The Turing Test and its Role in Artificial Intelligence
Part 1: The Career of Turing's Paper

Introduction

This is the first part of a two-part paper on the intellectual construct called the Turing Test (henceforth just "Test"), and the role it has played, mainly in Artificial Intelligence, but also to some extent in Robotics, Epistemology, the Philosophy of Mind, and related disciplines and projects. This first part deals with direct appeals to the Test by AI workers, and various interpretations that have been made of it; the second, to appear here in the near future, deals with the most determined and thorough-going attempt to realize the Test, the Loebner Prize Competition, and with the principal attack on the Test, John Searle's Chinese Room thought experiment.

I am not concerned here with any of the above-mentioned disciplines and projects in their own right, but only with their relation to the Test; nor do I deal with the many uses that have been made of the Test by novelists, playwrights, and others for whom it is a stage property or rhetorical trope. This disclaimer is necessary because the Test has acquired a life of its own, quite apart from the philosophical and scientific setting that Turing originally put it in. The image of an investigator seated at a computer terminal, matching wits with an unseen adversary in an attempt to unmask him, has become practically part of our folk memory, and a staple element of modern thrillers in television, books, plays, and movies. To explore the role that the Test plays in such contexts would be an interesting project, but is not the present one.

I have had the benefit of some comments from reviewers of an earlier draft of this paper, and may be able to forestall some further criticism by a declaration here at the outset. One reviewer admonished me for having definite positions on the several controversial issues I will be dealing with here, implying that a true scholar would simply present "the facts," and let them speak for themselves. I strongly disagree; to my mind, that would be like an account of the role that the phlogiston theory played in the development of physics that failed to note that the theory had been thoroughly discredited, or an account of witchcraft that refrained from taking sides on the question of whether witches really exist. I don't know which would be worse: a writer who suppressed his views on such questions in order to seem "objective," or one who honestly thought those questions were still open.

My idea of a scholarly account of a controversial issue is one in which the writer makes his position quite explicit, and supports that position with clear reasoning and whatever evidence he can find, and I have tried to do that here. Any account of a controversial issue that is not organized around a thesis, a definite point of view, is not a history but merely a chronology of more or less random facts; at best, a trove of information that a real historian might find use for in producing a real history. I am well aware that my views are not shared by everyone, and I do not claim to have uttered the last word on the subject; anyone who disagrees with my argument is free to publish a rebuttal, which I will read with interest. If I have reasoned badly anywhere, or distorted or omitted relevant facts, I will be grateful for civil corrections, but I do not apologize for having a point of view, and presenting the subject as seen from it.

A final note: to keep this paper within the limits set by *Annals*, I have had to deal very summarily with some topics, particularly with some of the more esoteric interpretations of the Test and of Searle's Chinese Room thought experiment; readers who wish to delve deeper into these matters are invited to contact me via e-mail.

Origins: Turing's 1950 Paper

In the October 1950 issue of the British quarterly *Mind*, Alan Turing published a 28-page paper titled "Computing Machinery and Intelligence" [Turing 1950]. It was recognized almost instantly as a landmark. A striking indication of its success is that in 1956, less than six years after its publication in a periodical of small circulation, 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 has since prevailed over what Turing really said, James Newman, editor of the anthology, silently retitled the paper "Can a Machine Think?" when he reprinted it.) Since then it has become the most reprinted, cited, quoted, misquoted, paraphrased, alluded to, and generally referenced philosophical paper ever published.

Turing claims there that suitably programmed digital computers (not, as Newman's retitling would suggest, any old machine) will by about the year 2000 come to be generally accepted as thinking. In preparing his readers to accept this idea, he explains what a digital computer is, presenting it as a special case of the 'discrete state machine'; he offers a capsule explanation of what 'programming' such a machine means; and he refutes—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 what it is today, a document as significant in the history of Western thought as the Mayflower Compact, Luther's Ninety-five Theses, or the Communist Manifesto. The part that has seized our imagination, to the point where thousands who have never seen the paper nevertheless clearly remember it, is that in which he introduces a test for determining whether a computer is thinking; a test he calls the Imitation Game, but which the world has since decided to call the Turing Test .

