The Turing Test and its Role in Artificial Intelligence
Part 2: Searle's Chinese Room TE, Loebner's Contest

Introduction

In the first part of this study (*Annals*, insert date), I examined the Turing Test (henceforth simply the "Test") directly, closely analysing key passages of the paper in which it was presented, and noting how the Test has been used by the AI community. In this second part, I will examine what may be called the offspring of the Test: first, what is generally regarded as the strongest attack on it, Searle's Chinese Room thought experiment, and second, the most determined effort yet made to perform it, the Loebner Prize Competition. At the end, I offer some general conclusions that seem to be warranted by the study as a whole.

The Chinese Room thought experiment

In 1980 John Searle, professor of philosophy at UC Berkeley, published a paper[1] in which he sought to discredit not just the Test but it the entire program that he called 'strong AI'—that which claims that a computer can be said to think. He encapsulated his argument in the form of the Chinese Room thought experiment (TE), in which he asks us to 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 (Searle gallantly volunteers himself, but anyone who knows English but not Chinese will do), 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 someone who knows nothing of what goes on within the black box that is the Chinese Room, but observes that it produces answers in Chinese to questions in Chinese, 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. What this TE shows, Searle claims, is that an ability to replace one string of symbols by another, however meaningful and responsive—that is, to answer questions correctly—can be done without an understanding of those symbols. The bearing of this TE on the Test is clear: it shows that an ability to provide good answers does not imply thinking. In fact, the TE *is* the Test, minimally revised so as to guarantee that there is no understanding within the Room of the inputs given it. It is too bad that Searle, in order to give us that guarantee, made the entity in the Room human (and a professor of philosophy, at that), because in doing so he opened the door to claims that the human's intelligence somehow has something to do with the translation process, even though he knows no Chinese. Much irrelevant speculation might have been nipped in the bud if Searle had made the inhabitant of the Room a trained chimpanzee (or even a simple, pure-hardware gadget). Although we would miss a categorical verbal assurance from the chimp that he didn't know Chinese, we would not be tempted to suppose that some general property of human intelligence was producing Chinese translations.

[1] Searle 1980; since the paper is most easily available in Hofstadter 1981, citations will be to its reprinting in that volume.

I think that the logic of the Chinese Room TE is valid, and that the many counterarguments[2] that have been put forth are all faulty in one way or another. But Searle has handled the controversy, for the most part, in such a way as to undercut his own success. He has well said, “The original Chinese room argument is so simple that its point tends to get lost in the dozens of interpretations, comments, and criticisms to which it has been subjected over the years.”[3] Indeed it does—which makes one wonder why he has sometimes joined his critics in elaborating his argument to the point where it becomes near incomprehensible.

Far from simplifying his TE as far as possible, so as to minimize possible objections and misunderstandings, Searle seems sometimes almost to revel in its elaboration and the inclusion of gratuitous flourishes and curlicues. Simply by putting a human being into the Chinese room, he invites mildly muddle-headed critics (Motzkin) to wonder, irrelevantly, in what language that human is thinking, and the moderately muddle-headed (Hofstadter) to suggest that in doing so he is falling victim to the 'homunculus' fallacy[4]; by bringing into it the handling of texts in Chinese, he suggests to the terminally muddle-headed (Papineau) that he is proposing an ability to translate Chinese as a reasonable criterion to use in evaluating the claim that machines think, and so on. When his critics take advantage of the openings he presents them with to becloud the simple point his TE really makes, Searle, instead of rejecting their moves and insisting on returning to the root issue, sometimes plays their game, and caps their fancies with some of his own.

For example, Searle and his critics, between them, introduce further *personae* into the Chinese room: they postulate that the room's inhabitant is a woman (no reason given; perhaps feminism is at work, and they believe that the Chinese room is a happy place where women should be given their turn); that there are other characters ('demons') involved, who are always—again, for no clear reason—male; that the whole Chinese room should be put inside a robot; that the collection of elements in the TE (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, and so on without visible end.

Here is a specimen of Hofstadter at work for the reader who quite understandably suspects I'm making all this up, or at least highly embroidering it:

> 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.[5]

Hofstadter offers no reason why we should follow him in his ascription of what seem to be wholly gratuitous characteristics to the *dramatis personae* of the TE (whose ranks are already swollen by 'demons' whose *raison*

[2] Virtually all of them are considered in Searle’s paper and Hofstadter’s reply to it; Preston 2002 contains a few new counter-arguments, but none that need detain us here. To see my analysis of the few new ones, send me an e-mail message.

[3] Motzkin 1989, page 45.

[4] The homunculus fallacy is the ascribing to a 'little man'—a homunculus—somewhere within a system such intelligence as the system seems to display, only to raise the question of where the homunculus himself gets that property. When the answer is that the homunculus, in turn, has an even littler man within himself, it is evident we are dealing with a problem of infinite regression. The man in the Chinese room, however, is not a homunculus; he is not introduced in order to account for the intelligence apparently at work in that room, so the question of accounting for *his* intelligence does not arise.

