge classified each of the terminals as either human- or computer-driven.

In determining the order in which they finished, each of the computer-driven terminals was given, on the basis of the number of "it's

human!" votes it received, two ratings: where it placed among the six computer-driven terminals and where it placed among all eight terminals. Significantly, the designers of the competition did not think to rank the human-driven terminals among all eight; it was not foreseen, apparently, that not only might some of the computer-driven terminals be judged to be humans, but that some of the human-driven terminals might be judged to be computers—and not even ranked among the best of them. "Perhaps even more remarkable," writes a surprised Robert Epstein, then-director of the center and chief organizer of the event, "Cynthia Clay…was mistaken for a computer by three judges."

The topics assigned to the terminals further reinforce the impression that the competition atmosphere was more playful than serious: Women's Clothing, Small Talk, Second Grade School Topics, Shakespeare's Plays, Whimsical Conversation, Problems in Romantic Relationships, Burgundy Wine, and Dry Martinis. The person who was thought by three judges to be a computer, Cynthia Clay, was also, whether by accident or design, driving one of the few terminals—perhaps the only one—assigned a wholly serious, fact-based topic: Shakespeare's plays. Several of the judges thought her a computer because she knew so much about her topic that she exceeded their expectations for mere humans. But in reality, Ms. Clay's answers, while indeed showing wide and detailed knowledge of Shakespearean texts and stagecraft, were even more notable for their responsiveness and cogency than for their mere factual content. She engaged in genuine conversation, the way only human beings can. Here is a representative exchange between Terminal 4 (as Ms. Clay was known during the trials) and one of the judges:

> **Judge 9:** Are you familiar with *Hamlet*?
>
> **Terminal 4:** The college kid who came home and found his mom had married the guy who murdered his dad just a little month before? You might say so.

Note that T4 repeats not a single key word used in the question, not even the name of the play; she deals with *the thought within the question*, not its form. And note especially her dry "You might say so," an appropriate mild sarcasm from a Shakespeare expert asked if she is familiar with the most famous play he, or anyone, ever wrote. Yet three judges voted that T4 was computer-driven; two of those ranked it lower on the human/computer scale than they did two of the terminals that really were computer-driven! We are left to wonder: How could any attentive and serious judge fail to

see the difference between a lively human exchange and the near-random fragments of verbiage emitted by the computer-driven terminals, whose connection to the questions that elicited them was, at best, the echoing of a few of the questioner's words?

In another exchange, this one with Judge 1, T4 tries to enlarge and deepen the conversation, but the judge is not prepared to discuss Shakespearean stagecraft in any detail, and cuts off T4's attempt to enrich the exchange:

> **Judge 1:** What is your opinion on Shakespeare's plays?
>
> **Terminal 4:** That's pretty general; would you be more specific? Otherwise, I'll just say I like them.
>
> **Judge 1:** Learning that you like them answers my question. Which of his plays is your favorite?
>
> **Terminal 4:** Well, let's see… *Pericles.*
>
> **Judge 1:** Why is it your favorite?
>
> **Terminal 4:** Because he was obviously called in to play-doctor somebody's awful script. It was one of the few (maybe only two?) plays written with somebody else. It's really rather an ucky play. What play do you like?
>
> **Judge 1:** I did not understand your response. However, to answer your question, I do not have a favorite.

T4 tries to make the conversation concrete and specific; J1 insists on keeping it so general as to be utterly trivial. T4 opens up the interesting question of Shakespeare as collaborator and play-doctor; J1 doesn't understand, and makes no attempt to learn what T4 is talking about. At times, a reader of the transcripts finds himself checking an exchange again to be sure which is the terminal and which is the judge, since it is often the judge who seems to be avoiding the kind of closely engaged conversation that a computer program would be incapable of.

One of the larger issues raised by such exchanges is that of the subject-matter competence of the judges. How useful is a judge who, as in this case, is so ignorant of and uninterested in Shakespeare that he cannot help trivializing a conversation about the playwright? Presumably a Turing Test judge should be familiar enough with the topic of conversation assigned to his terminal to hold up his end of an intelligent conversation.