Having postulated that programming will advance to the point where a computer can be made to respond to questions posed by an interrogator very much as a human would, he then claims that that development offers us a way to determine whether such a computer is thinking. What he proposes is an experiment in which the interrogator asks questions of a hidden entity, which might be either a computer or another human being, and is then required to decide, solely on the basis of the answers given to his questions by that hidden entity, whether he had been interrogating a computer or a human[1]. If the results of this experiment—that is, of the Test—show that interrogators cannot, on the basis of such trials, distinguish humans from computers any better than they can distinguish, say, men from women by the same means, then, Turing claims, we have no good reason to deny that the computer that deceived them was thinking.

Turing does not *argue* for the idea that an ability to convince some 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—and this resort to mere assertion is a problem that has troubled many. Some of his defenders have tried to supply the underpinning that Turing himself apparently saw no need for by arguing that in the Test, he merely asks us to judge the unseen entity the same way we regularly judge humans: if they answer our questions in what seems to us a reasonable way, we say they're thinking. Let us be fair, they argue, and apply the same criterion in deciding whether other entities are thinking.

[1] See Appendix 1, "The Preliminary Version of the Test: Man or Woman?," for discussion of some details.

But this defense fails, because we do *not* decide on the basis of his ability to reply reasonably to our questions that someone we see as human can think—we accept any human being on sight and without question as a thinking being, just as we distinguish a man from a woman on sight.[2] If, after some verbal exchanges with a human newcomer, questions were to arise about whether he can think, we would consider whether the words he had addressed to us seemed to spring from some autonomous process within him, rather than being mere echoes of words we had addressed to him; if they did, we would regard that as confirming our assumption that he was thinking. We might, if his words were incoherent, judge him to be stupid, injured, drugged, or drunk; we would not question his ability to think.

If, on the other hand, his responses seemed to be nothing more than reshufflings and echoes of the words we had addressed to him, or if they seemed to be not so much addressing as parrying or evading our questions, we would decide either that he was not acting in good faith, or was gravely brain-damaged and thus accidentally deprived of his birthright ability to think. In short, 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. What we require if we are to regard an interlocutor as a thinking being is that his responses be autonomous; to think is to think for yourself.

The *immediate* and *default* ground, however, of our belief that a human being is thinking has nothing to do with answering questions; it is that we can *see* he is human; and once we see him as such, we are extremely reluctant to give up that belief, and do so only in most unusual and highly contrived circumstances. And that belief is only further confirmed if the words he addresses to us are not rehashings of the words we just said to him, but words we did not use, did not think of, possibly would never have thought of—in short, if they are not derivative, implying that they issue from someone doing the same thing we are: thinking. We believe that our interlocutor has a mind, just as we have, when it is clear from his answer that he is dealing—even if badly—with the meaning of what we said to him, and not with the words we happened to use.

Only if a human interlocutor fails to respond so, and over a considerable period of time, would we entertain the possibility that he was not thinking, and even then we would attribute that condition not to a constitutional lack of ability to think, but to some sort of pathological interference with that ability. For good or ill, when we know we are dealing with a human, the question of whether he is a thinking entity does not even arise—we have to strain our imagination to come up with a situation in which it might—whereas in the Test, that is the first and only question. Perhaps our automatic attribution of thinking ability to anyone who is visibly human is deplorably superficial and lacking in rigor, but nevertheless, that is what we do. And though we can debate forever the exact nature of "thinking," it is by definition what the human mind does. So if we are to credit some unseen entity with being a "thinker," that entity had better respond as we expect a human would, and make us see it, in our mind's eye, as a human being.

This preliminary discussion of what human beings commonly mean by *thinking* is necessary because Test defenders have erroneously argued, as noted, that it simply applies to other entities the criterion by which we recognize human as thinkers, and because Turing pins his claim that computers can think on a prediction about how people will come to use that term in the near future:

[2] In addition to this negative reason for rejecting the 'single standard' argument, there is a positive one: we can *observe* what the computer really is doing, and understand, especially if we are programmers, why it deceives many into supposing it is thinking. But pursuing this thesis would take us on a long digression from the present line of argument, so it will not be followed up at this point; instead, a reference will be given below for those interested .