[5] Hofstadter 1981, page 377.

d'être is equally unclear), unless, like him, we value "color." In TEs even more than in most intellectual constructs, entities are not to be multiplied without necessity, but Hofstadter points to no such necessity, nor seems to realize one is needed. 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, as noted, seems happy to play this game, suggesting still further variations; at one point in setting up his TE, he says, “Now just to complicate the story a little, imagine that …”[6]. He gets quite carried away by the brainstorming spirit, and quite careless of the fact that the force of his original TE is diluted by every variation and elaboration he entertains. (The one occasion known to me in which he soberly and properly refuses to join the game of "let's see how many layers of confusion and distraction we can smother the Chinese room with," and redirects attention to the simple and irrefutable point of the original, is in his brief rebuke to Motzkin, quoted above)

Searle has invited misinterpretation of every sort by building an unnecessarily elaborate TE, when the need is for the simplest one that will establish his proposition (which is, to repeat, that some results usually obtainable only by the exercise of thought and understanding can be obtained without them). The Chinese room TE does demonstrate this, and is logically sound; but pedagogically and polemically it stumbles. The ideal TE does not stir up debate or cause critics to make clever rejoinders; it is not "provocative," nor does it "make you think"—it knocks you down with its simple irrefutability. It convinces in at least the sense that it leaves critics speechless; their hearts may remain unwon, but they can think of no way to refute it. The amount of debate that the Chinese Room TE has occasioned is the measure of its failure by this standard; and it fails, or is at least much less successful than it could be, because Searle is careless with just that aspect about which some critics thought he was being dangerously clever, the rhetorical.

The idea of the Chinese Room TE, stripped as it should be to its bare bones, is this: suppose that the first sine-function table ever had just been developed, and existed in one copy only. The man who secretly possessed that sole copy, though so completely unmathematical as to be unable to add or subtract, could nevertheless make a living, even a killing, by selling instant sine values to everyone who needed them. Without his services, others would have to compute from scratch the sine function of a given angle each time they needed it, while our lucky table-owner would merely need to look up the given angle in his table, and read off its sine-function value immediately. His clients would credit him with being a whiz at mathematics, if not a positive magician. And AI champions, even after being informed of the real situation, would insist that in some sense he was, after all, a mathematician—or at least that he and his table together constituted a mathematician.

The man in the Chinese Room is the man just described, provided with a different argument/function table. His new table does not contain angles and their corresponding sine values, but other graphics—we observers may call it the Chinese-questions/Chinese-answers table, but the man in the room doesn’t even know enough to call it that; to him it’s just the input-graphic/output-graphic table. And just as he acquired an undeserved reputation as a lightning calculator of sine values by responding instantly to any request, so he will now acquire one as a brilliant Sinologist by responding in perfect Chinese to Chinese-language questions.

[6] Hofstadter 1981, page 355.

“System”?—what system?

As AI partisans have attempted to defend the Test by mistakenly claiming that it simply applies to an unseen entity the same criterion we apply to humans, so they have attempted to refute the Chinese Room TE by arguing "All right, none of the *parts* of the Chinese room understands Chinese, but the *whole*—the 'system'—*does*." The 'system' retort—by far the most frequently mounted argument against the TE—is nevertheless erroneous, and on two levels.

First, if there is a ‘system’ at all, it is not for Searle’s critics to define it. A system is a assemblage of parts so organized as to obey some purpose to which all of them are subordinated; what are the parts of the Chinese Room TE, and what is their purpose? If there is any system, its parts are all and only those that Searle has created, and its purpose is Searle’s, which is to refute the doctrine of hard AI; if critics want to talk about the ‘Chinese Room system,’ that’s what they should be talking about. What warrant, then, have Searle’s critics for excluding from the system those people outside the Room, but within the TE, who are asking the questions in Chinese, and reading the answers submitted by the Room’s occupant? None at all, so far as I can see; they are excluded only because they do not lend themselves to the critics’ purpose. If they are admitted into the system, then the claim that the system understands Chinese becomes true—but trivially true, not true in the sense the critics would have it. And what of us, the audience to whom the TE is addressed—are we not, as such, part of the system? If not, why not? And if we are, we become witnesses to the fact that the Chinese speakers outside the room are mistaken in their belief that the man inside understands Chinese.

But I waive this line of argument against the system defense, because it is even more enlightening, and more destructive of that defense, to allow the critics to define the system as they wish, arbitrarily including some of the elements of the TE and excluding others. Given their head, the critics have come up with a ‘system’ that simply expresses in another form their incorrigible and indefensible conviction that if a result usually associated with intelligence is being produced, intelligence must be found somewhere in the 'system.' But the validity of this notion is precisely what is in question in this TE, so the 'system' argument is inadmissible. Its faults may be seen more clearly if it is put in syllogistic form:

- Since the Chinese Room produces output that requires competence in Chinese, it must either contain or constitute an entity with such competence;

- The Chinese Room is a system, or whole, none of whose *parts* is competent in Chinese;

- Therefore, the Chinese Room as a *whole* must be competent in Chinese.

Since the major premise is just an assertion of the very point at issue, it commits the fallacy of *petitio principii*, or begging the question, and the argument is therefore invalid. (Note that what is refuted here is not the conclusion, but this particular argument for it. It remains possible in principle that the conclusion is true; what is not true is that this argument proves it.)

The users of the ‘system’ argument try to prop it up with an analogy: none of the parts of the human brain, they point out, exhibit thinking, only the brain as a whole does so. Just so (they claim) the *parts* of the Chinese room may be mindless, but the *whole* thinks. But an essential element is missing from the analogy. We *know* that the brain is the physical organ of thought; the only question is whether it produces it in some circumscribed portion, or acts *en bloc*. This makes it legitimate to conclude, if an exhaustive search reveals no such portion, that the whole brain is what thinks, but we may not 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 a Chinese Room, only that active thinking is finished and done with before it 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, now perhaps dead for centuries, intended.[7]

In part 1 of this study, Raj Reddy was quoted as saying "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." In a way Reddy is right; those who say that machine intelligence is in the future do have the tense wrong—but so does Reddy in demanding that we speak of it in the present tense. Machine intelligence is in the *past*; when a machine does something intelligent, it is because some extraordinarily brilliant person(s), some time ago (perhaps millennia ago) found a way to preserve some fragment of intelligent action so that later, more ordinary people could perform it, or even build machines that could perform it. We, the more ordinary people who have inherited these treasures, have our own little triumphs: we can at least perform the procedures whose steps were codified for us by our brilliant ancestors, and even translate the steps of those procedures—we call them algorithms—into sequences of even simpler steps—programs—that our machines can carry out.