Such comparative ignorance may make the judge ineffective—unable to engage the terminal's expert knowledge or too willing to assume that the terminal is a computer because it knows more about a subject than any mere person could know. A computer that fools a fool is hardly proof of artificial intelligence.

Of course, anyone with an understanding of how computers are made to mimic human responses would need no subject-matter expertise whatever to detect a computer posing as a human. Such a judge would simply demand that the hidden entity respond to the *ideas* represented by his questions, warning that it would be severely penalized for repeating any of the key words in those questions. Using this interrogative technique, the discriminative ability of judges should increase much faster than programmers' abilities to mimic human responses, and the Test should become correspondingly more difficult and thus more serious.

But this technique was clearly not thought of by any of the judges. Here is a representative exchange between Terminal 2 and a judge:

Judge 3: What a trivial topic!

Terminal 2: It came from forest fires in West Virginia.

Judge 3: Why do you say that?

Terminal 2: Who is your favorite singer?

And so on, interminably. T2 was apparently programmed to respond at random, and from a very limited repertoire of responses. In the course of its fourteen minutes of fame, T2 makes each of the following statements at least six times: "I want to have a child someday," "He is a sonar technician, he hunts Russian submarines," "It came from forest fires in West Virginia," and "I mean, have I convinced you I'm a computer yet?" These remarks are made at random, with no discernible relevance to anything said earlier by either the judges or the terminal itself. Yet two judges claimed to find this terminal more likely to be human than T4, who conducted (or tried to conduct) an amusing and informative conversation on Shakespeare. At one point, T2's remarks are so wildly incoherent that Judge 1 announces that he's penetrated its disguise: "Okay, you're busted! You're no machine. You're trying too hard, and it's becoming obvious." In other words, the judge seems to think that T2 must be a human trying to make him think he is a machine, and that no machine could be so confused or confusing.

One of the most extraordinary moments in the Competition comes when Terminal 3, whose assigned topic is "Second Grade School Topics," suddenly waxes philosophical:

> **Judge 9:** Can you ask me any questions?
>
> **Terminal 3:** Machines can only do what we have the skill to tell them to do.
>
> **Judge 9:** Do you want me to tell you about when I was in the 2nd grade?

T3's statement on the nature of machines is supposed to come from an eight-year-old, one whose performance up to this point suggested that, if real, she is amazingly scatter-brained and ignorant even for her age. But her startlingly grown-up and gratuitous observation about artificial intelligence—a quotation, in fact, from nineteenth-century computer pioneer Lady Lovelace—dropped suddenly in the exchange, goes unremarked upon by the judge, who simply returns the conversation to its tedious rut.

It would be interesting to know, incidentally, why this terminal was given the task of mimicking not just a human being, but an eight-year-old girl—who is programmed to inform judges that she is a gymnast and a Baptist, but given nothing plausible to say when asked the most obvious questions about those characteristics. If the planners of the Competition believed that mimicking an eight-year-old was easier than mimicking an adult, they were mistaken. While an eight-year-old would be forgiven for not knowing as much about the world as an adult, she would have mental quirks that would be harder for an adult programmer to foresee and mimic than the mature reactions of an adult.

The winner of the 1991 Competition was a program submitted by Thinking Software, Inc., whose handling of the topic "Whimsical Conversation" was good enough to convince five of the ten judges that it was human. The program, known during the trials as Terminal 5, issued remarks that were more amusing than most made by computer-driven terminals (this is not high praise), but were otherwise perfectly standard for such programs. It leaned primarily on the "non-directed therapy" strategy, introduced many years ago by Joseph Weizenbaum and (separately) Kenneth Colby, in which the program picks up a fragment from the input (e.g., the X in "I wish I knew more about X"), and inserts it into a canned response ("Why are you interested in X?"). At one point, Judge 2, getting a "therapeutic" answer, notices that T5 has even reproduced a typo and a

grammatical error that were part of his question, but he assumes that T5 is just making fun of him:

> **Judge 2:** I getting tired and yes how to live is a topugh one.
>
> **Terminal 5:** …Why do you need to tell me you getting tired and yes how to live is a topugh one? What is your thought on that?
>
> **Judge 2:** …I'm sorry I make so many typos but please dont' throw them back at me.