> 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.[3]

And as the text goes on to make clear, he 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. This has made the Test highly problematic for AI enthusiasts, who want to enlist Turing as their spiritual father and philosophic patron, because while they have programmed the computer to do things that might have astonished even him, they have not gotten it to pass his Test—that is, to converse with humans in such a way as to compel general recognition of it as a thinking entity.

The relationship of the AI community to Turing, therefore, is 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 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. So when that community feels in need of some illustrious forebear who will 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 in particular 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 important in AI work, if they ever were, 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 exactly what most AI workers do with it. Examples of such treatment will be offered after a more detailed examination of what Turing actually wrote.

The Specifics of the Test

After introducing the general idea of the Test, Turing goes on to offer a presumably representative fragment of the Test in action, and in doing so, creates an imaginary interrogator who is just as inept as the actual interrogators whose attempts, half a century later, at such detective work will be reviewed in some detail in the second part of this study. As I hope the preceding discussion has established, the key to successful discrimination between a programmed computer and a human is to ask the unseen entity the sort of questions that humans find easy to answer, but that an AI programmer will find impossible to predict and deal with, and to use such questions to unmask evasive and merely word-juggling answers. It is instructive to examine Turing's suggested line of questioning with this in mind (pages 434-435):

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

[3] (Turing 1950) page 442. Since all quotations from Turing, unless stated otherwise, will be from this paper, full citation form will be omitted from now on; only page numbers, from its original publication in *Mind*, will be offered.

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, but simply switches 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 plays chess at all would be stumped by an end-game so simple that a mate-in-one was available; second, it introduces an assumption that cannot be automatically allowed: namely, that the computer is to play to win.

It may seem rather pedantic and pettifogging to call attention to, and disallow, all these simple assumptions; after all, they amount to no more than ordinary common sense. *Exactly*. Turing's sample questions seem almost deliberately designed to keep the interrogator from understanding what he's dealing with—but perhaps that hardly matters, since he endows the computer with enough cleverness to fool the interrogator forever.

I have rationalized and simplified—but not, I think, oversimplified—Turing's rather rambling and sketchy description of the Test. I have also overlooked, or treated very summarily, several other highly problematic aspects of Turing's general thesis because I am concerned here with just those of his arguments that AI workers and philosophers have made use of. (For those who want a quick indication of my personal view of AI, it is that computers are general-purpose algorithm executors, and that their apparent intelligent activity is simply an illusion that some of us suffer from because we do not fully appreciate the way in which algorithms capture and preserve the fruits of intelligence. A full discussion of this position is to be found in my book

Binding Time [Halpern 1990].) There are in particular two problematic aspects of Turing's paper that need be noticed here: first, the glaring contradiction between his initial refusal to respect the common understanding of the words *machine* and *think*, and his appeal at the conclusion of his argument to just such common usage; and second, his making *surprise* the characteristic sign of an AI success.

The first of these points is quickly illustrated. At the very outset of his paper, he tells us (page 433):

> 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 a few pages later he gives us, in the words quoted earlier, what he regards as a non-absurd criterion: that by the end of the 20th century, a 'Gallup poll' would show that the public now accepts that the computer can think. To which the answer is, of course, "And what of it? If the common usage of 1950 is unworthy of serious consideration, why is that of 2000 deserving of greater respect?" Turing's lofty, not to say arrogant, repudiation of common usage gets promptly forgotten as soon as he imagines an era in which common usage supports his thesis.

And in any case, Turing's prediction has failed. We are now well beyond the date by which usage should have changed as Turing predicted, and no such change has occurred. 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 a person "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, and no one who has been told that the reason his hotel reservation has been lost is that "the computer messed up" is likely to use the term "thinking machine" except sarcastically.

Although Turing's prediction has failed—and not simply because 50 years was not quite enough—the Test remains a living issue in almost all discussions of AI if only because Turing said something definite, something that one can get one's teeth into, whether to applaud or attack. No one else has proposed any specific criterion by which the success of AI can be measured, so the Test, even when found less than satisfactory, holds the field for want of alternatives—as they say, you can't beat something with nothing. The promises and predictions made by virtually all other AI champions are too nebulous to get a grip on; all they usually tell us is that great progress has already been made, even more will soon be made, and wonderful things, things that will change human life radically, are on the way. In comparison with these outbursts of cheerleading and self-congratulation, Turing's proposal is a model of definite, concrete specificity (judged by standards more rigorous than those of the AI debate, of course, the Test is shoddy—riddled with ambiguities, arbitrary definitions, and unsupported assertions).