And in that same Pickwickian sense, those who claim that the Chinese Room 'system' understands Chinese even if none of its visible elements do, are right—or at least would be, except that 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 the sine-function salesroom depends ultimately on the original geniuses who originated the methods, linguistic or mathematical, of which we are the heirs. So enlarged, the system may be said to understand, but this hardly helps the AI enthusiasts; no one will be impressed by being assured that even if no part of a computer that has passed the Turing Test really understands what it's doing, the complete system, which includes every logician and mathematician at least as far back as the Babylonians, does understand.

Since Hofstadter has taken it upon himself in his and Dennett's *The Mind's I* to lead the counterattack against Searle by marshalling and deploying most of the counterarguments offered by all its critics, it is convenient to deal with those counterarguments by quoting or paraphrasing his presentations of them. But bear in mind that this is not an analysis of arguments due to one lone professor of computer science, but of those of virtually the entire array of critics of Searle's TE. The main thrust of Hofstadter's counterattack is the 'system' argument, which we have already disposed of, eked out with bits and pieces of others. What results is a kind of mosaic of fragments, not always consistent with each other, cemented together with many observations which, true or false, are of little or no apparent relevance.

In his attempted refutation of Searle, he likens those who scoff at AI to those who scoffed at non-Euclidean geometry when it was first presented to the world, and says, "About fifty years later, however, non-Euclidean geometry was rediscovered and slowly accepted."[8] Hofstadter's analogy is not one that does the AI cause much good: first, the critical fifty-year period has, for AI, already passed, and without success—it was in 1950 that Turing described his Test. Second, non-Euclidean geometry has *not* been accepted in the sense of becoming the way we now perceive the world; it is merely a mathematical tool that scientists find useful in explaining some

[7] The logic of the 'system' argument appears also in Agatha Christie's *Murder in the Calais Coach*, where Hercule Poirot, seeing that no one of the possible suspects in a murder can have committed it, deduces that *all* the suspects must have conspired to commit it. The difference, of course, is that Poirot had an undoubted corpse to account for, just as brain surgeons can be sure they're dealing with the organ of thought.

[8] Hofstadter 1981, page 374.

phenomena that lie well outside ordinary human experience. Third, his analogy overlooks all the many theories and notions that were scoffed at by narrow, pedestrian, unimaginative minds, and later turned out to be deserving of ridicule. If "they all laughed at Christopher Columbus," they had some reason to—some of his ideas were wrong. His stumbling on the American continent, or at least some of its offshore islands, while searching for a route to the Indies may distract our attention from his wrong ideas, but they remain wrong.

Before dealing further with Hofstadter, we must briefly revisit Searle's text itself. As observed before, he seems unable to resist the needless elaboration of his TE, and in doing so to offer a multitude of opportunities for misunderstanding to his critics, few of which they fail to take. What is more—*much* more—he elaborates his description of what is going on in the Chinese Room to the point where he himself sometimes gets confused, and offers explanations that make no sense. In the paraphrase of his TE offered at the outset, I trimmed away almost all the fat and gristle, but a few examples need to be looked at because they play a part in the debate between Searle and his critics. For example, he postulates not just the Chinese questions given to the man in the Room, and the Chinese answers produced by that man, but several other elements as well: a "story" in Chinese which is to be the subject of the questions, and which is passed into the room as what Searle calls "a second batch of Chinese script," but which plays no part whatever in the action, because the man in the room cannot read Chinese. Searle also throws in a set of stories in English, and questions and answers in English about *those* stories—elements added just to highlight the difference between answering questions that one understands and responding to questions that are, in effect, just arbitrary arguments to an argument/function table. (The full details of the Chinese Room TE as originally presented by Searle are so convoluted that a purely verbal analysis is futile; one forgets the beginning by the time one has reached the end. I offer in Appendix A a tabular presentation of them for the benefit of the exceptionally conscientious reader.)

The way most critics have dealt with the irrelevancies in the Chinese Room TE is to ignore them. Whether they have acted with conscious wisdom in pruning away the excrescences, or been protected from them by their hasty and not too careful reading, at least they have not, most of them, gotten bogged down by them. But Hofstadter has noted some of them, at least, and gotten duly bogged down. He has noted that among the paraphernalia in the Chinese Room is that batch of Chinese script that Searle calls a "story," and he jumps on this with both feet. In his "Reflections"—that is, his extended counterattack on Searle—he writes, "[Searle's] reader is urged to identify with a human being executing by hand the sequence of steps that a very clever AI program would allegedly go through as it read stories in Chinese in a manner sufficiently human-seeming as to be able to pass the Turning test."[9] Hofstadter has somehow gotten the idea that the man in the room has read the "story" that Searle has thrust into the room, or is somehow simulating what a computer program would do if it were trying to deal with such a story—his own writing isn't perfectly clear here—and is answering questions on the basis of what he has gleaned from that reading, or has come up with as a result of simulating a computer program. But the man in the room knows no Chinese; he has not read the story, because he cannot; and nothing whatever in the TE depends on that story. (It is an example of poetic justice that the man in the room is lumbered with this utterly useless mass of unintelligible stuff, because the man in the room is the very one who created the mess, Searle.)

