of US$100,000 for the prize money, along with some additional funds from his company, Crown Industries, to help with expenses. The quest for the thinking computer had begun. In this article, I will summarize some of the difficult issues that were debated in nearly 2 years of planning that preceded the first real-time competition. I will then describe that first event, which took place on November 8, 1991, at The Computer Museum in Boston and offer a summary of some of the data generated by that event. Finally, I will speculate about the future of the competition – now an annual event, as Hugh envisioned – and about its significance to the AI community.

Planning for the event was supervised by a special committee, first chaired by I. Bernard Cohen, an eminent historian of science who had long been interested in the history of computing machines. Other members included myself, Daniel C. Dennett of Tufts University, Harry R. Lewis of the Aiken Computation Laboratory at Harvard, H. M. Parsons of HumRRO, W. V. Quine of Harvard, and Joseph Weizenbaum of MIT. Allen Newell of Carnegie-Mellon served as an advisor, as did Hugh Loebner. After the first year of meetings, which began in January of 1990, Daniel Dennett succeeded I. Bernard Cohen as chair. The committee met every month or two for 2 or 3 h at a time, and subcommittees studied certain issues in between committee meetings. I think it is safe to say that none of us knew what we were getting into. The intricacies of setting up a real Turing Test that would ultimately yield a legitimate winner were enormous. Small points were occasionally debated for months without clear resolution.

In his original 1950 proposal the English mathematician Alan M. Turing proposed a variation on a simple parlor game as a means for identifying a machine that can think: A human judge interacts with two computer terminals, one controlled by a computer and the other by a person, but the judge does not know which is which. If, after a prolonged conversation at each terminal, the judge cannot tell the difference, we would have to say, asserted Turing, that in some sense the computer is thinking. Computers barely existed in Turing's day, but, somehow, he saw the future with uncanny clarity: By the end of the century, he said, an "average interrogator" could be fooled most of the time for 5 min or so.

After much debate, the Loebner Prize Committee ultimately rejected Turing's simple two-terminal design in favor of one that is more discriminating and less problematic. The two-terminal design is troublesome for several reasons, among them: The design presumes that the hidden human – the human "confederate", to use the language of the social sciences – is evenly matched to the computer. Matching becomes especially critical if several computers are competing. Each must be paired with a comparable human so that the computers can ultimately be compared fairly to each other. We eventually concluded that we could not guarantee a fair contest if we were faced with such a requirement. No amount of pretesting of machines and confederates could assure adequate matching. The two-terminal design also makes it difficult to rank computer entrants. After all, they were only competing against their respective confederates, not against each other.

We developed a multiterminal design to eliminate these problems: approximately ten judges are faced with an equal number of terminals. The judges are told

that at least two of the terminals are controlled by computers and at least two by people. Again, the judges do not know which terminal is which. Each judge spends about 15 min at each terminal and then scores the terminals according to how human-like each exchange seemed to be. Positions are switched in a pseudorandom sequence. Thus, the terminals are compared to each other and to the confederates, all in one simple design.

Other advantages of this design became evident when we began to grapple with scoring issues. We spent months researching, exploring, and rejecting various rating and confidence measures commonly used in the social sciences. I programmed several of them and ran simulations of contest outcomes. The results were disappointing for reasons we could not have anticipated. Turing's brilliant paper had not gone far enough to yield practical procedures. In fact, we realized only slowly that his paper had not even specified an outcome that could be interpreted meaningfully. A binary decision by a single judge would hardly be adequate for awarding a large cash prize – and, in effect, for declaring the existence of a significant new breed of intelligent entities. Would some proportion of ten binary decisions be enough? How about 100 decisions? What, in fact, would it take to say that a computer's performance was indistinguishable from a person's?

A conceptual breakthrough came only after we hit upon a simple scoring method. (R. Duncan Luce, a mathematical psychologist at the University of California, Irvine, was especially helpful at this juncture.) The point is worth emphasizing: The scoring method came first, and some clear thinking followed. The method was simply to have each judge rank the terminals according to how human-like the exchanges were. The computer with the highest median rank wins that year's prize; thus, we are guaranteed a winner each year. We also ask the judges to draw a line between terminals he or she judged to be controlled by humans and those he or she judged to be controlled by computers; thus, we have a simple record of errors made by individual judges. This record does not affect the scoring, but it is well worth preserving. Finally, if the median rank of the winning computer equals or exceeds the median rank of a human confederate, *that computer will have passed (a modern variant of) the Turing Test*. It is worth quoting part of a memo I wrote to the committee in May of 1991 regarding this simple approach to scoring:

Advantages of this method

1. It is simple. The press will understand it.
2. It yields a winning computer entrant.
3. It provides a simple, reasonable criterion for passing the Turing Test: When the [median] rank of a computer system equals or exceeds the [median] rank of a human confederate, the computer has passed.
4. It preserves binary judgment errors on the part of individual judges. It will reveal when a judge misclassifies a computer as a human.
5. It avoids computational problems that binary judgments alone might create. A misclassified computer would create missing data, for example.
6. It avoids theoretical and practical problems associated with rating scales.

Other issues were also challenging. We were obsessed for months with what we called "the buffering problem". Should we allow entrants to simulate human typing foibles? Some of us – most notably, Joseph Weizenbaum – said that such simulations were trivial and irrelevant, but we ultimately agreed to leave this up to the programmers, at least for the first contest. One could send messages in a burst ("burst mode") or character-by-character ("chat mode"), complete with misspellings, destructive backspaces, and so on. This meant that we had to have at least one of our confederates communicating in burst mode and at least one in chat mode. Allowing this variability might teach us something, we speculated.

We knew that an open-ended test – one in which judges could type anything about any topic – would be a disaster. Language processing is still crude, and, even if it were not, the "knowledge explosion" problem would mean certain defeat for any computer within a very short time. There is simply too much to know, and computers know very little. We settled, painfully, on a restricted test: Next to each terminal, a topic would be posted and the entrants and confederates would have to communicate on that one topic only. Judges would be instructed to restrict their communications to that one topic, and programmers would be advised to protect their programs from off-topic questions or comments. Entrants could pick their own topics, and the committee would work with confederates to choose the confederates' topics. Moreover, we eventually realized that the topics would have to be "ordinary." Expert systems – those specializing in moon rocks or the cardiovascular system, for example – would be too easy to identify as computers. In an attempt to keep both the confederates and judges honest and on-task, we also decided to recruit referees to monitor both the confederates and the judges throughout the contest.

This sounds simple enough, but we knew we would have trouble with the topic restriction, and we were still debating the matter the evening before the first contest. If the posted topic is "clothing", for example, could the judge ask, "What type of clothing does Michael Jordan wear?" Is that fair, or is that a sneaky way to see if the terminal can talk about basketball (in which case it is probably controlled by a human)?

Should we allow the judges to be aggressive? Should graduate students in computer science be allowed to serve? Again, many stimulating and frustrating debates took place. Both to be true to the spirit of Turing's proposal and to assure some interesting and nontrivial exchanges, we decided that we would select a diverse group of bright judges who had little or no knowledge of AI or computer science. We attracted candidates through newspaper ads that said little other than that one had to have typing skills.

In short – and I am only scratching the surface here – we took great pains to protect the computers. We felt that in the early years of the contest, such protection would be essential. Allen Newell was especially insistent on this point. Computers are just too inept at this point to fool anyone for very long. At least that was our thinking. Perhaps every fifth year or so, we said, we would hold an open-ended test – one with no topic restriction. Most of us felt that the computers would be trounced in such a test – perhaps for decades to come.

We agreed that the winner of a restricted test would receive a small cash award and bronze medal and that the cash award would be increased each year. If, during

an unrestricted test, a computer entrant matched or equaled the median score of a human, the full cash prize would be awarded, and the contest would be abolished.

Other issues, too numerous to explore here, were also discussed: How could we assure honesty among the entrants? After all, we were dealing with a profession known widely for its pranks. Should the confederates pretend to be computers or simply communicate naturally? We opted for the latter, consistent with Turing. Should we employ children as confederates in the early years? Should professional typists do the judges' typing? How aggressive should the referees be in limiting replies? Should entrants be required to show us their code or even to make it public? We said no; we did not want to discourage submissions of programs with possible commercial value.

Our final design was closely analogous to the classic double-blind procedure used in experimental research: The prize committee members were the "investigators". We knew which terminal was which, and we selected the judges, confederates, and referees. The referees were analogous to "experimenters". They handled the judges and confederates during the contest. They were experts in computer science or related fields, but they did not know which terminal was which. The judges were analogous to "subjects". They did not know which terminal was which, and they were being handled by people with the same lack of knowledge.

Over time, formal rules were developed expressing these ideas. Announcements were made to the press, and funding for the first contest was secured from the Sloan Foundation and the National Science Foundation. Technical details for running the show were coordinated with The Computer Museum in Boston, which agreed to host the contest. Applications were screened in the summer of 1991, and six finalists were selected by the prize committee in September. Confederates, judges, and referees were selected in October.

1.2 The 1991 Competition

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 and conjectured while they watched eight conversations unfold in real time on large screens. A moderator – A. K. Dewdney of *Scientific American* – roamed the auditorium with a cordless microphone, interviewing and commenting. Four film crews and dozens of reporters and photographers bristled and flashed and shouldered each other to get the best angles. Food flowed all day.

The judges and terminals were set up in a roped-off area outside the auditorium. You could view them directly behind the ropes if the journalists would let you by – or on a large screen setup in the auditorium. Groups of about 20 chairs were positioned around each screen in the auditorium, and the moderator encouraged the members of the audience to move to another screen whenever the judges switched terminals – about once every 13 min. The confederates were stationed in a remote area of the museum behind a guarded door. Dennett and others made some opening remarks midday, and the real-time competition itself took about 3 h in the afternoon.

Some technical problems got things off to a frustrating start. Two of the ten terminals would not function properly, knocking out two of the four confederates we had planned to use. We went forward with the six entries and two confederates (recall that everyone except the members of the prize committee and some technicians knew simply that at least two of the eight working terminals were controlled by computers and that at least two were controlled by people).

Our ten judges were diverse, bright, and without expertise in computer science, just as we had planned. Seven were female, and three were male. They ranged in age from 19 to 48; six were in their 20s, one was in her 30s, and two were in their 40s. All were proficient in English, yet one was born in the Philippines, and three had spent most of their lives in foreign countries: Ethiopia, Australia, and West Africa, respectively. Seven had B.A.'s, two of the seven had masters degrees, and two were currently in graduate school. One was high school educated, and two were current college students. Four claimed to have some background in psychology or philosophy. Those not currently in school had a variety of professions: auto mechanic, graphic designer, hypnotist, teacher, and so on.

All communications were by modem. Entrants communicated from PCs or mainframes located throughout the country, and even the confederates' communications were looped through phones outside the building. The judges and audience were told this so that transmission errors would be discounted as clues about the identity of the terminals.

As we expected, the computers performed relatively poorly for the most part. Consider the following exchange, for example, between Judge 2 and Terminal 10 (the topic was "Dry Martinis"):

01:36:42 Judge 2: Of what is gin made?
01:36:52 Terminal 10: I'M NOT SURE.
01:36:58 Judge 2: OK Tell me any thin else you know of about martinis.
01:37:21 Terminal 10: I REALLY DON'T KNOW.
01:37:28 Judge 2: What foods do martinis go with.
01:37:44 Terminal 10: COULD YOU ASK THE QUESTION A DIFFERENT WAY?
01:37:52 Judge 2: Why?
01:37:58 Terminal 10: I REALLY CAN'T SAY. OH WELL.
01:38:04 Judge 2: Ok with what foods shoul onew drink a martini.
01:38:28 I REALLY CAN'T SAY. OH WELL.[1]

Five consecutive replies were unresponsive to the judge's questions; the last two replies were repeated verbatim, a dead giveaway. Frequent typing errors by the judge may have had some effect. Unfortunately, this pattern was common. Consider this exchange (the topic was "Second Grade Topics"):

[1] Copyright, 1991, Cambridge Center for Behavioral Studies. Reprinted by permission. Note that the quotes from the 1991 Loebner Competition transcripts are included here character-for-character as they occurred during the contest.