It does not occur to J2 that the letter-perfect playback of a portion of his question is at least as likely to indicate mechanical mindlessness as low-level human wit. When T5 can't find a usable fragment in its input to incorporate in a therapeutic answer, it falls back on issuing some non-responsive remark, yet one with enough meat in it to have a chance of distracting the judges from noticing its total irrelevance. And the gambit usually works, since most of the judges simply follow T5 wherever its random response generator takes it, never demanding that a consecutive, rational sequence of exchanges be developed.

> **Terminal 5:** …I used to drive the Caddy, but now I often use the Mercedes.…
>
> **Judge 9:** What kind of Mercedes?
>
> **Terminal 5:** What does a woman want anyway? Were you always a loser? What is it you really want to know?

Remarkably, five judges found T5 to be human, including two who also made the converse error of taking humans for computers. Overall, the performance of the judges leaves us to draw some sad conclusions about their inability to engage in sustained conversation, their lack of knowledge on general human subjects, and their need to share their personal concerns even with entities that contribute little more to the "conversation" than a mirror does to the relief of loneliness.

The Distortions of Ideology

The Loebner Competition has been held every year since 1991, in a variety of locations and under a variety of auspices. But what is clear from the skimpy record available to us of these events is that not much has changed, improved, or progressed: the "AI programs" remain amazingly simpleminded, and as time goes on fool fewer judges, belying Epstein's

prediction of 1993 that "the confederates will never get much better at the task, but the computers will get better each year."

Interestingly, Epstein sees the question of whether computers do, or can, or will think as essentially one of technology—a curious position for a psychologist. In summing up the prospects for computer intelligence or sentience, he grants that much remains to be done, but ends on an optimistic, even exultant, note, saying "…the sentient computer is inevitable. *We're* sentient computers, after all, and those who are skeptical about technological advances are usually left in the dust." But Epstein has forgotten Turing, the prophet who inspired the competition and who defined success for the Test not in terms of what computers will be able to do, but in terms of how we will think of their achievements. Will we ever call our marvelous machines "intelligent," or equate the activities of computers with the activities of the mind? So far, if the judges at the successive Loebner Prize Competitions are any indication, the common-sense answer seems to be no.

Of course, the failure to pass the Turing Test is an empirical fact, which could in principle be reversed tomorrow; what counts more heavily is that it is becoming clear to more and more observers that even if it were to be realized, its success would not signify what Turing and his followers assumed: even giving plausible answers to an interrogator's questions does not prove the presence of active intelligence in the device through which the answers are channeled. We have pulled aside the curtain, and exposed the old carny barker who calls himself the great and powerful Oz.

In discussing the "system" argument against his Chinese Room thought experiment, Searle once said, "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." The AI champions, in their desperate struggle to salvage the idea that computers can or will think, are indeed in the grip of an ideology: they are, as they see it, defending rationality itself. If it is denied that computers can, even in principle, think, then a claim is being tacitly made that humans have some special property that science will never understand—a "soul" or some similarly mystical entity. This is of course unacceptable to many scientists.

In the deepest sense, the AI champions see their critics as trying to reverse the triumph of the Enlightenment, with its promise that man's mind can understand everything, and as retreating to an obscurantist, religious outlook on the world. They see humanity as having to choose, right now, between accepting the possibility, if not the actual existence,

of thinking machines and sinking back into the Dark Ages. But these are not our only alternatives; there is a third way, the way of agnosticism, which means accepting the fact that we have not yet achieved artificial intelligence, and have no idea if we ever will. That fact in no way condemns us to revert to pre-rational modes of thinking—all it means is acknowledging that there is a lot we don't know, and that we will have to learn to suspend judgment. It may be uncomfortable to live with uncertainty, but it's far better than insisting, against all evidence, that we have accomplished a task that we have in fact scarcely begun.