If progress toward the Test is no longer the measure of progress in AI, in fact, it is difficult to say just what AI is. The computer pioneer Maurice V. Wilkes, himself a winner of the Turing Award, has written[4], "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

[4] Wilkes 1992

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."

Wilkes' negative conclusion evoked several letters of rebuttal from AI workers, including one[5] from Patrick J. Hayes, then president of the American Association for Artificial Intelligence. As is traditional in such matters, these letters are strong on indignation and scorn, less so in citing specific achievements and arguments that show Wilkes was wrong, or in putting forward any alternative to Turing's as a goal for AI. Hayes does not, in his letter, 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 Turing: ignore his specific proposal even while donning his mantle to cover your own nakedness.

As Wilkes pointed out, no progress had been made by 1992 toward passing the Test (nor has any been made since), and most AI workers are not even aiming at that objective any more. In the absence of any generally accepted alternative goal, this has made it practically impossible to say what is and what is not AI. In practice, any new software that comes out of an institution with "AI" in its title, or 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 programs.

A typical complaint, from Martha Pollack, professor at the AI Laboratory of the University of Michigan, and executive editor of the *Journal of Artificial Intelligence Research*: "It's a crazy position to be in. 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."[6] Professor Pollack is right; that's just what we—we educated observers of the computing scene—do; and in so doing, we show that we're good observers. The AI community will continue to suffer such frustration so long as AI means 'anything done by people who call themselves AI workers.' Since there is no generally accepted definition for "AI", we outside the AI world just use our common sense in deciding what is worthy of that rather pretentious name, and most of us have decided that very little if anything that we've seen so far is worthy.

One AI worker who believes he has avoided the problem that the Test poses is Douglas Lenat, former professor of computer science at Stanford, and founder and president of Cycorp, who said[7] "The Turing test is a red herring. Anthropomorphizing a computer program isn't a useful goal." Since 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, it is not clear why his own project is not to be 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, but to make it actually possess the "common sense" that the Test computer only pretends to have.

[5] Hayes 1992

[6] Pollack 2003, page 3.

[7] Lenat 2001, page 82.

Another computer scientist, Peter Wegner of Brown University, avoids the problem in a refreshingly candid, even rather startling way: considering whether, in the light of Searle's Chinese Room thought experiment, the computer can be said to understand, he wrote[8],

> The bottom line is that we can answer the question 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 is a point of view that brushes aside both Turing and his critics; *he* is told that his supposedly tough, no-nonsense operational approach to AI is really just another fuzzy-minded, metaphysical piece of wool-gathering, and *they* that their answer is rejected because, true or false, it 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's just to keep the boys in the lab happy and productive.

But this kind of Machiavellian manipulativeness seldom works (Machiavelli himself couldn't make it work; he failed to win power or great influence). Those AI workers who are trying to pass the Test or otherwise demonstrate machine intelligence do so because they believe that success would mean they had really achieved AI, and would be widely honored for doing so. If they learn that the doctrine that machines can think is simply a carrot that is 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, the argument over whether passing the Test constitutes the achievement of machine understanding will not be settled—they might well feel that they were simply being lied to, and react in just the wrong way, from Wegner's 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 have a good effect on other people, you don't publish a letter announcing your plan. (If only Machiavelli had refrained from publishing *The Prince*, he might have gotten somewhere, but he preferred authorship to political success.)

A final—and vitally important—point to note is that many appeals to the Test are made without explicit mention of it; instead, we hear of *surprise* as the decisive consideration in determining whether a computer is thinking. Again and again one hears AI champions point out that the computer has done something that surprised us (or even surprised *them*), 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 what he had thought a human was instead a computer, or his surprise at learning that a computer has performed some feat that he had thought only humans could perform, is the very essence of the Test—unless the interrogator calls an unseen entity a human, and is then surprised to learn that it is actually a computer, the Test has—for Turing—failed. As a result, many AI workers have appealed to the criterion of surprise without being aware that they're following Turing; the influence of the Test is so pervasive that many who have never read Turing, and think they are working on entirely different lines from his, are nevertheless his epigones.