02:53:22 Judge 4: What do you study at school?
02:53:43 Terminal 3: I don't know.
02:53:51 Judge 4: What do you like most about school?
02:54:16 Terminal 3: As I said – I don't know.
02:54:25 Judge 4: What topic do you know the most about?
02:54:42 Terminal 3: I don't know.
02:54:51 Judge 4: What kind of gymnastics do you do?
02:55:10 Terminal 4: As I said – I don't know.
02:55:16 Judge 4: I thought you were a gymnast.
02:55:27 Terminal 3: Learning is fun.[2]

The winning entry, "Whimsical Conversation", submitted by an independent programmer, Joseph Weintraub of New York City, performed substantially better than the others. It seemed, fairly often, to make contact with the judge's comments and queries. Moreover, it simulated the rhythms and corrections of human typing nearly perfectly. Consider the following exchange:

01:32:25 Judge 6: Hello, earth person!
01:33:03 Terminal 5: Hello there judge, are you ready to have some fun? How ya doin, judge? Hello – how's the weather up there today?
01:33:35 Judge 6: Dark clouds with occasional bursts of sunshine (or rather camera flashes).
01:34:45 Terminal 5: The trouble with the rat race is that even if you win you're still a rat. Whimsy can be flimsy, right? How important is that to you?
01:35:26 Judge 6: I can't say I understand the question…but the whimsy of a poet is one thing and the whimsy of most of us is another.
01:36:51 Terminal 5: How do you know you can t say you understand the question…but the whimsy of a poet is one thing and the whimsy of most of us is another.
01:37:37 Judge 6: Are you repeating me just for laughs?
01:38:04 Terminal 5: I'm sorry, I do sometimes repeat myself…perhaps it's because i can't find my electronic organizer![3]

In this exchange, the program reflects back one response wholesale, just as Weizenbaum's ELIZA did decades ago. But it recovers fairly impressively from the blunder with an apology, simulating typing foibles all the while.

The overall result was fairly predictable, but there were also some surprises. The two human confederates, whose topics were, respectively, "Women's Clothing" and "Shakespeare's Plays", scored in first and second place, respectively. Not far behind was Weintraub's whimsical program; thus, Weintraub won the 1991 contest. He was awarded a bronze medal (bearing a likeness of Turing on the front and Loebner on the back) and a check for US$1,500. Weintraub, who tended his equipment in New York

[2] Copyright, 1991, Cambridge Center for Behavioral Studies. Reprinted by permission.

[3] Copyright, 1991, Cambridge Center for Behavioral Studies. Reprinted by permission.

during the contest, took questions from the press by speakerphone after the announcement was made. An entry from Peter Colby and Kenneth Colby, of PARRY fame, scored just below Weintraub's. In third place among the computers was a submission from Michael Mauldin of the Center for Machine Translation at Carnegie-Mellon.

The surprises were notable. First, five of the ten judges (Judges 2, 3, 4, 9, and 10) mistook Weintraub's program for a human. Judge 3 rated it above one human confederate (Terminal 1), and Judge 10 rated it above both human confederates (Terminal 1 and 4). The Colbys' program ("Problems in Romantic Relationships") was mistaken for a person by two of the judges, and another program ("Second Grade School Topics") was misclassified by one judge. Perhaps even more remarkable, Cynthia Clay, the human confederate whose topic was Shakespeare's plays, was mistaken for a computer by three judges. Judge 10 placed her lower in rank than a computer (Terminal 5), and two judges (1 and 5) placed her lower in rank than two computers. Note that Cynthia's responses were buffered; that is, her responses occurred in a burst, suggesting computer output. Furthermore, she was an expert on Shakespeare, and she often quoted lengthy passages verbatim. Several judges remarked that her replies seemed too expert to be human.

As Turing anticipated, the contest tells us as much, or perhaps even more, about our failings as judges as it does about the failings of computers. People's preconceptions about the limits of computers – and of people – strongly biases their judgments.

At the start of the contest, members of the audience were given forms to help them do their own judging. The forms asked for basic demographic information, as well. Seventy-seven forms were collected at the end of the contest. Based on this sample, audience ratings may be summarized as follows:

- Audience rankings matched those of the judges, and the rankings of those who claimed expertise in computer science did not differ substantially from the rankings of those who did not claim such expertise. For the 66 respondents who ranked all eight terminals, Terminals 1 and 4 were again ranked first and second, respectively, and Terminal 5 ("Whimsical Conversation") was again ranked third. Curiously, the other five terminals were ranked equally; that is, they were, on average, indiscriminable.
- 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. For example, the winning computer, "Whimsical Conversation", was labeled a human by only five out of the 77 respondents (ten did not reply, leaving 61 correct classifications), and Cynthia Clay (Terminal 4) was misclassified as a computer by only five respondents (seven did not reply, leaving 65 correct classifications). The other human confederate, although ranked higher by both judges and audience, was misclassified at nearly the same rate. Once again, expertise in computer science had so systematic effect.

With James C. Pabelico, a student at the University of California, San Diego, I attempted a search for objective factors that could predict the judges' ratings – in other words, that measured the apparent intelligence of an entity communicating

over a computer terminal. Simplistic factors such as word length, sentence length, number of syllables per word, and number of prepositions were not predictive. Neither were various measures of readability, such as Flesch Reading Ease, Gunning's Fog Index, and Flesch-Kincaid Grade Level. The Weintraub and Colby programs, for example, had Flesch-Kincaid Grade Levels of 2 and 6, respectively; the two humans had scores of 3 and 4.

So why did Weintraub's program win? And how did it fool half the judges into thinking it was a person? Unfortunately, it may have won for the wrong reasons. It was the only program, first of all, that simulated human typing foibles well. Another program simulated human typing so poorly that it was instantly recognizable as a computer on that basis alone; no human could possibly have typed the way it was typing.

Perhaps more notable, Weintraub's program simulated a very curious kind of person: the jester. We allow great latitude when conversing with jesters; incomprehensible, irrelevant responses are to be expected. We are equally tolerant of young children, developmentally disabled individuals, psychotic patients, head-injured individuals, and absentminded professors. Weintraub's program may have succeeded simply because his terminal was labeled "whimsical conversation". The prize committee discussed this possibility, and considerable concern was expressed. In 1992, the committee favored programs that had clear subject matters.

1.3 Speculations

I believe that when a computer passes an unrestricted Turing Test, humankind will be changed forever. From that day on, computers will be companions to the human race – and extraordinary companions indeed. For starters, they will be efficient, fast, natural-language interfaces to virtually all knowledge. They will be able to access and evaluate enormous amounts of data on an ongoing basis and to discuss the results with us in terms we can understand. They will think efficiently 24 h a day, and they will have more patience than any saint.

Thinking computers will also have new roles to play in real-time control. Everything from vacuum cleaners to power plants has a dumb computer in it these days; some day, smart computers will share in the decision-making. Over networks or even airwaves, thinking computers will be able to coordinate events worldwide in a way humans never could.

Thinking computers will be a new race, a sentient companion to our own. When a computer finally passes the Turing Test, will we have the right to turn it off? Who should get the prize money – the programmer or the computer? Can we say that such a machine is "self-aware"? Should we give it the right to vote? Should it pay taxes? If you doubt the significance of these issues, consider the possibility that someday soon *you will have to argue them with a computer*. If you refuse to talk to the machine, you will be like the judges in *Planet of the Apes* who refused to allow the talking human to speak in court because, according to the religious dogma of the planet, humans were incapable of speech.

The Internet has added another dimension to the Turing Test. It is only natural that we think of the Internet as a tool that serves humanity, but someday sentient computers will undoubtedly see it as their natural home. It is not inconceivable that within milliseconds of achieving sentience, that first remarkable entity will dive into the Internet to learn, to grow, and to assure its own survival. Once in the Net, it will be impossible to disable, and its subsequent effect on the human race is anyone's guess. Internet II – the UltraNet now functioning at some major institutions – will provide an even larger nest.

Some people, including members of the original Loebner Prize Committee, believe that computers will never cross this threshold. But 40 years of reading science fiction novels, 35 years of programming, and nearly 30 years of studying psychology has me convinced that 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.

Loebner himself was open-minded when the contest was set in motion, perhaps even skeptical. But he also offered the most outrageous prediction of all. Some day, he said, when the human race is long dead, a mechanical race will remember us as deities. After all, we are the creators, are we not?

I think the quest for the thinking computer will eventually become as intense as the quest for the Holy Grail. The stakes are similar. A program that passes the Turing Test will be worth a fortune. Just ask it.

Even within the first year, committee members talked about expanding the contest at some point to include Turing-like tests of robotics, pattern recognition, and speech recognition and synthesis. In a week-long tournament, computers would compete against people in each domain. The ultimate outcome? Well, have you seen or read *I, Robot*?

I may be overly optimistic about the future of artificial intelligence. Certainly, several of my colleagues, much older and, by definition, much wiser than I, have told me so. But we will all have fun exploring the possibilities – even if, someday, and for reasons I cannot now imagine, we are forced to conclude that the Turing Test cannot be passed.

Afterword This is a slightly modified version of an article first published in 1992 in AI Magazine. The Loebner Prize Committee continued to direct the event for the first four contests. During the planning for the fifth contest, Hugh Loebner asked the committee members to change the rules in ways they found objectionable, and the committee disbanded. The contest has been held every year since, however, under Loebner's direction. The 16th Annual Loebner Prize Competition was held on September 17, 2006, at University College London. Four computer programs participated, and the winner was a program named "Joan," created by Rollo Carpenter of Icogno Ltd.

Reference

Turing, A. M., 1950, Computing machinery and intelligence, *Mind* **50**(236): 433–460.

Chapter 2
Alan Turing and the Turing Test

Andrew Hodges

Abstract The study of Alan Turing's life and work shows how the origin of the Turing Test lies in Turing's formulation of the concept of computability and the question of its limits.

Keywords Alan Turing, Turing machine, computability

2.1 Introduction

Alan Mathison Turing
Born: 23 June 1912, Paddington, London
Died: 7 June 1954, Wilmslow, Cheshire

The short and extraordinary life of the British mathematician Alan Turing embraces the foundations of mathematics, the origin of the computer, the secret cryptological war against Nazi Germany, and much more. For the modern public, his name is perhaps most strongly associated with yet another context: the philosophy of Mind and Artificial Intelligence (AI). Specifically, he is immortalised in the "Turing Test" for intelligence, which Turing himself called "the imitation game".

The famous test appeared in Turing's paper, "Computing machinery and intelligence", published in October 1950 in the philosophical journal *Mind* (Turing, 1950). Turing was then employed at Manchester University, where the world's first stored-program computer had been working since June 1948. Turing had not been appointed to produce philosophical papers; his primary function was to create and manage the first software for the computer. However, he had also continued his research in mathematics, and had been drawn into discussion with the scientific philosopher Michael Polanyi, who encouraged him to publish his views on what Turing called "intelligent machinery" (Turing, 1948). The point of this 1950 paper,

Wadham College, University of Oxford

R. Epstein et al. (eds.), *Parsing the Turing Test.*

of which the imitation game was only a part, was to argue that intelligence of a human level could be evinced by a suitably programmed computer. The imitation game lent definiteness to the idea of being as intelligent as a human being.

Turing's 1950 paper did not arise in isolation, and the purpose of this biographical sketch is to set Turing's test in the context of his life and work. The 1950 paper was an important summary of his views for the general philosophical public, but he had been developing those views for many years. It would also be a mistake to think of Turing as a mathematician making a detached comment on the potential of computers. He was very fully engaged in the development of modern computer science, both in theory and in practice.

2.2 The Turing Machine

Indeed the 1950 paper had itself an important autobiographical element, although Turing did not emphasise its personal aspect. Much of the early part of the paper involved an exposition of the concept of computability. The definition of computability was Turing's own great achievement of the pre-war period, when he was a young Fellow of King's College, Cambridge University. In 1936, when he was only 23, he announced the definition of what soon became known as the *Turing machine*, thus giving a precise and convincing mathematical definition of an "effective method". In Turing's paper "On computable numbers, with an application to the Entscheidungsproblem" (Turing, 1936), he gave a discussion of his definition, arguing convincingly that it encompassed the most general possible process for computing a number.

By so doing, he satisfactorily answered Hilbert's decision problem for the provability of mathematical theorems. The paper did more: it also defined the concept of a *universal* Turing machine and hence the principle of the modern stored-program computer. The universal machine was another important element of the theory Turing needed to explain in his 1950 paper: the point is that all Turing machines can be thought of as programs for a single universal machine. Most striking, perhaps, is that the Turing machine, formulated 14 years before the "Turing Test", was also based on a principle of imitation. The "machine" was modelled by considering what a human being could do when following a definite method. According to Turing's argument, it encompassed everything that could be done by a human calculator with finite equipment, but allowed an unlimited supply of paper to write on (formalised into a paper tape marked off in squares) and unlimited time.