But this is simply Searle's confused presentation overlaid with the further confusion contributed by Hofstadter (and several other critics of Searle's); it cannot be repeated too often that there are only three active elements in the Chinese Room: the questions entering it (the "arguments"), the lookup by the man inside using his lexicon (the "argument/function table"), and the answers he produces (the "functions"). Everything else is meaningless, obstructive rubbish. But on the basis of this misunderstanding of what is already a structure that seems to have

[9] Hofstadter 1981, pages 373-4.

been designed by Ludwig of Bavaria, Hofstadter proceeds to claim that Searle has violated the rules for TEs: "To the Systems-Reply advocates, Searle offers the thought that the human being in the room (whom we shall from now on refer to as 'Searle's demon") should simply memorize, or incorporate all the material on the 'bits of paper'. As if a human being could, by any conceivable stretch of the imagination, do this."[10]

Why is the man in the room to be called a "demon" from now on? Hofstadter does not deign to tell us, but two reasons suggest themselves: first, it is, for a critic of the TE, a step in the direction of taking charge of the case, as a trial lawyer always attempts to impose his viewpoint, his terminology, and his pace on a trial. Second, it suggests another famous thought-experimental demon, Maxwell's—and *that* demon is one who supposedly accomplishes something we think impossible: defeating the laws of thermodynamics—so if Searle, or the man representing him, is also a 'demon,' he is already half-guilty by association (who knows, maybe he lied when he said he didn't know Chinese!).

But more important than this attempt to prejudice the discussion with tendentious terminology, Hofstadter is trying to discredit Searle's TE—as he understands it—by claiming that some condition Searle has postulated is unrealistic. This shows that Hofstadter is unfamiliar with the conventions that apply to TEs: the only things ruled out of them are those that are *logically* impossible. The merely unrealistic or extremely difficult are not ruled out, because doing so would make TEs impossible at all; the reason why we have TEs is precisely that some experiments are, from a practical point of view, non-performable. The only constraint on a TE is that the conditions described must be possible *in principle*; practical objections, such as those imposed by human frailty, mortality, and other such limitations are out of place. In some cases, of course, ignoring such considerations means that the TE in question may be unconvincing—the audience can always decide that in this particular case such contingencies are of the essence, and cannot be ignored. When they do, it means that we have a case where constructing a valid TE is simply impossible—and such cases are common, which is why good TEs are uncommon. But the contention that the Chinese Room TE is such a case needs some justification, and Hofstadter offers none except to say that the feat of memorization or simulation of a computer program that he thinks—mistakenly—that it involves is not practically realizable. This would not be a fatal objection even if he were correct in supposing that such a feat were required of the man in the room.

The really mind-boggling thing about this objection of Hofstadter's, however, is that it is being made in the course of a defense of the most non-performable TE of all: the Turing Test. And he is not alone in this act of sawing off the branch on which he sits; several of Searle's critics in Preston (2002) also take a dig at the Chinese Room TE for being 'unrealistic' or 'unrealizable' without realizing that in doing so they are discrediting all TEs, and the Test particularly.

The Loebner Competition: Trying to Implement the Turing Test in Real Life

As noted earlier, the Test has entered the general imagination, even of those with no special interest in computing or philosophy: without making any effort to find such things, I have come across a novel (Rogers 1982) and a play (McEwan 1980) based on the Test, and have no doubt that a deliberate search would uncover many more treatments of it in imaginative literature. But its fame has generated more practical consequences as well; among those whose imaginations have been captured by it is Mr. Hugh Loebner, president of Crown Industries, Inc., a light manufacturing company in New Jersey. Mr. Loebner wanted to know how the Test

[10] Hofstadter 1981, page 375.

would work out in practice, and he was rich and enterprising enough to mount an effort to find out. He subsidized an annual competition, originally to be held each year in Boston under the auspices of the Cambridge Center for Behavioral Studies, for the purpose of identifying and rewarding that computer program that best approximates the program Turing postulated. The first such competition, held in 1991, was exceptionally well documented both by its official *Transcripts* (Cambridge 1991) and the many press reports, and repays close examination. (The competition has been held each year since then, in a variety of locations, but nothing new has emerged from later ones, and the first, because of its novelty, is by far the best reported and documented; for those reasons, I have chosen to concentrate on that first occasion.)

Most readers of Turing have regarded the Test as a thought-experiment (TE)—that is, a plan for an experiment that is in principle performable, but which is carried out only in the imagination because it presents serious difficulties in practice. Chief among such difficulties are cost; the impracticality of gathering all necessary resources at one time and place; and social or ethical obstacles. (The first two of these need no explanation; the last may be illustrated by Schroedinger's classic quantum-theory experiment, which might involve the death of a cat if carried out.) The Cambridge Center has acted on the apparent belief that the Test is a thought-experiment that has never been carried out only because funds and institutional support were lacking; with Loebner's munificence supplying the one, and the Center's energy and organizing skills the other, nothing, they seem to think, stands in the way of its performance.

But very few hypotheses can yield TEs; the only kind that qualify are the exceptional ones that involve so few variables, and permit so rigorous an isolation of those few variables from disturbing factors that all qualified judges can agree that the elements of the experiment are controllable, and hence unproblematic. The best known thought-experiments, such as the Einstein-Podolsky-Rosen in physics, or Searle’s 'Chinese Room' in artificial intelligence, feature experimental setups that have very few 'moving parts' or degrees of freedom; we readily grant that they could be performed without violating known laws or begging the questions they were designed to decide. Furthermore, acceptable thought-experiments yield decisive answers; their possible outcomes are few—usually just two—and the conclusions to be drawn from those outcomes are unambiguous.