[8] Wegner 1987, page 7.

The Turing Test's Double Life in AI Circles

Professors Marvin Minsky of M.I.T. and John McCarthy of Stanford are considered the two founders of Artificial Intelligence as a formal discipline or research program, and both are still active as this is written. Minsky, writing a survey article[9] in 1961, gives classic examples of both the dropping of Turing's name to clinch an argument, and of the appeal to surprise. He defends the idea that computers may be capable of thinking with the words "…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, and then caps his argument with the words, "Turing ... gives a very knowledgeable discussion of such matters." He quotes no specific words of Turing's, however, because he is simply invoking Turing as an authority figure, not really referring to any particular point of his argument. But in 2003, Minsky expressed his disappointment and frustration at the lack of progress made by AI toward the goals that Turing had set, saying "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."[10]

Another eminent AI champion (like Minsky, a winner of the Turing Award, and a former president of the American Association for Artificial Intelligence), Raj Reddy of Carnegie Mellon University, thinks otherwise. In a paper[11] published in 1996, he begins with the usual bow to Turing, then says, "Since its inception, AI has made steady progress…" In elaborating on this progress, 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"; he nowhere mentions any attempt to pass the Test or do anything remotely like it. He has a short way with dissenters; referring to Hubert Dreyfus' well-known attack on AI, *What Computers Can't Do*, he writes:

> 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' 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—if they can't make the computer smarter, they can try to improve the ratio by making people seem dumber. (This tendency of AI enthusiasts to manipulate or misinterpret the human intelligence:machine intelligence ratio has been observed also by Jaron Lanier[12]: "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.") The idea seems to be 'if you can't beat 'em, claim they're not worth beating.' And Dreyfus, as a professor of philosophy, might not agree that he has "never done serious research on human intelligence"—nor does he think that computer intelligence is "in the future," he thinks it's in Cloud-Cuckoo-Land.

[9] Minsky 1961, page 27, note 38.

[10] Minsky 2003, pages 1-2.

[11] Reddy 1996.

[12] Lanier 2001, at page D6.

Another example of this line of argument: in a survey of the AI controversy[13], Judith Grabiner writes:

> Dreyfus' point that we do not really understand how people think, so we cannot model the process with computers, was turned against him by Yorick Wilks, who said that we were so far from understanding the processes people use in thinking that the only way we even know that other human beings think is by the Turing test.

Wilks (not to be confused with Maurice Wilkes, quoted earlier) offers us here a *reductio ad absurdam* of the argument, discussed earlier, that the Test simply applies to unknown entities the criterion we use in judging other humans: the 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?

To return to Reddy: he pays lip service to Turing, but his real guide and philosopher is Herbert Simon:

> 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."[14]

Whatever the value of the accomplishments that Simon and Reddy point to, they have nothing to do with Turing or the Test. As for intelligence, Simon's (and by inference Reddy's) definition of that property would make it very hard to deny intelligence to virtually any machine, implement, or tool; if anything that does something that, done by a human, would be called 'intelligent,' then what is *not* intelligent? If we have a gadget for de-veining shrimp, does it not do what, done by a human, would have to be called intelligent?

A final example (final only because there is a limit to patience and energy, not because further examples would be hard to find): in a paper[15] on SYNCHEM, a program that generates paths for the synthesis of chemical compounds, Gelernter and his colleagues claim that it should be given credit for true intelligence because "From the beginning SYNCHEM always performed above our reasonable expectations at each stage of its development." (How reasonable their expectations were if they always proved wrong, is a question they do not consider.) Here it is the surprise of the program designers themselves, rather than that of a disinterested observer or interrogator, that is offered as evidence of AI. One can only imagine how much more intelligence would have to be imputed to the program if its designers were even more forgetful or lacking in insight than they were.

[13] Grabiner 1986, at page 115.

[14] Reddy 1995.

[15] Gelernter 1977.