The basis in human calculation was emphasised in Turing's arguments. The "squares" of the Turing machine "tape", for instance, originated in Turing's explanation as the squares of a child's exercise book. The atomic operations of scanning, marking, erasing, and moving to left and right were likewise related to human actions. Most importantly, the finitely many "configurations" of a Turing machine were related to the finite number of states of mind, or finite memory, of a human calculator. This very bold appeal to modelling "states of mind" by states of a machine seems already to anticipate the thesis of machine intelligence in 1950.

Should we then say that Turing in 1950 was only restating the implications of what he had claimed in 1936?

A simple reading of the story would support this view. It might be argued that from 1936 onwards, Turing steadfastly sought and found ways to implement his theory in practice. In 1937 Turing began a practical interest in electromagnetic relays to implement logical operations, an interest very different from anything expected of a Cambridge pure mathematician (Hodges, 1983). After 1939, this interest turned into one of immense practical importance. Turing's ingenious logic was translated into working electromagnetic machinery at Bletchley Park, Buckinghamshire, the centre of British code-breaking operations. His algorithm for breaking the Enigma-ciphered German messages, as embodied in the British "Bombe", lay at the centre of its success. He personally headed the work of deciphering German naval messages, and led the development of methods of astonishing efficiency. He was introduced to American work at the highest level, and to the most advanced electronic technology. In this way he learned that electronic storage and electronic circuits could make an effective and fast practical version of the "paper tape" and configurations needed for a universal machine. He learned electronics for himself, again a highly unconventional step. Turing emerged from the war in 1945 full of enthusiasm for engineering a practical version of the universal machine – in modern parlance a stored-program computer. By a remarkable sequence of events, Turing was taken on by the National Physical Laboratory, London, with exactly this commission. His plan was submitted in March 1946 (Turing, 1946). As well as pioneering key ideas in computer hardware and software design, it mentioned the idea of the machine showing "intelligence", with chess-playing as a paradigm. This germ then rapidly developed into the program set out in the 1950 paper.

2.3 Intelligence and Intuition

Although basically correct, a subtle adjustment to this basic story is required, and it is one that casts light on the structure and content of the 1950 paper. A reading of that paper will show that Turing was highly aware of the natural objection that machines cannot do those things which are by nature non-mechanical: those human actions that require initiative, imagination, judgment, cause surprise, are liable to errors, and so forth. Much of his argument was directed to the claim that machines would, in fact, be capable of all these apparently non-mechanical things.

But there was no reflection of this claim in his 1936 work, "On computable numbers" (Turing, 1936). The "states of mind" that Turing considered were only those employed when the mind is occupied on performing a definite method or process. There was no reference to imagination or initiative in the "states of mind".

So we can ask: at what point in his biography did Turing adopt the idea that computable operations would encompass everything normally called "thinking" rather than only "definite methods"? Or equivalently, when did he first consider that operations, which are in fact the workings of predictable Turing machines, could nevertheless appear to have the characteristics of genuine intelligence?

Turing wrote very little about his own intellectual development, and his writings do not give a direct answer to this question. However, there are two important stages in his work not mentioned in the above account, which when considered in their context suggest a plausible answer: namely at some point after 1938 and probably in about 1941.

During the 2 years Turing spent at Princeton, from 1936 to 1938, he was investigating the logic of the *uncomputable*. Turing's exposition (Turing, 1939) described the "formulae, seen intuitively to be correct, but which the Gödel theorem shows are unprovable in the original system". In this pre-war period, Turing left open the possibility that the mind had an "intuitive" power of performing uncomputable steps beyond the scope of the Turing machine (a thesis, incidentally, that was always held by Gödel himself).

But in the period around 1941 when the immediate crisis of the Enigma problem was resolved, Turing began to discuss with colleagues at Bletchley Park the possibility of machines playing chess (Hodges, 1983). Chess-playing was in fact a civilian analogue of what they were doing in their secret military work, in which mechanical methods of great sophistication were outdoing some aspects of human intuition. It seems, taking the view as expressed in Hodges (1997, 2002), that Turing probably decided in about 1941 that the scope of computable operations was in fact sufficient to account for those mental operations apparently "non-mechanical" by the standards of ordinary language, and even the apparently uncomputable operations of truth recognition.

There was possibly another wartime influence on his development: Turing's general exposure to modern ideas such as the neural physiology of the brain and the behaviourist model of the mind. McCulloch and Pitts (1943) related their logical model of neurons to Turing's computability; Turing returned the compliment by referring to their work. Turing was developing the picture of the brain as *a finite discrete state machine*. In a sense this was only a small step from the "finitely many states of mind" of 1936. But it went further because Turing's postwar idea was that *all* mental functions of the brain could be accommodated in this model, and not just those of a mind following a definite rule. As we shall see, Turing framed an argument to explain how this mechanical picture of the brain could be reconciled with the counter-arguments from Gödel's theorem.

Very possibly it was this new conviction that made him so enthusiastic, in the closing stages of the war, about the prospect of building an electronic version of the universal machine. He was not highly motivated by building a computer to work out programmed mathematical operations. His interest was more in the nature of the mind. Informally, when speaking of his computer plans in 1945, he called them plans for "building a brain" (Hodges, 1983).

2.4 Intelligent Machinery

With this in mind, we can examine in more detail his very first written mention of "intelligent" machinery (Turing, 1946). One should first note how remarkable it was that Turing should put a speculative claim about intelligence in a purely technical,

practical report. However, this was entirely typical of his *modus operandi*. One should next appreciate that Turing always relished the paradox, even apparent contradiction in terms, involved in speaking of "intelligent" machinery. First he explained how the computer could be programmed to calculate chess moves. He continued, underlining the paradox:

> This … raises the question 'Can a machine play chess?' It could fairly easily be made to play a rather bad game. It would be bad because chess requires intelligence. We stated … that the machine should be treated as entirely without intelligence. There are indications however that it is possible to make the machine display intelligence at the risk of its making occasional serious mistakes. By following up this aspect the machine could probably be made to play very good chess.

This mysterious reference to "mistakes", which could have made no sense to anyone reading this report, was explained in a talk of February 1947 (Turing, 1947). Here the idea of "mistake-making" appeared in the context of the objection to the prospect of machine intelligence posed by Gödel's theorem. This objection (which appears as "The Mathematical Objection" in the 1950 paper) is that no Turing machine (i.e., computer program) can do as well as a human being. The human being can see the truth of mathematical assertions which cannot be proved by following rules based on formal axioms.

Turing's post-war argument (the point of view he probably arrived at in about 1941) is, however, that human beings do *not* reliably see the truth of such statements. Mathematicians, their brains being discrete state machines, can only employ an algorithm. Gödel's theorem tells us that no algorithm can coincide with truth-seeing in every case, and so the algorithm is bound to fail sometimes. But if it is accepted that the mathematician is not infallible, and will sometimes fail, it follows that machines – also implementing algorithms, and therefore also making mistakes – may do equally well. To illustrate the theme of doing equally well, Turing appealed to the concept of "fair play for machines". This concept was essentially the idea of the imitation game. The 1950 scenario merely added dramatic detail. Thus, the imitation game had its origins in the wartime debate in Turing's own mind about how to reconcile Gödel's theorem and the apparently non-mechanical actions of human minds with his discrete state machine model of the brain.

After 1947, Turing continued to a wider and more constructive discussion of how machines might perform apparently non-mechanical tasks: how completely unintelligent micro-operations might add up to intelligent processes. This investigation was presented in an internal report: "Intelligent Machinery", for the National Physical Laboratory (Turing, 1948). It was not published until 1968, but was in many ways the basis of his better-known and less technical 1950 exposition. One interesting feature of this 1948 report is its evidence of a wartime inspiration for his new ideas. Turing referred to images of the writer Dorothy Sayers, to illustrate the commonly accepted meaning of "mechanical" behaviour. The book he quoted was one he was reading at Bletchley Park in 1941. Turing also tellingly described 1940 as the date after which machines were no longer restricted to "extremely straightforward, possibly even to repetitious, jobs". He must have had his own Enigma-breaking Bombe, and other highly sophisticated code-breaking operations, in mind.

In this report, Turing characterised intelligence as requiring "discipline", which he identified with the programmability of a universal machine, plus a residue of "initiative". Initiative now played the role that "intuition" had done in 1938: mental actions apparently going beyond the scope of a "definite method". How was initiative to be found within the scope of computable operations, and so implemented on a computer?

Turing suggested various possibilities all based on *imitating* human brains: learning, teaching, training, and searching. From the outset of his design work in 1945, Turing had been enthusiastic for exploiting the feature of a stored-program computer that it allows for a program thus stored to be manipulated in the same way as data. These ideas took his enthusiasm further, by having the machine actively modify its own programs, to arrive at functions which had never been planned or envisaged by a programmer. Turing emphasised that at a more fundamental level the concept of "a machine changing its own instructions" was "really a nonsensical form of phraseology", but it was convenient. The upshot of his argument was that by one means or another, and probably using many methods in combination, a computer could be made to simulate the mental functions of human brains.

From a purely biographical point of view, it is remarkable that someone so original, and whose individual qualities had generally been stoutly resisted by his social environment, should arrive at the conclusion that creativity is a computable process, something that could be achieved by a computer. But it was where he was led by his guiding idea of the brain as a finite machine, whose operations must be computable however different they appeared from what people had hitherto thought of as "mechanical" in nature.

2.5 The Imitation of Mind

This 1948 work was the background to the 1950 paper, in which Turing made a more public claim than ever before that intelligence could be evinced by computing machinery: i.e., belonged to the realm of computable processes. It was also a more ambitious claim than ever, since by provocative forays into the world of the Arts with witty talk of Shakespeare and sonnets, Turing made it quite clear that he was not restricting "intelligence" to some special science-minded arena. The famous test, pitting human against machine in demonstrating intelligence, embodied the "fair play" announced in 1947. The setting of the test, however, with its remote text-based link, did have a further functional significance. It was supposed to give a way of distinguishing those things Turing considered relevant to intelligence, as opposed to other human faculties involving their many senses and physical actions. It is probably in drawing this distinction that Turing showed the least certainty, and this aspect of his paper has attracted the most criticism.

Returning, however, to Turing's central idea, it should be emphasised that Turing never imagined that the structure of the brain would resemble that of a computer, with a central control unit processing instructions serially. The crucial point here

lies in Turing's exposition of the universal machine concept (Turing, 1939). It follows from his argument that provided the operation of "thought" is equivalent to the operation of *some* discrete state machine, it can be simulated by a program run on a single, universal machine, i.e., a computer. Thus Turing threw all his emphasis on the development of what would now be called software, rather than on the engineering of a brain-like object.

This point can be further refined. Turing's description of computability in the 1950 paper was all based on the finite capabilities of real, finite machines, illustrated by an account of the Manchester computer as it then stood. His claim was that the simulation of thought did not even require the full scope of computable functions, only that infinitesimal fraction of them which could be run using only a finite amount of "tape". (As a technical point, Turing's description did not even mention the "tape". This is because a finite tape can be absorbed into the instruction table of a Turing machine, which in turn he identified with the storage of a computer such as the Manchester computer. This resulted in him rather confusingly describing the full gamut of computable processes as requiring an "infinite store". This is, however, just the unlimited supply of tape as prescribed in 1936, not an infinite instruction table.) He suggested a necessary storage capacity of 10^9 bits, which of course is far surpassed by modern personal computers.

A fortiori, there was no suggestion in this paper of anything beyond the scope of computability. There were three areas of Turing's discussion where mathematics beyond the computable was raised, but in each case the thrust of Turing's argument was that computable operations would suffice. One of these was the Gödel argument, actually rather more weakly addressed in this than in his earlier papers, but still concluding that it had no force. The second lay in Turing's discussion of "the continuity of the nervous system". He claimed that the brain's basis in continuous matter, rather than being a discrete machine, was again no argument against the computability of thought: a discrete system could approximate a continuous one as closely as desired. The third was the concept of randomness, which Turing introduced without any serious definition. His illustration used "the digits of π" as a random sequence, and this is par excellence a computable input.