By all these criteria, the Test fails to qualify as a thought-experiment. First, its apparatus is full of unknowns at every point. Turing dealt with some of these arbitrarily: how often must the judges guess a computer to be human before we accept their results as significant; at what date is common linguistic usage to be sampled, and so on. The Center settles the others equally arbitrarily: how many judges are there to be, how are they to be chosen, what instructions are to be given them, how long are the trials to last, and so on. Second, neither Turing nor the Center deals with the full range of possible outcomes (in discussing these, I waive all the foregoing objections to the idea that the Test has, in a scientific sense, any outcomes at all). What conclusions are we justified in reaching if the judges are generally successful in identifying humans as humans, 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? Even more interesting, what if judges are frequently mistaken, but in a way just the reverse of that expected by Turing—that is, what if they frequently mistake humans for computers? (This last possibility is no red herring; three Competition judges made this mistake, discussed here later.)

Again, one of the silent postulates of the Test is that of computer-naive judges, judges who would know virtually nothing of AI and its claims, and listen to the answers their questions elicited from the hidden entities without prejudice either way. 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 today who, while quite representative of "general educated opinion," have had no interaction to speak of with cleverly programmed computers, no encounter with the notion of 'thinking machines'? There is also an element of vanity involved: a judge who guesses wrong may feel a bit silly if he judges a computer to be human, and even sillier to mistake a human for a computer. Finally, there is the problem of getting the judges to take their task with total seriousness; as the *Transcripts* and other publications 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 good-faith best to reach considered, sober judgments.

For all these reasons (and a good many more, discussed in [Halpern 1990]), I conclude that the Test is not a proper thought-experiment, let alone the basis of an actually performable experiment. It is instead a thought-drama—a very effective piece of imaginative writing, or fantasy, that cannot bear critical analysis. It is an especially effective piece of rhetoric when its targets are scientists and mathematicians because it treats the notion of rhetoric, and even ordinary non-rhetorical language, with disdain, and such readers are highly susceptible to the appeal of anti-rhetorical rhetoric (they are especially susceptible, but all of us are to some degree: all political candidates today have learned to tell us how much they despise politics). If I am right, then the Competition is, taken on its own terms, a nullity; there is nothing to be learned from it about whether a computer can fairly be said to think. There may be something of interest to be learned from it nevertheless, but only if we put it in an entirely different context; it would appear to have more to tell us if we look at it as a psychological experiment of the type of Stanley Milgram's celebrated exploration of human willingness to accept orders given by authority-figures as a warrant for doing the strangest things.

An attempt at a careful reading of the *Transcripts* of the 1991 Competition can be somewhat frustrating; it does not pretend to be more than a verbatim record of the exchanges between the judges and the terminals, and 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 seem to be some inconsistencies in its reports of how the various judges voted after the trials. In the account that follows, therefore, I have had to infer some facts from internal evidence, and take others from accounts in the press and elsewhere ([Stipp 1991a], [Stipp 1991b], [Epstein 1992]). I do not think that any vital facts are omitted or misrepresented, but the reader should be aware that I have had to use outside sources to supplement and understand the document under review.

With this caveat given, then, the essential facts are these: there were eight terminals in the 1991 Competition, of which six were later revealed to be driven by computers, and two by humans. There were ten judges, all from the Boston area, all "without extensive computer training." Each terminal was given 14 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 he had interacted with (my "he" here is conventional; seven of the ten judges were women) as 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 got, 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 might be judged to be computer-driven—and not even placed among the best of them. A memo from Dr. Robert Epstein, the director of the Cambridge Center and organizer-in-chief of the Competition, to the committee that supervised the planning of the Competition, reads in part [Epstein 1992, p. 3]:

> [The scoring method here proposed] preserves binary judgment errors on the part of individual judges. It will reveal when a judge misclassifies *a computer as a human*. [emphasis supplied]

No mention of the possibility of the converse error. Later in the same paper, Epstein reviews the results of the first Competition, starting with the observation that "The surprises were notable"; after listing several of these, he goes on to say "Perhaps even more remarkable, Cynthia Clay ... was mistaken for a computer by three judges." All this makes it clear that this outcome was assigned negligible probability when the Competition was planned, if indeed it was thought of at all. In the event, the program that won the Competition was thought human by one judge who also mistook both the human-driven terminals for the computer-driven kind.

The overall results, to anyone reading the *Transcripts*, are hard to understand except on the hypothesis that the judges, like most of those involved, were simply having a good time. Epstein makes a point of this in his background paper, writing [Epstein 1992, p.5] "The first contest fulfilled yet another desire of the prize committee. It was great fun. It was an extravaganza. A live audience of 200 laughed and cheered. ... Food flowed all day." ("Flowing" seems a strange action for food; something more usually done by liquids.) The topics assigned to the terminals further reinforce the impression that the atmosphere was one of playfulness; they were: Women's Clothing, Small Talk, Second Grade School Topics, Shakespeare's Plays, Whimsical Conversation, Problems in Romantic Relationships, Burgundy Wine, and Dry Martinis. The human 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 who thought her a computer did so because she knew so much about her topic that she exceeded their expectations for mere humans. But Ms Clay's answers to the questions put to her, while indeed showing wide and detailed knowledge of Shakespearean texts and stagecraft, were even more remarkable for responsiveness and cogency than for mere factual content.

Could it have escaped the notice of any attentive and serious judge that the conversation he was holding with Terminal 4 (as Ms Clay was known during the trials) was just that, a coherent and consecutive exchange of views and information with a lively, responsive partner? Could such a judge have failed to see the difference between such a conversation, and the near-random fragments of verbiage emitted by the computer-driven terminals, whose connection to the questions that elicited them was, at best, that they echoed a few of the questions' words? Here is a representative exchange between Ms Clay and one of the judges (here and throughout, except as otherwise noted, I have corrected typos, spelling errors, and other mechanicals):

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 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. Nevertheless, 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!

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 depth, 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.