Conclusions to Part 1

We have now considered the classic paper by Turing in which the Test was first described; some variant readings of that paper; and the typical use to which the Test and Turing's reputation have been put by the AI community;. What conclusions may we draw after so much preparation? I suggest the following:

- Turing, for all his initial scorn for common usage in determining what 'thinking' is, wound up making his claim for machine intelligence depend on such common usage after all—but the common usage of 2000 rather than 1950. His intellectual honesty made him realize that if machines were to be credited with intelligence, it would have to be done by convincing the general public that what the computer was doing was thinking *in their own sense*, not is some special *ad hoc* sense contrived to support the claims of AI.

- No progress worth speaking of has been made to date in realizing Turing's goal, and few AI workers are attempting to reach it any more. The actual goal of AI workers today, whether they articulate it or not, is to come up with surprising new applications that will, if only briefly, astonish the public—and then to go back to the laboratory, when that astonishment has faded away, to produce another. A few of the AI pioneers who were inspired by Turing, like Minsky, are still yearning after Turing's dream, but they have little influence on what is called AI today.

- Turing's prestige is virtually the only part of his legacy that is of use to the AI community, and almost all members of that community exploit that prestige on public and ceremonial occasions, such as receiving awards, delivering keynote talks, assuming office in AI organizations and departments, and applying for grants and support from funding sources.

- It may be, as the letter from Wegner suggests, that some developers of advanced software find it important to believe that they are engaged in a project whose goal is machine intelligence, and one that has been making significant progress toward that goal. If so, the present study may be harmful to their morale, and hence to the progress of software; it may be prudent to keep it from working programmers.

APPENDIX 1: The Preliminary Version of the Test: Man or Woman?

Turing does not set the familiar man-or-machine scene immediately, but prepares his reader for it by first sketching an analogous one in which the interrogator questions an unseen human pair, one male and the other female, and tries to tell from their answers which is which. (The problem of identification in this case is further complicated by the rule that the woman is to tell the truth, while the man may lie at will.) Even Hodges, his greatly admiring biographer, wonders[16] why Turing offers this preliminary. It not only adds nothing of value, but in a way undercuts his main thesis—he will want us to agree that if an interrogator cannot tell the difference between the answers of a human and those of a machine, human and machine are intellectually the same, but he is not, presumably, suggesting that a similar failure to distinguish a man from a woman means that sexual differences are imaginary.

He confuses some readers by failing to make an explicit transition from the first to the second scenario, and

[16] Hodges 1983, page 415.

muddies the water further by failing to make clear, when passing from the man-or-woman test to the human-or-machine one, whether this is a switch to be made while a Test session is in progress, and if so, whether the interrogator is to be informed that this shift has been made. These questions should not have been troublesome to readers, because only one interpretation makes sense: if the switch is made during a session, or if the interrogator is allowed to think after the switch is made that his task is still to decide which is man and which woman, when in reality it is now to distinguish human (of whichever sex) from machine, whatever decision he comes to will be of no value for Test purposes. Accordingly, I conclude that Turing did not mean to swap entities in mid-session, which would only confuse the interrogator and make his reactions meaningless, but was simply careless, when writing his paper, in jumping silently from the first version to the next.[17]

Trivial though it is, this carelessness has caused much confusion among some readers, as is evident from the number who, in futile efforts to reconcile that preliminary version with the definitive one, have produced some truly bizarre interpretations of Turing’s proposal. These readers might have spared themselves many mental contortions by asking themselves how a test for distinguishing a man from a woman would be relevant to the question of whether machines can think, which is what Turing and the rest of the world thought his paper was about.

Among those confused was Peter Naur, a computer scientist of sufficient distinction to have been considered for the Turing Award: he takes the first version of the Test—the man-or-woman scenario—for the main event rather than a preliminary, saying[18]:

> This [closer analysis of the imitation game] is called for in part by the seemingly widespread misunderstanding that the game requires the interrogator to detect whether the respondent is a human being or a machine, while in fact Turing's interrogator faces two respondents and has to detect their sexes.

When Turing switches to the real Test, in which the interrogator is to distinguish human from machine, Naur suggests that Turing's purpose in introducing the preliminary version is to confuse the interrogator into thinking that his task is still to distinguish man from woman, even after a computer has been substituted for one of them[19] :

> One may suspect that Turing's motivation for his choice of the forms of the game was that he would consider the simpler form [in which the interrogator decides whether a single interlocutor is human or mechanical] too difficult for the machine. In fact, by the form of his game Turing will turn the interrogator's attention away from the real issue, the difference between man and machine, toward a pseudo issue, the sex difference.