In fact, Turing's exposition ran through two stages, reflecting what has been suggested above as his "1936" and "1941" stages of development. First came the concept of computable functions, thought of as planned instructions (Turing, 1950), and then followed the finite discrete state machine picture. However, as he argued, these differently pictured machines could alike be implemented on a universal machine, the computer. This same two-part structure came into his final constructive proposals for the development of machine intelligence. Turing imagined rule-based programming (rather like expert systems as later developed), but also the "child machine" learning for itself. Turing concluded by recommending that "both approaches should be tried": he never suggested a rigid dichotomy between top-down and bottom-up approaches, which was later to divide Artificial Intelligence research so deeply.

In summary, Turing was able to claim:

> I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning.

This prophecy of the power of the computable was, of course, to stimulate the Loebner Prize competitions as the dateline of 2000 approached.

2.6 After the Test

No account of Alan Turing would be complete without a mention of his last years. "Computing machinery and intelligence" was shot through with courtroom images of juries and trials; they were prophetic. Turing was arrested for his affair with a young Manchester man in 1952. All homosexual behaviour was then criminal. He was seriously disturbed by the punishment that ensued: his brain was "treated" with oestrogen. But his mind did not atrophy in this new period. In 1950 he had begun serious work, involving use of the Manchester computer, on a new theory of non-linear partial differential equations, proposed as "the chemical basis of morphogenesis". This and other research continued vigorously despite the interruption. (So did his personal life, which as usual he refused to adjust to the expectations of society.)

The question arises as to whether there were further developments in Turing's ideas about machine intelligence after 1950. There is an indication that there were. In the following year he gave a popular talk on BBC radio (Turing, 1951). It was basically a version of what he had set out in the 1950 paper. He explained the principle that any mechanical process could be simulated by a program run on a single machine, the computer: in particular, he had in mind the function of the brain. But this time he inserted an important *caveat* that had not been made in 1950. The machine to be simulated

> should be of the sort whose behaviour is in principle predictable by calculation. We certainly do not know how any such calculation should be done, and it was even argued by Sir Arthur Eddington that on account of the indeterminacy principle in quantum mechanics no such prediction is even theoretically possible.

Copeland (1999) has rightly signalled the importance of this new point, but his critical context suggests a link with the "oracle", a particular kind of uncomputable function used in Turing's 1938 work (Turing, 1939). But Turing made no reference to this "oracle" when admitting this possibly fatal flaw in his argument about the brain as a discrete state machine. The question he was raising was whether the space-time properties of quantum-mechanical physics could be captured by a discrete state machine model. And this was a question which went back to his earliest serious thought, being related to the work by Eddington and von Neumann that he was reading in 1928–1932, especially that of Eddington.

In 1932 Turing had speculated, influenced both by Eddington, and by trauma in his personal life, that quantum mechanics underpinned free will (Hodges, 1983). The relationship between von Neumann and Turing has enjoyed much attention

because of the question of who first had the idea of a practical universal machine (Hodges, 1983; Davis, 2000). Less well known is that Turing's first serious research study was of von Neumann's work on the foundations of quantum mechanics. Von Neumann clarified the measurement or *reduction* process in quantum mechanics; it is this which is not predictable. Seventy years later, there is no agreed or compelling explanation of how or when such "reduction" occurs. In 1953–1954, Turing wrote to his friend and colleague Robin Gandy that he was trying to invent a new quantum mechanics, and raised questions about the reduction principle, as discussed in Gandy (1954). Probably he was trying to find a predictable theory of the reduction process, to close the loophole in his argument for the computability of brain processes. However, he died in 1954 before announcing any result.

His last published paper (Turing, 1954), was again semi-popular, appearing in *Penguin Science News*. This did not mention quantum mechanics, but it returned to the pure mathematics of computability (which had recently gained new life with advances in algebra), and gave an account of Gödel's theorem. His conclusion was surprisingly unlike that pronounced in 1950; he said that Gödel's theorem showed that "common sense" was needed in interpreting axiomatic systems, and this time the intuitive "common sense" was not asserted to be a something a machine could show as well as a human being. The year 1950 seems to have marked the peak of his confidence about the prospects for machine intelligence, but it is impossible to know how his views would have developed had he lived longer.

In recent years, Roger Penrose has taken up the two themes that Turing found most difficult to fit into his thesis of computable mental functions – Gödel's theorem and the quantum-mechanical reduction process – and has said that they must be connected (Penrose, 1989, 1994, 1996). Penrose holds that the function of the brain *cannot* be simulated by a computer program, because of its quantum-mechanical physical basis. Thus, for entirely materialist reasons, no truly intelligent behaviour will ever be simulated by a computer; the full Turing Test will never be passed. Many commentators have attacked this conclusion, but it must be said that the topics Penrose emphasises are those that Turing himself found central to his arguments.

We are now so used to the properties of digital machines that it is hard to imagine the world in which Turing conjured them from pure imagination in 1936. However, it is crucial to see that what Turing offered in 1950, based on this earlier work, was something that went beyond the traditional mind-matter debate, and beyond loose science-fiction talk about humans and robots. It had a new solid substance in the digital or discrete-state machine model, made clear as never before. This structure, however, had a non-obvious limitation expressed by Gödel's theorem and Turing's own discoveries in the theory of computability. Turing always had these questions of limits in mind. Turing's universal machine now seems to sweep all before it, and continues to captivate us with the apparently never-ending range of applications that it can encompass. Turing's own excitement for this project, his game-playing enthusiasm and iconoclastic humour, live on in every conversation-program writer of the present day. But it should be remembered that Turing's imitation game actually first arose as the "fair play" argument for escaping the force of Gödel's theorem and the serious puzzle posed by the limits of what can be computed.

References

Copeland, B. J., 1999, A lecture and two radio broadcasts on machine intelligence by Alan Turing, in: *Machine Intelligence* **15**, K. Furukawa, D. Michie, and S. Muggleton, eds., Oxford University Press, Oxford, pp. 445–475.

Davis, M., 2000, *The Universal Computer*, Norton, New York.

Gandy, R. O., 1954, Letter to M. H. A. Newman, in: *The Collected Works of A M. Turing: Mathematical Logic*, R. O. Gandy and C. E. M. Yates, eds., North-Holland, Amsterdam (2001).

Hodges, A., 1983, *Alan Turing: The Enigma*, Burnett, London, Simon & Schuster, New York, new editions: Vintage, London (1992), Walker, New York (2000).

Hodges, A., 1997, *Turing, a Natural Philosopher*, Phoenix, London, Routledge, New York (1999); included in: *The Great Philosophers*, R. Monk and F. Raphael, eds., Weidenfeld & Nicolson, London (2000).

Hodges, A., 2002, Alan M. Turing, in: *Stanford Encyclopedia of Philosophy*, E. Zalta, ed.; http://plato.stanford.edu/entries/turing.

McCulloch, W. S. and Pitts, W., 1943, A logical calculus of the ideas immanent in nervous activity, *The Bulletin of Mathematical Biophysics* **5**: 115–133.

Penrose, R., 1989, *The Emperor's New Mind*, Oxford University Press, Oxford.

Penrose, R., 1994, *Shadows of the Mind*, Oxford University Press, Oxford.

Penrose, R., 1996, Beyond the doubting of a shadow, Psyche electronic journal; http://psyche.csse.monash.edu.au/v2/psyche-2–23-penrose.html.

Turing, A. M., 1936, On computable numbers, with an application to the Entscheidungsproblem, *Proceedings of the London Mathematical Society*, Series 2, **42**: 230–265.

Turing, A. M., 1939, Systems of logic defined by ordinals, *Proceedings of the London Mathematical Society*, Series 2, **45**: 161–228.

Turing, A. M., 1946, Proposed electronic calculator, report for the National Physical Laboratory, published in A. M. Turing's ACE report of 1946 and other papers, B. E. Carpenter and R. W. Doran, eds., MIT Press, Cambridge, MA (1986).

Turing, A. M., 1947, Lecture to the London Mathematical Society on 20 February 1947, published in A. M. Turing's ACE report of 1946 and other papers, B. E. Carpenter and R. W. Doran, eds., MIT Press, Cambridge, MA (1986).

Turing, A. M., 1948, Intelligent machinery, report for the National Physical Laboratory, published in *Machine Intelligence* **7**, B. Meltzer and D. Michie, eds. (1969).

Turing, A. M., 1950, Computing machinery and intelligence, *Mind* **50**: 433–460.

Turing, A. M., 1951, BBC radio talk, published in *Machine Intelligence* **15**, K. Furukawa, D. Michie, and S. Muggleton, eds., Oxford University Press, Oxford (1999).

Turing, A. M., 1954, Solvable and unsolvable problems, *Science News* **31**: 7–23.

Chapter 3
Computing Machinery and Intelligence

Alan M. Turing

Editors' Note: The following is the article that started it all – the article by Alan Turing which appeared in 1950 in the British journal, *Mind*. Accompanying the article are three running commentaries by Kenneth Ford, Clark Glymour, and Pat Hayes of the University of West Florida; Stevan Harnad of the University of Southampton; and Ayse Pinar Saygin of the University of California, San Diego, designated respectively by the symbols:♠,♣, and♥. A fourth commentary by John Lucas of Merton College, Oxford, is found in Chapter 4.

3.1 The Imitation Game

I propose to consider the question, "Can machines think?"♣ This should begin with definitions of the meaning of the terms "machine" and "think". The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. 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 the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by

Manchester University

♣Harnad: Turing starts on an equivocation. We know now that what he will go on to consider is not whether or not machines can think, but whether or not machines can do what thinkers like us can do – and if so, how. Doing is performance capacity, empirically observable. Thinking is an internal state. It correlates empirically observable as neural activity (if we only knew which neural activity corresponds to thinking!) and its associated quality introspectively observable as our own mental state when we are thinking. Turing's proposal will turn out to have nothing to do with either observing neural states or introspecting mental states, but only with generating performance capacity indistinguishable from that of thinkers like us.

R. Epstein et al. (eds.), *Parsing the Turing Test.*

another, which is closely related to it and is expressed in relatively unambiguous words.♠♣♥

The new form of the problem can be described in terms of a game which we call the "imitation game".♠♣ It is played with three people, a man (A), a woman

♠ FORD, GLYMOUR, AND HAYES: Turing derides deciding the question by an empirical survey of which sorts of objects the word "think" or its synonyms are positively applied to. Presumably, in 1950 people rarely if ever said of machines that they think, and few people in 1950 would have said that any machine, then or in the future, could *possibly* be said to think. The procedure is absurd because what people say, even what almost everyone agrees in saying, is often wildly wrong: a century before almost everyone would have agreed to the proposition that angels think.

♣ HARNAD: "Machine" will never be adequately defined in Turing's paper, although what will eventually be known as the "Turing Machine," the abstract description of a computer, will be. This will introduce a systematic ambiguity between a real physical system, doing something in the world, and another physical system, a computer, simulating the first system formally, but not actually doing what it does: an example would be the difference between a real airplane – a machine, flying in the real world – and a computer simulation of an airplane, not really flying, but doing something formally equivalent to it, in a (likewise simulated) "virtual world."

A reasonable definition of machine, rather than Turing Machine, might be any dynamical, causal system. That makes the universe a machine, a molecule a machine, and also waterfalls, toasters, oysters, and human beings. Whether or not a machine is man-made is obviously irrelevant. The only relevant property is that it is "mechanical" – i.e., behaves in accordance with the cause-effect laws of physics.

"Think" will never be defined by Turing at all; it will be replaced by an operational definition to the effect that "thinking is as thinking does." This is fine, for thinking cannot be defined in advance of knowing how thinking systems do it, and we do not yet know how. But we do know that we thinkers do it, whatever it is, when we think and we know when we are doing it (by introspection). So thinking, a form of consciousness, is already ostensively defined by just pointing to that experience we all have and know.

Taking a statistical survey like a Gallup poll instead, to find out people's opinions of what thinking is would indeed be a waste of time, as Turing points out – but then later in the paper he needlessly introduces the equivalent of a statistical survey as his criterion for having passed his Turing Test!

♥ SAYGIN: Operational definition: a definition of a theoretical construct that is stated in terms of concrete, observable procedures (Pelham, 1999). While some readers believe the imitation game is only a thought experiment, I think it is pretty clear that Turing is proposing an operational definition for machine thought. One could argue whether this is the best way to test for machine intelligence, but that would be a discussion of construct validity, i.e., the quality of someone's operational definitions, not the existence or lack thereof.