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 him? Assuming for the sake of argument that the Test is a potential basis for a real experiment, I will suggest that a Test judge should be familiar enough with the topic of conversation assigned to his terminal to hold up his end of an intelligent conversation. When a judge interacts with an interlocutor who obviously knows far more about their assigned topic than he does, he cannot engage in the sustained exchanges that could make it unmistakably clear whether the terminal was human- or computer-driven. It may also make the judge rather defensive and touchy; as the exchange reproduced above illustrates, he may not even be willing to take advantage of the terminal's superior knowledge to ask questions of it, but may simply rebuff the terminal's advances with a curt "Learning that you like them answers my question." Having more subject-matter competence themselves would also help judges avoid the mistake of the several who thought T4 must be a computer because it knew more about Shakespeare than any mere person could know.

(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 it that it would be severely penalized for repeating any of the key words in those questions. As this technique becomes widely understood and employed, the discriminative ability of judges will increase much faster than programmers' abilities to mimic human responses, and the Test will become correspondingly harder to pass, or even to stage.)

Two judges decided that T2 was more human than T4; here is a representative exchange between T2 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 an amusing and informative conversation on Shakespeare, so far as the judges permitted her to.

At one point, T2's remarks are so wildly incoherent that Judge 1 announces that he's penetrated its disguise:

Judge 1: "Okay, you're busted! You're no machine. You're trying too hard, and it's becoming obvious."

From this remark, it would seem that some contestants were told that it was not their task to convince the judges that they were human, but simply to confuse the judges. Judge 1, at least, seems to think here that T2 must be a human trying to make him think he is a machine, and also believes, apparently, that no machine could be so confused or confusing. Or perhaps the judge was simply having fun. The mood throughout, as already noted, was light-hearted, to say the least; I have criticized the judges for being, in general, unprepared, but a few may have prepared only too well for the discussions about Burgundy wine and dry Martinis. Which brings up the issue of judges' seriousness, as least as important as that of their subject-matter expertise; judges must have judicial temperament as well as learning. Assuming *per impossibile* that the Test is worth carrying out at all, judges must be given incentive to do their job seriously; there must be some real reward for judging right, and real penalty for judging wrong.

One of the most extraordinary moments in the Competition comes when Terminal 3, whose assigned topic is "Second Grade School Topics" (and who does so poor a job of impersonating an eight-year old girl that even of these remarkably laid-back judges, only one thought it was human-driven), suddenly becomes a channel for the spirit of Ada, Lady Lovelace:

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 dictum is supposed to come from an eight-year old, whose performance up to this point suggests 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 goes unremarked by J9, who simply returns the conversation to the tedious rut it has been in so far.

It would be interesting to know, incidentally, why this terminal was given the task of mimicking not just a human, but an eight-year old girl—who is further made to inform the judges that she is a gymnast and a Baptist, but given nothing plausible to say when asked the obvious questions about those characteristics. If it was supposed by the planners of the Competition that the task of mimicking an eight-year old was easier than that of 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 reactions of an adult. And in fact only one judge voted that T3 was human; for these judges, a triumph of canny skepticism. Overall, T3 came in second from last in the voting.

A moment's reflection, too, might have made the judges wonder how likely it was that an eight-year old, with all the tendencies to get bored or silly or suddenly tired and cranky typical of that age, would be allowed to take the part of 'hidden entity' in the Competition. The fact that a child of that age would probably not be a decent typist could be dealt with by the provision of an adult typist at her end, but the other factors just mentioned make it very hard to believe that T3 could have been driven by an eight-year old child. Did the judges not think of this, or did they dismiss the objection, feeling that only a party-pooper would admit so pedestrian a consideration?

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. Usually it leaned on the "non-directed therapy" strategy, introduced many years ago by Weizenbaum and (separately) 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 in it a typo and a grammatical error that were part of his question, but assumes that T5 is making fun of him (in the exchange that follows, I reproduce the relevant parts literally, and indicate by ellipses the omission of irrelevant parts):

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 a human trying to be witty (and just how funny or clever is it to throw his typos or minor grammatical lapses back in the face of your conversational partner?). Among the five judges who found T5 human, two also made the converse error of taking humans for computers: J3 thought one of the actually human-driven terminals was computer-driven; J10 thought *both* the human-driven terminals were computer-driven.

When T5 can't find a usable fragment in its input to give a therapeutic answer, it falls back on issuing some remark that is non-responsive, but has enough meat in it to have a chance of distracting the judges from noticing its total irrelevance. And the gambit usually works; most of the judges simply follow T5 wherever its random response generator takes it, never asking 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?"

Some rather sad conclusions emerge from the *Transcripts* about the quality of general conversation among many of our fellow citizens. First, they neither expect nor require that a conversation stick to any topic or train of thought, even for as short a stretch as two or three remarks on each side. Second, they not only possess no organized bodies of knowledge on any topic outside their personal lives, but find it hard to believe that anyone does. When they meet an entity that clearly has such a body of knowledge, they find the fact so remarkable that it may be easier for them to suppose that entity a computer than a fellow human—even though that entity shows itself capable and desirous of carrying on a witty, literate, and cumulatively developing dialogue. Third, they are often so hungry to have someone listen to their personal concerns that they will put up with a partner who contributes little more to the "conversation" than a mirror does to the relief of loneliness.

Perhaps the most valuable fact reported by Epstein in the paper [Epstein 1992] from which I have already quoted several times is one whose full significance he does not seem to realize: "Unlike the judges, members of the audience rarely misclassified the terminals, perhaps because members of the audience could communicate with each other; judges could not." Whatever the value of his conjecture as to why mere spectators succeeded so much better than the official judges at correctly distinguishing humans from computers, their success suggests that something is wrong with the plan of the Competition. The judges, remember, are supposed to be representative of "general educated opinion"—that is, of just the kind of people who presumably constituted the audience. (The audience may have included many people with more computer training than the judges, but that does not matter; Epstein has run statistical analyses to discover whether such training is a factor, and assures us that "expertise in computer science had no systematic effect.") If the audience came to conclusions different from those of the judges—and they were strikingly different—then there is an anomaly to be resolved. In fact, it was the audience, not the official judges, who met the criteria set up by Turing; they were the real judges, and it is their judgments that ought to be studied.