Another gem of misreading comes from Harry M. Collins[20], a professor of sociology at the University of Bath:

[17] Professor Maurice Wilkes (Wilkes 1985, p. 197) writes that Turing told him, in a letter, that he was “not very pleased with [his paper]”; unfortunately, Turing did not make it clear just what aspect of the paper he was dissatisfied with.

[18] Naur 1986, page 176.

[19] Naur 1986, page 183.

[20] Collins 1990, page 181.

> The Turing Test looks simple, but has hidden subtleties. The first thing to notice is that the machine is to mimic a man who is imitating a woman—not a man being himself.

My own view is that Turing felt that the real Test, involving as it did a conversation between a man and a machine, would seem so strange to the reader of 1950 as to require an introductory example peopled with more familiar human characters. He accordingly sketched such an example, but took no care to point out that it was a merely introductory example, meant only to prepare the reader to understand the fully described man-or-computer test to which the rest of his paper was devoted. (It doesn't help that reader, though, that this scientist who so scorned the imprecision of words like 'thinking' sometimes uses 'man' to mean 'male, as distinct from female' and sometimes 'human, as distinct from machine.') In any case, little more is said in his paper about the man-or-woman test; it appears there just long enough to cause some confusion among hasty readers—at page 442, he tries very briefly to relate the preliminary version of the Test to the fully developed, familiar one, gets into something of a muddle, and then mercifully drops it for good.

APPENDIX 2: On the definition of key terms: *thinking, mind, intelligence* and so on.

Some readers, especially those trained in mathematical and engineering disciplines, will be disturbed by the use here of some undefined or at least imperfectly defined terms, the chief examples being those just given in the heading. How can we discuss a topic when the key terms are not even well defined, they will protest. There is a short, specific answer to that question, and a somewhat longer, general one; I will offer both, since some find the specific more satisfactory, some the general.

The specific answer is that in the absence of a universally accepted formal definition of those terms, we must work with their common senses—this is what Turing himself did, after expressing his mathematician's disgust at having to do so. And much to his credit, he fully accepted that necessity; unlike many of his later followers, he did not shamelessly ask us to adopt an *ad hoc* definition so that the computer could be said to be 'thinking' even though it was not doing what we normally mean by that term. Sophisticated defenders of AI—see (Moore 2003), *passim*—often, in their desperate effort to save the appearances, suggest that while machines may not think in the same sense that humans do, they may think in some other way—their own special machine way, presumably—so that refusal to grant that they're thinking is a kind of cultural chauvinism, almost a violation of their civil rights. This suggestion is to be rejected. *Thinking* is, by decree of common usage, what human minds do; if machines do something else, then what they do, however surprising and wonderful, is not thinking. As I noted earlier, Turing at least tacitly accepted that thinking was something we all recognize when we see it, even if we cannot come up with a precise definition of it, and stipulated that the triumph of AI would be signaled by the readiness of the educated public to extend that term to include what the computer was doing.

The more general answer to those disturbed by discussions carried on in the absence of formal definitions is that an insistence on starting a discussion of any profound and speculative subject matter with such definitions is fatal even to informal conversation, let alone serious investigations and debates. For thousands of years, we in the Western world (and perhaps elsewhere) have been discussing Justice and Beauty and Truth and a few other such subjects, not only without preliminary agreement on the meaning of the capitalized terms, but without arriving, even after millennia of consideration, at rigorous and universally agreed-on definitions of them. We're able to do so, to the scandal of the mathematically inclined, because those subjects are ones on which we have common experience and common intuitions. In most cases we can use what philosophers call ostensive definitions for them—we can't rigorously define them, but we can point to examples that almost all of us can agree on. It would indeed be good to have formal definitions for these important concepts—and it's in the hope of arriving at them that we have been discussing them for so long: precise definitions of profound concepts are not what you begin your discussions with, they're what you hope to end up with. In the meantime, we do the best we can with what we've got, and what we've got is good enough to be getting on with.

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: NewEessays 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?”], vol. IV, pp. 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.