♠ FORD, GLYMOUR, AND HAYES: Turing's use of the singular here may be misleading, as we will see. There are many versions of "the" imitation game, and Turing himself seems to slide between them without giving the reader adequate notice. It might be best to take this paragraph as a description of a family of "games" that share a common theme: a real exemplar and an imitator, respectively trying to help and to fool a judge.

♣ HARNAD: Another unfortunate terminological choice: "Game" implies caprice or trickery, whereas Turing in fact means serious empirical business. The game is science, the future science of cognition – actually a branch of reverse bioengineering. "Imitation" has connotations of fakery or deception too, whereas what Turing will be proposing is a rigorous empirical methodology for testing theories of human cognitive performance capacity (and thereby also theories of the thinking that presumably engenders it). Calling this an "imitation game" (instead of a methodology for reverse-engineering human cognitive performance capacity) has invited generations of needless misunderstandings (Harnad, 1992).

(B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman.♣ He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A". The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be

"My hair is shingled, and the longest strands are about nine inches long."

In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten.♣ The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as "I am the woman, don't listen to him!" to her answers, but it will avail nothing as the man can make similar remarks.

We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is

---

♣ HARNAD: The man/woman test is an intuitive "preparation" for the gist of what will eventually be the Turing Test, namely, an empirical test of performance capacity. For this, it is first necessary that all non-performance data be excluded (hence the candidates are out of sight). This sets the stage for what will be Turing's real object of comparison, which is a thinking human being versus a (nonthinking) machine, a comparison that is to be unbiased by appearance.

Turing's criteria, as we know, will turn out to be two (though they are often confused or conflated): (1) Do the two candidates have identical performance capacity? (2) Is there any way we can distinguish them, based only on their performance capacity, so as to be able to detect that one is a thinking human being and the other is just a machine? The first is an empirical criterion: Can they both do the same things? The second is an intuitive criterion, drawing on what decades later came to be called our human "mind-reading" capacities (Frith and Frith, 1999): Is there anything about the way the candidates go about doing what they can both do that cues me to the fact that one of them is just a machine?

Turing introduces all of this in the form of a party game, rather like 20-Questions. He never explicitly debriefs the reader to the effect that what is really at issue is no less than the game of life itself, and that the "interrogator" is actually the scientist for question (1), and, for question (2), any of the rest of us, in every one of our daily interactions with one another. The unfortunate party-game metaphor again gave rise to needless cycles of misunderstandings in later writing and thinking about the Turing Test.

♣ HARNAD: This restriction of the test exclusively to what we would today call email interactions is, as noted, a reasonable way of preparing us for its eventual focus on performance capacity alone, rather than appearance, but it does have the unintended further effect of ruling out all direct testing of performance capacities other than verbal ones; and that is potentially a much more serious equivocation, to which we will return. For now, we should bear in mind only that if the criterion is to be Turing-indistinguishable performance-capacity, we can all do a lot more than just email!

played like this as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think?"♠♣♥

♠ FORD, GLYMOUR, AND HAYES: This question can be understood in several ways, depending on what one takes Turing to mean by "this game." It is usually understood to mean a version of the game where, in the terminology of 1951 (published in 1950, exposition in 1951), a "man" – and a machine each try to persuade the judge that they are the human being. However, taken literally, and we see no reason not to, "the game" refers to the game just described, and then Turing seems to be proposing a comparative test of the ability of a man to pretend to be a woman, as against the ability of a computer to pretend to be a woman (the contestant in each case being a real woman, of course). This reading – the "gender test," in contradistinction to the "species test" usually assumed – may seem strange, but it has a number of subtle advantages, including the fact that the judge is not given a predisposition to be particularly alert for signs of non-human behaviour, and the fact that the players, man and machine, both have imitation tasks.

The critical question is whether, in typed conversation, a computer can pass as a woman as convincingly – and as *unconvincingly* – as can a man. In the gender test version, as posed, the test is not only of conversational competence, but also of a special kind of knowledge: the computer must "know" what it is like for a man to try to converse like a woman (Hayes and Ford, 1995).

Psychological speculations aside, one might reasonably object that different men and women and judges would yield widely varying accuracies of judgement, or that a sufficiently skilled judge, given sufficient time, would be able to distinguish most men from most women, so that to qualify as thoughtful, the computer would have a very low bar.

Many writers assume the game should be played with the question of gender (being female) replaced by the question of species (being human), so that the judge is faced with the task of differentiating a human participant from a machine pretending to be human. Notice that under this most common interpretation, the Turing Test slides from a test for intelligence, to a test of the ability of humans to distinguish members of their own species from mechanical impostors. This version is often called the "species version," and is the most popular understanding of the Turing Test, but it was not Turing's. In the gender test, the judge is still thinking about the differences between women and men, not humans and machines. The hypothesis that one of his subjects is not human is not even in his natural space of initial possibilities. This judge has exactly the same problem to solve as a judge in the original imitation game and could be expected to bring the same attitudes and skills to the problem. For a discussion of the two versions and the advantages of Turing's own version see (Genova, 1994; Sterrett, 2000).

♣ HARNAD: Here, with a little imagination, we can already scale up to the full Turing Test, but again we are faced with a needless and potentially misleading distraction: Surely the goal is not merely to design a machine that people mistake for a human being statistically more often than not! That would reduce the Turing Test to the Gallup poll that Turing rightly rejected in raising the question of what "thinking" is in the first place! No, if Turing's indistinguishability criterion is to have any empirical substance, the performance of the machine must be equal to that of a human being – to anyone and everyone, for a lifetime.

♥ SAYGIN: Note that here the machine takes the part of A, the man. The man was trying to convince the interrogator that he actually was the woman. Now that the machine takes the place of the man in the game, will it be trying to convince the interrogator that it is a woman? The answer could be yes or no, depending on interpretation (Piccinini, 2000; Saygin et al., 2000; Traiger, 2000). As it is now generally understood, the Turing Test tries to assess a machine's ability to imitate a human being, rather than its ability to simulate a woman in an imitation game. Most subsequent remarks on Turing's work in the following 50 years, as reviewed in Saygin et al. (2000), ignore the gender issue, and if they discuss the imitation game at all, consider a game that is played between a machine (A), a human (B), and an interrogator (C) whose aim is to determine which one of the two entities he/she

3.2 Critique of the New Problem

As well as asking, "What is the answer to this new form of the question", one may ask, "Is this new question a worthy one to investigate?" This latter question we investigate without further ado, thereby cutting short an infinite regress.♠

The new problem has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man.♣ No engineer or chemist claims to be able to produce a material which is indistinguishable from the human skin. It is possible that at some time this might be done, but even supposing this invention available we should feel there was little point in trying to make a "thinking

---

is conversing with is the human. In many cases the imitation game is considered irrelevant and the discussion revolves around the vague idea of a digital computer "passing for" a human and the relation this possibility bears to Artificial Intelligence (AI). The imitation game is not an analogy Turing introduced but failed to properly execute, nor is it a joke or a commentary on gender roles in society (unlike that of [Genova, 1994; Lassègue, 1996]). But I will suggest below that the seemingly quirky gender-based imitation game is in fact an ideal and fair test for machine intelligence.

♠ FORD, GLYMOUR, AND HAYES: Turing's intellectual strategy is to replace questions of a traditional philosophical form, e.g., provide necessary and sufficient conditions for something to be intelligent, with related questions for which there is a hope of providing an answer empirically or mathematically.

♣ HARNAD: It would have had that advantage, if the line had only been drawn between appearance and performance or between structure and function. But if the line is instead between verbal and non-verbal performance capacities, then it is a very arbitrary line indeed and a very hard one to defend. As there is no explicit or even inferable defence of this arbitrary line in any of Turing's paper (nor of an equally arbitrary line between those of our "physical capacities" that do and do not depend on our "intellectual capacities"), I will take it that Turing simply did not think this through. Had he done so, the line would have been drawn between the candidate's physical appearance and structure on the one hand, and its performance capacities, both verbal and non-verbal, on the other. Just as (in the game) the difference, if any, between the man and the woman must be detected from what they do, and not what they look like, so the difference, if any, between human and machine must be detected from what they do, and not what they look like. This would leave the door open for the robotic version of the Turing Test that we will discuss later, and not just for the email version.

But before a reader calls my own dividing line between structure and function just as arbitrary, let me quickly agree that Turing has in fact introduced a hierarchy of Turing Tests here, but not an infinity of them. The relevant levels of this hierarchy will turn out to be only the following 5:

T1: The local indistinguishable capacity to perform some arbitrary task, such as chess. T1 is not really a Turing Test at all, because it is so obviously subtotal; hence the machine candidate is easily distinguished from a human being by seeing whether it can do anything else, other than play chess. If it cannot, it fails the test.

T2: The indistinguishable performance capacity in email (verbal) exchanges. This seems like a self-contained performance module, for one can talk about anything and everything, and language has the same kind of universality that computers (Turing Machines) turned out to have. T2 even subsumes chess-playing. But does it subsume star-gazing, or even food-foraging? Can the machine go and see and then tell me whether the moon is visible tonight and can it go and unearth truffles and then let me know how it went about it? These are things that a machine with email capacity alone cannot do, yet every human being can.

machine" more human by dressing it up in such artificial flesh.♣ The form in which we have set the problem reflects this fact in the condition which prevents the interrogator from seeing or touching the other competitors, or hearing their voices.♣ Some other advantages of the proposed criterion may be shown up by specimen questions and answers. Thus:

Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 s and then give as answer) 105621.
Q: Do you play chess?
A: Yes.
Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
A: (After a pause of 15 s) R-R8 mate.

The question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include.♣ We do not wish to penalize the machine for its inability to shine in beauty competitions,♣ nor to penalize a man for losing in a race against an aeroplane.♣ The conditions of our

---

T3: The indistinguishable performance capacity in robots (sensorimotor). This subsumes T2, and is (I will argue) the level of test that Turing really intended (or should have!).

T4: The indistinguishable external performance capacity, as well as internal structure and function. This subsumes T3 and adds all data that a neuroscientist might study. This is no longer strictly a Turing Test, because it goes beyond performance data, but it correctly embeds the Turing Hierarchy in a larger empirical hierarchy. Moreover, the boundary between T3 and T4 really is fuzzy: Is T3 or T4 blushing?

T5: The indistinguishable physical structure and function. This subsumes T4 and rules out any functionally equivalent but synthetic nervous systems: The T5 candidate must be indistinguishable from other human beings right down to the last molecule.

♣ HARNAD: Here Turing correctly rejects T5 and T4 – but certainly not T3.

♣ HARNAD: Yes, but using T2 as the example has inadvertently given the impression that T3 is excluded too.

♣ HARNAD: This correctly reflects the universal power of natural language (to say and describe anything in words). But "almost" does not fit the Turing criterion of identical performance capacity.

♣ HARNAD: This is the valid exclusion of appearance (moreover, most of us could not shine in beauty competitions either).

♣ HARNAD: Most of us could not beat Deep Blue at chess, nor even attain ordinary grandmaster level. It is only generic human capacities that are at issue, not those of any specific individual. On the other hand, just about all of us can walk and run. And even if we are handicapped (an anomalous case, and hardly the one on which to build one's attempts to generate positive performance capacity), we all have some sensorimotor capacity. (Neither Helen Keller nor Stephen Hawking are disembodied email-only modules.)

game make these disabilities irrelevant.♣ The "witnesses" can brag, if they consider it advisable, as much as they please about their charms, strength, or heroism, but the interrogator cannot demand practical demonstrations.♣

The game may perhaps be criticized on the ground that the odds are weighted too heavily against the machine. If the man were to try and pretend to be the machine he would clearly make a very poor showing. He would be given away at once by slowness and inaccuracy in arithmetic. May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection.♠♥

It might be urged that when playing the "imitation game" the best strategy for the machine may possibly be something other than imitation of the behaviour of a man. This may be, but I think it is unlikely that there is any great effect of this kind.

♣ HARNAD: Disabilities and appearance are indeed irrelevant. But non-verbal performance capacities are certainly not. Indeed, our verbal abilities may well be grounded in our non-verbal abilities (Cangelosi, 2001; Harnad, 1990; Steels and Kaplan, 1999). Actually, by "disability," Turing means non-ability, i.e., absence of an ability; he does not really mean being disabled in the sense of being physically handicapped, although he does mention Helen Keller later.