Closely related to this point is another whose potential for causing offense makes one wish it could be glossed over. But Turing himself brought it up, and if our topic is to be dealt with seriously, we cannot suppress it: it is the question of what constitutes "educated opinion," and whose opinions are to count about whether a hidden entity is thinking or not. Turing touches on the point in the course of refuting an objection to his thesis that, he says, is tantamount to solipsism; in rejecting the objection, he observes (page 446) "...it is usual to have the polite convention that everyone thinks." On the grounds that one does not refer to one of one's own beliefs as a "polite convention," I infer that Turing himself, as might be expected of a young scientific genius, did not accept the idea that "everyone thinks"—a conclusion supported by his contemptuous dismissal, at the outset of his paper, of common English-language usage. I think it likely that Turing would see the judges employed at the 1991 Competition as unqualified to render judgment as to whether hidden entities do or do not think.

Finally, I observe that 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 here; the prophet in whose name the Competition is being held defined success for the Test not in terms of what computers will be able to do, but of how we humans will think of their achievements. Let computers do everything Epstein and other AI visionaries dream of in their most euphoric moments—what counts for Turing is how we, their creators and programmers, talk about their activities: will we use for their behavior the same word we use for what humans do with their minds? That is what Turing thought; if Epstein disagrees, that is his right—but he must not claim to be carrying out the Turing Test.

In concluding that we can learn nothing about whether machines can think from this or any attempt to realize the Test, I do not mean to say that such efforts have no value at all. Artificial Intelligence is notoriously a research program of which only the by-products have value; the trick is to throw out the dull baby, and keep the sparkling bath water. From the Loebner Prize Competition we will learn nothing about whether machines think, but we may yet learn something about how and when humans do.

General Conclusions

We have now considered the classic paper by Turing in which the Test was first described; some variant readings of that paper; the use to which the Test and Turing's prestige have been put by the AI community; and two outgrowths of the Test, Searles' Chinese Room thought experiment and Loebner's attempt to realize the Test to the extent possible with today's technology and understanding. What conclusions may we draw after so much preparation? I suggest the following:

- The attempts by the Loebner Award contestants to realize the Test only make it painfully clear that we are no closer to that goal today than we were in Turing's day, even though we have many orders of magnitude more computing power than Turing dreamt of. The failure to realize the Test is of course only 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 it would—we know that no number of plausible answers to our questions imply 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.

- Searle, discussing the 'system' argument against his Chinese Room TE, says, "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."[11] 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 scientists (and even more to aspirants to the title "scientist"). 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, 'medieval' outlook on the world; deny that the computer can think, and you are in their eyes halfway to bringing back the Inquisition. They see humanity as having to choose, and right now, between accepting the possibility, if not the actual achievement, of machine thought, or entering a new dark ages. What I would urge on them is agnosticism—an acceptance of the fact that we have not yet achieved AI, and have no idea when if ever we will. That fact in no way condemns us to revert to pre-rational modes of thinking—all it means is that there is a lot we don't know, and that we will have to learn to suspend judgement. It may be uncomfortable to live with uncertainty, but it's far better than insisting, against all evidence, that the emperor is well dressed.

- Like the cult leaders who endlessly prophesy the end of the world or other apocalyptic event, the champions of AI are forever holding out the promise that AI will someday—fifty years hence is a favorite target date—redeem all its claims. And strangely, AI champions are often allowed to get away with this; they keep saying, in effect, "Hey, everyone knows that eventually we'll be able to pass the Turing Test, or otherwise justify all our claims, so let's talk as if that were already done—the mere doing of it is a detail that we shouldn't get hung up on." In golf, that would be regarded as asking for a gimme on every hole, including all the tee shots.

Appendix A: The Full Chinese Room Apparatus

The table below presents the full inventory of the elements of Searle's Chinese Room, along with comments on each by the man in the Chinese Room.

Searle's name for element	Contents	Role in the Chinese Room TE	Comment by the man in the Room
SCRIPT (Batch 1)	Chinese writing.	none apparent	I have no idea what this batch of graphics is for.
STORY (Batch 2)	Chinese writing, plus English rules for "correlating this batch with batch 1"	none apparent; I cannot read the story, since it's in Chinese, and what it means to "correlate" this mass of meaningless characters to	More graphics that mean nothing to me, plus some rules that I can read, but not apply; not clear how this differs from either the rules that

[11] Hofstadter 1981, page 359.

		the SCRIPT — itself meaningless— is a mystery.	accompany the QUESTIONS, or the PROGRAM.
QUESTIONS (Batch 3)	Chinese writing, plus English rules for “correlating this batch with batches 1 and 2” and for forming answers.	The rules for forming answers to given questions need be stated only once, since they’re the same for all the questions: “find these graphics in your lexicon, copy to an output slip the graphics associated with them.”.	At last, something I can do something with: here are graphics to be replaced with other graphics according to the rules in English.
ANSWERS	The Chinese characters I return in response to the Answers.	OK	The graphics I return for those given me in Batch 3.
PROGRAM	This is all the rules given in the earlier batches.	Not clear why this category is needed.	
Stories in English		none apparent.	These three items are here, apparently, just to contrast answering questions I understand and those I don’t
Questions in English			
Answers in English			

REFERENCES

Anderson, David (1989), *Artificial Intelligence and Intelligent Systems: the Implications.* New York: John Wiley & Sons.