♣ HARNAD: This would definitely be a fatal flaw in the Turing Test, if Turing had meant it to exclude T3 – but I doubt that is what he meant. He was just arguing that it is performance capacity that is decisive (for the empirical problem that future cognitive science would eventually address), not something else that might depend on irrelevant features of structure or appearance. He merely used verbal performance as his intuition-priming example, without meaning to imply that all "thinking" is verbal and only verbal performance capacity is relevant.

♠ FORD, GLYMOUR, AND HAYES: It is clear that Turing intended passing the Turing Test to be an uncontroversial criterion *sufficient* for thought, not a necessary one. He allows machines to be inhumanly capable, for example. Indeed, the few electronic computers which existed at the time he was writing were already inhumanly capable in doing arithmetic, which is of course why large funds were expended in designing and constructing them.

The Turing Test is, however, a poorly designed experiment, depending entirely on the competence of the judge. As Turing noted above, a human would be instantly revealed by his comparative inadequacies in arithmetic unless, of course, the computer were programmed to be arithmetically incompetent. Likewise, according to media reports, some judges at the first Loebner competition (1991), a kind of Turing test contest held at the Computer Museum in Boston, rated a human as a machine on the grounds that she produced extended, well-written paragraphs of informative text at dictation speed without typing errors. (Apparently, this is now considered an inhuman ability in parts of our culture.)

♥ SAYGIN: Turing did not live long enough to reply to most critiques of his work, but in this paper he foresees many criticisms he thinks may be made by others and formulates some advance arguments against them (§6). Nevertheless, even those issues Turing diligently addresses have been raised in subsequent discussions. For example, he has subsequently been criticized both for his test being too anthropomorphic and limited (Millar, 1973), and on the basis that playing the imitation game is just one thing an intelligent machine can do and is not general enough for purposes of intelligence granting (Gunderson, 1964).

In any case there is no intention to investigate here the theory of the game, and it will be assumed that the best strategy is to try to provide answers that would naturally be given by a man.

3.3 The Machines Concerned in the Game

The question which we put in §1 will not be quite definite until we have specified what we mean by the word "machine". It is natural that we should wish to permit every kind of engineering technique to be used in our machines.♣ We also wish to allow the possibility than an engineer or team of engineers may construct a machine which works, but whose manner of operation cannot be satisfactorily described by its constructors because they have applied a method which is largely experimental.♣♥ Finally, we wish to exclude from the machines men born in the usual manner.♣ It is difficult to frame the definitions so as to satisfy these three conditions. One might for instance, insist that the team of engineers should be all of one sex, but this would not really be satisfactory, for it is probably possible to rear a complete individual from a single cell of the skin (say) of a man.♠ To do so would be a feat of biological technique deserving of the very highest praise, but we would not be inclined to regard it as a case of "constructing a thinking machine".♣ This prompts

---

♣ HARNAD: This passage (soon to be withdrawn!) implies that Turing did not mean only computers: any dynamical system we build is eligible (as long as it delivers the performance capacity). But we do have to build it, or at least have a full causal understanding of how it works. A cloned human being cannot be entered as the machine candidate (because we did not build it and do not know how it works), even though we are all "machines" in the sense of being causal systems (2000, 2003).

♣ HARNAD: Here is the beginning of the difference between the field of AI, whose goal is merely to generate a useful performance tool, and cognitive modelling (CM), whose goal is to explain how human cognition is generated. A device we built without knowing how it works would suffice for AI but not for CM.

♥ SAYGIN: Turing would clearly allow many kinds of machines to pass the test, and more importantly, through various means. Several researchers opposed this idea, especially the latter point, holding that restrictions should be placed on internal information processing if a machine is to be granted thought (Block, 1981; Gunderson, 1964). Is Turing happy to grant intelligence to any old hack that may be programmed to play the imitation game? Or is he so confident that the problem is too hard that he is willing to take the risk of having a quick and dirty solution?

♣ HARNAD: This does not, of course, imply that we are not machines, but only that the Turing Test is about finding out what kind of machine we are, by designing a machine that can generate our performance capacity, but by a functional means that we understand because we designed them.

♠ FORD GLYMOUR, AND HAYES: Turing's anticipation of cloning was not out of the blue. In the period in which this paper was written, he had a strong interest in mathematical biology; especially in morphogenesis. He published one paper on the topic and wrote a number of others. They are available in his Collected Papers.

♣ HARNAD: This is because we want to explain thinking capacity, not merely duplicate it.

us to abandon the requirement that every kind of technique should be permitted. We are the more ready to do so in view of the fact that the present interest in "thinking machines" has been aroused by a particular kind of machine, usually called an "electronic computer" or "digital computer". Following this suggestion we only permit digital computers to take part in our game.♣

♣ HARNAD: This is where Turing withdraws the eligibility of all engineering systems but one, introducing another arbitrary restriction – one that would again rule out T3. Turing said earlier (correctly) that any engineering device ought to be eligible. Now he says it can only be a computer. His motivation is partly, of course, the fact that the computer (Turing Machine) has turned out to be universal, in that it can simulate any other kind of machine. But here we are squarely in the T2/T3 equivocation, for a simulated robot in a virtual world is neither a real robot, nor can it be given a real robotic Turing Test, in the real world. Both T2 and T3 are tests conducted in the real world. But an email interaction with a virtual robot in a virtual world would be T2, not T3.

To put it another way, with the Turing Test we have accepted, with Turing, that thinking is as thinking does. But we know that thinkers can and do more than just talk. And it remains what thinkers can do that our candidate must likewise be able to do, not just what they can do verbally. Hence, just as flying is something that only a real plane can do, and not a computer-simulated virtual plane, be it ever so Turing-equivalent to the real plane – so passing T3 is something only a real robot can do, not a simulated robot tested by T2, be it ever so Turing-equivalent to the real robot. (I also assume it is clear that Turing Testing is testing in the real world: a virtual-reality simulation [VR] would be no kind of a Turing Test; it would merely be fooling our senses in the VR chamber, rather than testing the candidate's real performance capacity in the real world.)

So the restriction to computer simulation, though perhaps useful for planning, designing and even pretesting the T3 robot, is merely a practical methodological strategy. In principle, any engineered device should be eligible, and it must be able to deliver T3 performance, not just T2.

It is of interest that contemporary cognitive robotics has not gotten as much mileage out of computer-simulation and virtual-worlds as might have been expected, despite the universality of computation. "Embodiment" and "situatedness" (in the real world) have turned out to be important ingredients in empirical robotics (Brooks, 2002; Steels and Kaplan, 1999), with the watchword being that the real world is better used as its own model (rather than virtual robots having to simulate, hence second-guess in advance, not only the robot, but the world too).

The impossibility of second-guessing the robot's every potential "move" in advance, in response to every possible real-world contingency, also points to a latent (and I think fatal) flaw in T2 itself: Would it not be a dead giveaway if one's email T2 pen pal proved incapable of commenting on the analogue family photos we kept inserting with our text? (If he can process the images, he is not just a computer, but at least a computer plus A/D peripheral devices, already violating Turing's arbitrary restriction to computers alone.) Or if one's pen pal was totally ignorant of contemporaneous real-world events, apart from those we describe in our letters? Would not even its verbal performance break down if we questioned it too closely about the qualitative and practical details of sensorimotor experience? Could all of that really be second-guessed purely verbally in advance?

This restriction appears at first sight to be a very drastic one. I shall attempt to show that it is not so in reality. To do this necessitates a short account of the nature and properties of these computers.♠♣

It may also be said that this identification of machines with digital computers, like our criterion for "thinking", will only be unsatisfactory if (contrary to my belief), it turns out that digital computers are unable to give a good showing in the game.♣

There are already a number of digital computers in working order, and it may be asked, "Why not try the experiment straight away? It would be easy to satisfy the conditions of the game. A number of interrogators could be used, and statistics compiled to show how often the right identification was given." The short answer is that we are neither asking whether all digital computers would do well in the game nor whether the computers at present available would do well, but whether there are imaginable computers which would do well.♠ But this is only the short answer. We shall see this question in a different light later.

♠ FORD, GLYMOUR, AND HAYES: The "universal machine" idea that a computer can be designed that can in principle simulate *all* other computers, is now widely understood, but it was not at all obvious when Turing was writing, and indeed the idea was widely derided or rejected as ludicrous. The multitude of purposes that computers could serve was little appreciated. A senior British government scientific advisor asserted that the entire country would only need four or five computers, on the grounds that they could be used only for generating elevation tables for naval gunnery. Even von Neumann thought the most important application of computers in mathematics would be to compute examples that would then give mathematicians intuitions about proofs. It seems safe to say that nobody, probably not even Turing, could have foreseen the many uses to which computers have been put in modern society.

The next few pages are a tour de force of exposition for the time Turing was writing, but will seem obvious to many people in this and future generations.

♣ HARNAD: The account of computers that follows is useful and of course correct, but it does not do anything at all to justify restricting the Turing Test to candidates that are computers. Hence this arbitrary restriction is best ignored.

♣ HARNAD: This is the "game" equivocation again. It is not doubted that computers will give a good showing, in the Gallup poll sense. But empirical science is not just about a good showing: An experiment must not just fool most of the experimentalists most of the time! If the performance-capacity of the machine must be indistinguishable from that of the human being, it must be totally indistinguishable, not just indistinguishable more often than not. Moreover, some of the problems that I have raised for T2 – the kinds of verbal exchanges that draw heavily on sensorimotor experience – are not even likely to give a good showing if the candidate is only a digital computer, regardless of how rich a database it is given in advance.

♠ FORD, GLYMOUR, AND HAYES: Again, a simple point that has often been misunderstood since.

3.4 Digital Computers♥

The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer.♠ The human computer is supposed to be following fixed rules♣; he has no authority to deviate from them in any detail. We may suppose that these rules are supplied in a book, which is altered whenever he is put on to a new job. He has also an unlimited supply of paper on which he does his calculations. He may also do his multiplications and additions on a "desk machine", but this is not important.

If we use the above explanation as a definition we shall be in danger of circularity of argument. We avoid this by giving an outline of the means by which the desired effect is achieved. A digital computer can usually be regarded as consisting of three parts:

(i) Store
(ii) Executive unit
(iii) Control

The store is a store of information, and corresponds to the human computer's paper, whether this is the paper on which he does his calculations or that on which his book of rules is printed. In so far as the human computer does calculations in his head a part of the store will correspond to his memory.

The executive unit is the part which carries out the various individual operations involved in a calculation. What these individual operations are will vary from machine to machine. Usually fairly lengthy operations can be done such as

---

♥ SAYGIN: Turing's treatment here and in the next section is one of the most concise, but clear explanations of basic theory of computing that exists – I think it could be useful for teaching purposes. Although Turing is one of the fathers of computer science, being a pioneer in a field does not in itself mean that one is able to speak coherently about that field at an introductory level. I think his success is partly due to the fact that he himself is able to see, in a way characteristic of an interdisciplinary scientist, the relations between the abstract notions of computation, different levels of application, behavioural manifestation, and philosophical analysis.

♠ FORD, GLYMOUR, AND HAYES: It is often said that computers were invented around 1940, but this claim would have sounded strange at that date. The bare term "computer" then meant a human being who (often aided by an electromechanical calculator) performed computations for a living, or in support of some other intellectual activity such as theoretical physics, astronomy, or code-breaking. Computational skill was highly prized, and to be called a "computer" in 1940 was a professional compliment, as it had been since at least the 1850s.

In fact, the famous astronomer Simon Newcomb wrote a recommendation letter for one of his calculators in which he said, "His mind is more like a mathematical machine than any I have ever encountered," which was high praise indeed. To explain the operation of an electronic computer (the adjective, now seeming redundant, is commonly dropped) in terms of rule-books used by human beings, was therefore a perfectly natural expository device. However, this device can be misleading when read with hindsight, since it can suggest that the "thinking" part of the computer is the part of it which corresponds in this expositional metaphor to the human computer, i.e., the "executive unit" or CPU, which is nowadays simply a piece of etched silicon. A similar misunderstanding is the layman's objection – which Turing mentions later – that computers "can only obey instructions."

♣ HARNAD: This goes on to describe what has since become the standard definition of computers as rule-based symbol-manipulating devices (Turing machines).

"Multiply 3540675445 by 7076345687" but in some machines only very simple ones such as "Write down 0" are possible.

We have mentioned that the "book of rules" supplied to the computer is replaced in the machine by a part of the store. It is then called the "table of instructions". It is the duty of the control to see that these instructions are obeyed correctly and in the right order. The control is so constructed that this necessarily happens.

The information in the store is usually broken up into packets of moderately small size. In one machine, for instance, a packet might consist of ten decimal digits. Numbers are assigned to the parts of the store in which the various packets of information are stored, in some systematic manner. A typical instruction might say:

"Add the number stored in position 6809 to that in 4302 and put the result back into the latter storage position."

Needless to say it would not occur in the machine expressed in English. It would more likely be coded in a form such as 6809430217. Here, 17 says which of various possible operations is to be performed on the two numbers. In this case the operation is that described above, viz. "Add the number…" It will be noticed that the instruction takes up ten digits and so forms one packet of information, very conveniently. The control will normally take the instructions to be obeyed in the order of the positions in which they are stored, but occasionally an instruction such as

"Now obey the instruction stored in position 5606, and continue from there" may be encountered, or again

"If position 4505 contains 0 obey next the instruction stored in 6707, otherwise continue straight on."

Instructions of these latter types are very important because they make it possible for a sequence of operations to be replaced over and over again until some condition is fulfilled, but in doing so to obey, not fresh instructions on each repetition, but the same ones over and over again. To take a domestic analogy. Suppose Mother wants Tommy to call at the cobbler's every morning on his way to school to see if her shoes are done, she can ask him afresh every morning. Alternatively she can stick up a notice once and for all in the hall which he will see when he leaves for school and which tells him to call for the shoes, and also to destroy the notice when he comes back if he has the shoes with him.

The reader must accept it as a fact that digital computers can be constructed, and indeed have been constructed, according to the principles we have described, and that they can in fact mimic the actions of a human computer very closely.♠♥

♠ FORD, GLYMOUR, AND HAYES: One can almost sense the frustration that Turing may have felt when trying to find a convincing way to persuade a sceptical audience that mechanical computation was indeed possible. Again, all these metaphors about Mother and Tommy seem curiously antiquated to a contemporary sensibility.

♥ SAYGIN: The analogies here are overly simplified for purposes of explanation. However, I think Turing does believe at some level that most human behaviour is guided by "programs" of the sort that one prepares to make machines perform actions. It is easy to criticize this view, claiming AI has not produced much in terms of intelligent behaviour based on programs of this sort, and that

The book of rules which we have described our human computer as using is of course a convenient fiction. Actual human computers really remember what they have got to do. If one wants to make a machine mimic the behaviour of the human computer in some complex operation one has to ask him how it is done, and then translate the answer into the form of an instruction table.♠ Constructing instruction tables is usually described as "programming". To "programme a machine to carry out the operation A" means to put the appropriate instruction table into the machine so that it will do A.♠

An interesting variant on the idea of a digital computer is a "digital computer with a random element". These have instructions involving the throwing of a die or some equivalent electronic process; one such instruction might for instance be, "Throw the die and put the resulting number into store 1000". Sometimes such a machine is described as having free will (though I would not use this phrase

---

there is much more to being human than just following rules. These latter views are not inconsistent with Turing's thought, and a careful reading of his work will reveal he is also aware that human behaviour is guided by a program much more complex than these analogies suggest, even when random elements are thrown into the picture. It is also likely to be a rather opaque, if not cryptic, program since it will be based on a lifetime of perception, sensation, action, learning and buildup on little innate substrate in a rather experience-driven manner over a long period of time. But it does not follow from the complexity and opacity of the "human behavior program" that runs on the brain that is qualitatively different from a computer program of the sort discussed here. The point here is not to defend what is sometimes called the "computational view of the mind," which, in the light of recent research in cognitive neuroscience, is too symbolic and restricted to account for the level of complexity needed to model human minds – I am pretty sure Turing would not subscribe to that view either. But creating such a program based on ideas from modern cognitive science research and theory (e.g., based on connectionism, dynamical systems theory, embodied cognition and theoretical neuroscience) could be consistent with Turing's views.

♠ Ford, Glymour, and Hayes: This sentence is prescient. Turing was probably thinking of iterative numerical computations of the kind that human computers did indeed perform, but in fact (with a generous interpretation of "instruction table") this is exactly how "knowledge-based systems" are constructed, which have proven capable of performing many tasks which were not previously considered to lie within the province of human computation, but instead to require other human abilities such as "intuition" or "judgement."

♠ Ford, Glymour, and Hayes: Again, describing programming as the construction of look-up tables now seems very archaic. We are now much more familiar with programming as centrally concerned with *language*: programs typically manipulate expressions which themselves may be further interpreted as code, and the actual physical machine may be many levels below all this programming, almost invisible to the human user and even to the programmer. What Turing is describing, and what was at the time the only method of programming available, is what we would now call "assembly-code" programming, an activity that only a few specialists ever practice. In most modern computers, virtually every instruction executed by the CPU was generated by some other piece of code rather than written by a human programmer. Writing assembly code requires an intimate knowledge of the inner workings of the computer's hardware. Turing was what would now be called a wizard or a hacker. Given his views on programming technique and hardware design, he would probably be horrified by the wastefulness of modern programming, in which billions of machine cycles are wasted waiting for human typists' fingers to hit the next key.

myself).♣ It is not normally possible to determine from observing a machine whether it has a random element, for a similar effect can be produced by such devices as making the choices depend on the digits of the decimal for π.

Most actual digital computers have only a finite store. There is no theoretical difficulty in the idea of a computer with an unlimited store. Of course, only a finite part can have been used at any one time. Likewise only a finite amount can have been constructed, but we can imagine more and more being added as required. Such computers have special theoretical interest and will be called infinitive capacity computers.

The idea of a digital computer is an old one. Charles Babbage, Lucasian Professor of Mathematics at Cambridge from 1828 to 1839, planned such a machine, called the Analytical Engine, but it was never completed. Although Babbage had all the essential ideas, his machine was not at that time such a very attractive prospect. The speed which would have been available would be definitely faster than a human computer but something like 100 times slower than the Manchester machine, itself one of the slower of the modern machines. The storage was to be purely mechanical, using wheels and cards.

The fact that Babbage's Analytical Engine was to be entirely mechanical will help us to rid ourselves of a superstition. Importance is often attached to the fact that modern digital computers are electrical, and that the nervous system also is electrical.♠ Since Babbage's machine was not electrical, and since all digital computers are in a sense equivalent, we see that this use of electricity cannot be of theoretical importance. Of course, electricity usually comes in where fast signalling is concerned, so that it is not surprising that we find it in both these connections. In the nervous system chemical phenomena are at least as important as electrical. In certain computers the storage system is mainly acoustic. The feature of using electricity is thus seen to be only a very superficial similarity. If we wish to find such similarities we should look rather for mathematical analogies of function.

3.5 Universality of Digital Computers

The digital computers considered in the last section may be classified amongst the "discrete state machines". These are the machines which move by sudden jumps or clicks from one quite definite state to another. These states are sufficiently different

---

♣ HARNAD: Nor would I. But surely an even more important feature for a Turing Test candidate than a random element or statistical functions would be autonomy in the world – which is something a T3 robot has a good deal more of than a T2 pen pal. The ontic side of free will – namely, whether we ourselves, real human beings, actually have free will – rather exceeds the scope of Turing's paper (Harnad, 1982b). So too does the question of whether a Turing test-passing machine would have any feelings at all (whether free or otherwise; Harnad, 1995). What is clear, though, is that computational rules are not the only ways to "bind" and determine performance: ordinary physical causality can do so too.

♠ FORD, GLYMOUR, AND HAYES: A similar superstition is the view that brains cannot be thought of as computers because they are made of organic material.

for the possibility of confusion between them to be ignored. Strictly speaking there are no such machines. Everything really moves continuously. But there are many kinds of machine which can profitably be *thought* of as being discrete-state machines. For instance, in considering the switches for a lighting system it is a convenient fiction that each switch must be definitely on or definitely off. There must be intermediate positions, but for most purposes we can forget about them. As an example of a discrete-state machine we might consider a wheel which clicks round through 120° once a second, but may be stopped by a lever which can be operated from outside; in addition a lamp is to light in one of the positions of the wheel. This machine could be described abstractly as follows. The internal state of the machine (which is described by the position of the wheel) may be q_1, q_2, or q_3. There is an input signal i_0 or i_1 (position of lever). The internal state at any moment is determined by the last state and input signal according to the table

		Last State	
		q_1	q_2	q_3
Input	i_0	q_2	q_3	q_1
	i_1	q_2	q_3	q_1

The output signals, the only externally visible indication of the internal state (the light) are described by the table

State	q_1	q_2	q_3
Output	o_0	o_0	o_1

This example is typical of discrete-state machines. They can be described by such tables provided they have only a finite number of possible states.

It will seem that given the initial state of the machine and the input signals it is always possible to predict all future states.[♣] This is reminiscent of Laplace's view that from the complete state of the universe at one moment of time, as described by the positions and velocities of all particles, it should be possible to predict all future states. The prediction which we are considering is, however, rather nearer to practicability than that considered by Laplace. The system of the "universe as a whole" is such that quite small errors in the initial conditions can have an overwhelming effect at a later time.[♠] The displacement of a single electron by a

[♣] HARNAD: The points about determinism are probably red herrings. The only relevant property is performance capacity. Whether either the human or the machine is completely predictable is irrelevant. (Both many-body physics and complexity theory suggest that neither causal determinacy nor following rules guarantee predictability in practise – and this is without even invoking the arcana of quantum theory.)

[♠] FORD, GLYMOUR, AND HAYES: In more modern terminology, the universe is in some sense *chaotic*. Chaos theory had not been developed when Turing was writing, of course.

billionth of a centimetre at one moment might make the difference between a man being killed by an avalanche a year later, or escaping. It is an essential property of the mechanical systems which we have called "discrete state machines" that this phenomenon does not occur. Even when we consider the actual physical machines instead of the idealized machines, reasonably accurate knowledge of the state at one moment yields reasonably accurate knowledge any number of steps later.

As we have mentioned, digital computers fall within the class of discrete-state machines. But the number of states of which such a machine is capable is usually enormously large. For instance, the number for the machine now working at Manchester is about $2^{165,000}$, i.e., about $10^{50,000}$. Compare this with our example of the clicking wheel described above, which had three states. It is not difficult to see why the number of states should be so immense. The computer includes a store corresponding to the paper used by a human computer. It must be possible to write into the store any one of the combinations of symbols which might have been written on the paper. For simplicity suppose that only digits from 0 to 9 are used as symbols. Variations in handwriting are ignored. Suppose the computer is allowed 100 sheets of paper each containing 50 lines each with room for 30 digits. Then the number of states is $10^{100 \times 50 \times 30}$, i.e., $10^{150,000}$. This is about the number of states of three Manchester machines put together. The logarithm to the base two of the number of states is usually called the "storage capacity" of the machine. Thus the Manchester machine has a storage capacity of about 165,000 and the wheel machine of our example about 1.6. If two machines are put together their capacities must be added to obtain the capacity of the resultant machine. This leads to the possibility of statements such as "The Manchester machine contains 64 magnetic tracks each with a capacity of 2,560, eight electronic tubes with a capacity of 1,280. Miscellaneous storage amounts to about 300 making a total of 174,380."♠♥

Given the table corresponding to a discrete-state machine it is possible to predict what it will do. There is no reason why this calculation should not be carried out by means of a digital computer. Provided it could be carried out sufficiently quickly the digital computer could mimic the behaviour of any discrete-state machine. The imitation game could then be played with the machine in question (as B) and the mimicking digital computer (as A) and the interrogator would be unable to distinguish them.♠ Of course, the digital computer must have an adequate storage

---

♠ Ford, Glymour, and Hayes: It is hard to compare this accurately with modern machines, but a typical laptop computer may have an active memory capacity of approximately 10|9, and a hard disc capacity of perhaps a hundred times more. Of course, not all of this huge capacity may be being used in a way that Turing would have thought sensible.

♥ Saygin: Revisiting the question of whether Turing was proposing the game as a real operational definition or test. It seems highly unlikely to me that a man proposing a thought experiment would spend such time, space and energy to explain not only what he means by "thinking" but also exactly what kind of machine a digital computer is.