Anon. (1984), AP wire story “Reagan Advisers Firm on ‘Star Wars’ Despite Doubts in Science Study,” *Los Angeles Times*, April 26, page 4.

Buchanan, Bruce G., Lederberg, Joseph & McCarthy, John (1976), *Three Reviews of J. Weizenbaum's `Computer Power and Human Reason'*, Stanford University Computer Science Department Report No. STAN-CS-76-577, (AD/A-044 713).

Cambridge (1991), *1991 Loebner Prize Competition: Official Transcripts*, November 8, 1991, Center for Behavioral Studies, Inc., The Computer Museum, Boston, Massachusetts.

Collins, Harry M. (1990), *Artificial Experts: Social Knowledge and Intelligent Machines*. MIT Press.

Epstein, Robert (1992), "The Quest for the Thinking Computer," *AI Magazine,* Summer 1992, pages 80-95.

Gelernter, H. L. et al. (1977), “Empirical explorations of SYNCHEM,” *Science*, 197, 4308, pages 1041-1049.

Gleason, Andrew M. (1978), “The World of Four Colors,” *Harvard Magazine* March-April, 1978, 21.

Goodman, Nelson, “Inductive Translation,” in *Problems and Projects* (Bobbs-Merrill, 1972), pp. 294-297.

Grabiner, Judith V. (1986), “Computers and the Nature of Man: A Historian’s Perspective on Controversies About Artificial Intelligence,” *Bulletin of the American Mathematical Society* (October 1986), pp 113-126

Gunderson, Keith (1985), *Mentality and Machines*, 2nd edn, Minneapolis: University of Minnesota Press. (Original edition: Doubleday Anchor Books, 1971.)

Halpern, Mark (1990), *Binding Time*. Norwood, NJ: Ablex Publishing Corp.

Hayes, Patrick J. (1992), letter to the editor, "ACM Forum," in *Communications of the ACM* (December 1992), pages 13-14.

Hodges, Andrew (1983), *Alan Turing: The Enigma*. London: Burnett.

Hofstadter, Douglas R. (1981), and Daniel C. Dennett (eds.), *The Mind's I.* Basic Books (also Bantam pb.)

Lanier, Jaron (2001), quoted in Natalie Angier, "Defining the Undefinable: Being Alive," *The New York Times* (December 18, 2001), pages D1 and D6

Lenat, Douglas (2001), quoted in *Wired* (November 2001), page

McEwan, Ian (1980), *The Imitation Game* (see Hermione Lee, "Cracking the Codes of Tyranny," *Times Literary Supplement*, April 25, 1980, page 467).

Minsky, Marvin (1961), "Steps toward artificial intelligence," *Proceedings of the IRE*, 8-30.

————————(2003), quoted in "AI Founder Blasts Modern Research," *Wired News* (May 13, 2003), pages 1-3, at pages 1 and 2.

Moor, James H. (2003), (ed.) *The Turing test: the elusive standard of artificial intelligence*. Kluwer Academic Publishers.

Motzkin, Elhanan, and John Searle (1989), "Artificial Intelligence: An Exhange," *New York Review of Books* (February 16), pp. 44-45.

Naur, Peter (1986) "Thinking and Turing's Test," *BIT* 26, 1986, pages 175-187.

David Papineau (1984), "The Significance of Squiggles," [review of Searle's Reith Lectures] *Times Literary Supplement*, December 14, 1984, p. 1442.

Perutz, Max F. (1985), "Brave New World," *New York Review of Books* (September 26), page 14.

Pollack, Martha (2003), in "AI Founder Blasts Modern Research," *Wired News* (May 13, 2003), pages 1-3 (www.wired.com/news/technology/0,1294,58714,00.html).

Preston, John, and Mark Bishop (2002), (eds.) *Views into the Chinese room: new essays on Searle and artificial intelligence*. Oxford: Clarendon Press.

Reddy, Raj (1995), "To Dream the Possible Dream," *Turing Award Lecture*, March 1, 1995

————(1996), "The Challenge of Artificial Intelligence," *IEEE Computer* (October 1996), pages 86-98.

Rogers, M. (1982), *Silicon Valley*. New York: Simon & Schuster.

Russell, Bertrand (1903), *Principles of Mathematics*. Cambridge: The University Press.

Searle, John (1980), "Minds, Brains, and Programs," *Behavioral and Brain Sciences 3* (1980), pages 417-457; reprinted in Hofstadter 1981.

Stipp, David (1991a), "Does That Computer Have Something on Its Mind?," *Wall Street Journal*, March 19, 1991, p. A22.

------------- (1991b), "Some Computers Manage to Fool People At Game of Imitating Human Beings," *Wall Street Journal*, November 11, 1991, p. B5B

Turing, Alan M. (1950), "Computing machinery and Intelligence," *Mind* LIX, 236, pp. 433-460. Reprinted in (among other places) J. Newman (ed.) *The World of Mathematics* (New York: Simon & Schuster,1956) [retitled "Can a machine think?"], IV, 2099-2123;
A. R. Anderson (ed.) *Minds and Machines* (Englewood Cliffs, NJ: Prentice- Hall, 1964), pp. 4-30;
D. R. Hofstader & D. C. Dennett (eds.) *The Mind's I* (New York: Basic Books, 1981), pp.53-67;
E. A. Feigenbaum & J. Feldman (eds.) *Computers & Thought* (New York: McGraw-Hill, 1963).

Wegner, Peter (1987), letter to the editor, *Abacus* (Spring 1987), pages 5-7.

Wilkes, Maurice V. (1985), *Memoirs of a Computer Pioneer*. The MIT Press.

————————(1992), "Artificial Intelligence as the Year 2000 Approaches," *Communications of the ACM* (August 1992), pages 17-20.