♠ Ford, Glymour, and Hayes: Here, Turing seems to be using the term "imitation game" in a very generic sense.

capacity as well as working sufficiently fast. Moreover, it must be programmed afresh for each new machine which it is desired to mimic.♥

This special property of digital computers, that they can mimic any discrete-state machine, is described by saying that they are *universal* machines. The existence of machines with this property has the important consequence that, considerations of speed apart, it is unnecessary to design various new machines to do various computing processes. They can all be done with one digital computer, suitably programmed for each case. It will be seen that as a consequence of this all digital computers are in a sense equivalent.♣

We may now consider again the point raised at the end of §3. It was suggested tentatively that the question, "Can machines think?" should be replaced by "Are there imaginable digital computers which would do well in the imitation game?" If we wish we can make this superficially more general and ask "Are there discrete-state machines which would do well?" But in view of the universality property we see that either of these questions is equivalent to this, "Let us fix our attention on one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate program, C can

♥ SAYGIN: Turing reminds us of the imitation game here. He is trying to emphasize the universality aspect of the discrete-state machine, but he does so using the original indistinguishability test we started with. There is an interesting twist: What is the interrogator looking for in this instantiation of the game? Which one is the machine? Which one is the mimicking digital computer? What would you even ask in order to differentiate between the two? It does not make much sense, unless Turing means they will play the gender-based imitation game. He never says we change the game into anything other than the gender-based game anyway. It may sound silly or pointless to compare how well two entities imitate a woman in a teletype conversation, but as I will elaborate below, tests and experiments often construct situations that do not have direct equivalents in real life (i.e., they do not always have high ecological validity).

♣ HARNAD: All true, but all irrelevant to the question of whether a digital computer alone could pass T2, let alone T3. The fact that eyes and legs can be simulated by a computer does not mean a computer can see or walk (even when it is simulating seeing and walking). So much for T3. But even just for T2, the question is whether simulations alone can give the T2 candidate the capacity to verbalize and converse about the real world indistinguishably from a T3 candidate with autonomous sensorimotor experience in the real world.

(I think yet another piece of unnoticed equivocation by Turing – and many others – arises from the fact that thinking is not directly observable, which helps us imagine that computers think. But even without having to invoke the other-minds problem (Harnad, 1991), one needs to remind oneself that a universal computer is only formally universal: it can describe just about any physical system, and simulate it in symbolic code, but in doing so, it does not capture all of its properties: Exactly as a computer-simulated airplane cannot really do what a plane does (i.e., fly in the real world), a computer-simulated robot cannot really do what a real robot does (act in the real world) – hence there is no reason to believe it is really thinking. A real robot may not really be thinking either, but that does require invoking the other-minds problem, whereas the virtual robot is already disqualified for exactly the same reason as the virtual plane: both fail to meet the Turing Test criterion itself, which is real performance capacity, not merely something formally equivalent to it!).

be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man?"♠♥

♠ FORD, GLYMOUR, AND HAYES: Turing implicitly uses what has become known as the Church-Turing thesis: The computable functions are all and only those computable by a Universal Turing Machine. The idea that one can make a given computer act like any other just by reprogramming it, given enough processor speed and memory capacity, is now a familiar idea in our culture; but persuading his audience of its reasonableness was probably one of Turing's most difficult tasks of exposition in 1951.

♥ SAYGIN: Notice that the woman has disappeared from the game altogether. But the objectives of A, B, and C remain unaltered; at least Turing does not explicitly state any change. To be precise, what we have is a digital computer and a man both trying to convince an interrogator that they are the real woman.

Why the fuss about the woman, the man, and the replacement? Turing does not seem the type of person who would beat around the bush for no reason. What is going on here?

One could say the gender-based imitation game is merely an analogy, serving the purpose of making the paper easier to understand. But in a paper that starts with the sentence, "Can machines think?" would something like "Let us take a machine and a man and see if the machine can convince interrogators that it is a human being via teletype conversations" be really much harder to process or understand? Or maybe Turing was simply careless and forgot to clarify that we are no longer talking about the gender-based imitation game. Given the level of detail and precision in Turing's writing (see Sections 4 and 5 of this paper), this is unlikely to be the explanation. Also bear in mind Turing is a successful mathematician, a discipline characterized by precision of definition and rigor in argument and generalization, which would make it unlikely that he is being sloppy.

Here is my explanation for the quirks of the game – I cannot guarantee that this is what Turing intended, but I think it is consistent with the way he thinks and writes. Neither the man in the gender-based imitation game nor any kind of machine is a woman. Furthermore what Turing proposes is essentially to compare the machine's success against that of the man – not to look at whether it actually "beats" the woman. The man and the machine are measured in terms of their respective performances and their performances are comparable because they are both simulating something which they are not. Even though it is regarded as obscure by many, the imitation game could be a carefully planned experimental design. It provides a fair basis for comparison: the woman (either as a participant in the game or as a concept) acts as a neutral point so that the two imposters can be assessed in how well they perform the imitation. In other words, Turing gives us a question that is to be examined via a carefully defined task, an experimental group (digital computers) and a control group (men). This setup looks more like an operational definition given in terms of an experimental design than anything else.

It might seem that we are a long way from such relatively minor methodological points being relevant. But at least two studies have shown that people's judgments of computers' conversational performance are substantially influenced by whether or not they know in advance that their conversational partners may be machines. In the 1970s, a group of scientists devised an electronic interviewing environment where experienced psychiatrists corresponded with both real-life paranoid patients and computer programs simulating paranoid behaviour through teletype. The judges were not told that some of the interviewees could be computer programs. Details can be found in Colby et al. (1972), but to summarize, the finding was that the psychiatric judges did not do better than chance guessing at identifying the computers from the human patients.

In a more recent study, we carried out an experiment to examine possible relationships between pragmatic violations and imitation game performance, using real excerpts from human–computer conversations (Saygin and Cicekli, 2002). Due to the design of the experiment, some subjects made pragmatic judgments on a set of conversations without being told there were computers involved

3.6 Contrary Views on the Main Question

We may now consider the ground to have been cleared and we are ready to proceed to the debate on our question, "Can machines think?" and the variant of it quoted at the end of the last section. We cannot altogether abandon the original form of the problem, for opinions will differ as to the appropriateness of the substitution and we must at least listen to what has to be said in this connexion.

It will simplify matters for the reader if I explain first my own beliefs in the matter. Consider first the more accurate form of the question. I believe that in about 50 years' time it will be possible, to program computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70% chance of making the right identification after 5 min of questioning.♠♣♥ The original

---

before they were told about the imitation game and asked to evaluate the computers' performance in the same conversations, while other subjects knew computers' involvement from the outset. We noted that even something seemingly trivial like having read the conversations only once without any bias prior to being asked to make decisions regarding the computers' behaviour had a differential effect on people's judgments. In particular, the analysis revealed that when people were faced with anomalies in the conversations, those who knew about computers' involvement tended to automatically think these were indicative of the conversational partner's identity (i.e., by the fact that it is a machine). On the other hand, unbiased subjects tried to work through the problematic exchanges in the same way they would in a pragmatically confusing conversation between humans. Now note that the gender-based imitation game is immune to the bias that knowledge of computer participation may bring. It allows the interrogators to work out pragmatic violations (and in general, exercise their naive psychology) the way they normally do; therefore, this design allows us to give the digital computers a fairer shot at performing well.

In sum, the gender-based imitation game is a good experimental design. It provides an operational definition (i.e., a larger question is replaced by a task we can evaluate). It is controlled; the task is simulating something both the experimental and control subjects are not. Furthermore, any bias the interrogator (which may be thought of as a measurement device) brings in will be based on gender expectations, which will tend not to affect the two groups differentially. Viewed in this light, the quirky imitation game may well be one of the few ways to fairly and experimentally assess machine thought.

♠ Ford, Glymour, and Hayes: Turing was right about the memory capacity of modern computers, but it is widely claimed that he was wrong in his Turing Test prediction: here we are, 50 years later, and where are the passers of his imitation game? However, notice that Turing says that it will be *possible*. That question is still moot: maybe it is possible. Certainly, computers have already performed many tasks that were previously thought of as requiring human sagacity of some kind. But in any case, very few contemporary AI researchers are seriously trying to build a machine to play Turing's imitation game. Instead they are concerned with exploring the computational machinery of intelligence itself, whether in humans, dogs, computers, or aliens. The scientific aim of AI research is to understand intelligence as computation, and its engineering aim is to build machines that surpass or extend human mental abilities in some useful way. Trying to imitate a human conversation (however "intellectual" it may be) contributes little to either ambition. Progress in AI is not measured by checking fidelity to a human conversationalist. And yet many critics of AI are complaining of a lack of progress toward this old ambition. But perhaps we should forgive the critics, as even many AI textbooks still offer the Turing Test as AI's ultimate goal, which seems akin to starting a textbook on aeronautical engineering with an explanation that the goal of the field is to make machines that fly so exactly like pigeons that they can even fool other pigeons (Ford and Hayes, 1998).

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

---

This is of course a huge area of controversy, far larger than we have space here to survey, but one point may be worth making. In making this 50-year prediction, Turing may have meant that a determined 50-year effort devoted to this single aim could succeed, a kind of imitation-game Manhattan project, or a fivefold expansion of the decade of national effort it took to get a human being to the moon and back. Certainly, given his wartime experiences of large-scale government-sponsored projects, this interpretation is not implausible; and later in the paper he suggests explicitly that it would be a 3,000-man-year project.

♣ HARNAD (p. 41): No doubt this party-game/Gallup poll criterion can be met by today's computer programs – but that remains as meaningless a demographic fact today as it was when predicted 50 years ago. Like any other science, cognitive science is not the art of fooling most of the people for some or most of the time! The candidate must really have the generic performance capacity of a real human being – capacity that is totally indistinguishable from that of a real human being to any real human being (for a lifetime, if need be!). No tricks: real performance capacity.

♥ SAYGIN (p. 41): More than 50 years have gone by since Turing wrote these often-quoted words, yet we are nowhere near "the goal." How could Turing, a man with such great vision and intellect, so grossly underestimate the time it would take to tackle the problems he left behind? I grant it that Turing underestimated either how hard the task at hand is, or how long it takes to carry out such a task. But I wonder sometimes if that is the whole story. Could he, in addition, have overestimated how hard future researchers would work at the problems? I think the latter has played more of a role than is commonly considered in the fact that we have a gaping hole between Turing's expectations and the current state of AI. Think about it: Can we really say we followed Turing's advice, gave it our all and it did not work? Or did we try shortcuts and little hacks and cheats and gave up in pursuit of "useful" AI when they did not work? The point here is not to criticize AI researchers for working on this or that topic. I only want to note that we do not know how much closer we would have been at developing AI systems that can communicate using natural language had we actually pursued it as a serious, full-time goal. Turing's tone in this paper leads me to think that the future he envisioned is based on scientists, philosophers, and programmers working hard and wholeheartedly towards the goal, patiently overcoming obstacles and making steady progress. What really happened in the AI arena was a buzz, a wave of optimism with many researchers believing that successful AI was right around the corner, finding the whole endeavor challenging but "cool," and wanting to make it work and make it work fast. However, when the problem proved too difficult to yield fruit soon, there was an ensuing burnout, which soon led to a lack of serious interest in endeavors such as the Turing Test. Some AI researchers even went as far as outwardly refusing to work on the Turing Test, defending that it belongs in history books rather than current research agendas, indeed calling it "harmful for AI" (Hayes and Ford, 1995).

♣ HARNAD: It is not meaningless, it is merely indecisive: What we mean by "think" is, on the one hand, what thinking creatures can do and how they can do it, and, on the other hand, what it feels-like to think. What thinkers can do is captured by the Turing Test. A theory of how they do it is provided by how our man-made machine does it. (If there are several different successful machines, it is a matter of normal inference-to-the-best-theory.) So far, nothing is meaningless. Now we ask: Do the successful candidates really feel, as we do when we think? This question is not meaningless; it is merely unanswerable – in any other way than by being the candidate. It is the familiar old other-minds problem (Harnad, 1991).

machines thinking without expecting to be contradicted.♠♣ I believe further that no useful purpose is served by concealing these beliefs. The popular view that scientists proceed inexorably from well-established fact to well-established fact, never being influenced by any improved conjecture, is quite mistaken. Provided it is made clear which are proved facts and which are conjectures, no harm can result. Conjectures are of great importance since they suggest useful lines of research.♠♣

I now proceed to consider opinions opposed to my own.

3.6.1 The Theological Objection

Thinking is a function of man's immortal soul.♣ God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think.

♠ FORD, GLYMOUR, AND HAYES: That prediction has in a sense been vindicated: the idea of thinking machines has indeed entered popular culture, and is widely discussed as either a present reality or an imminent one. What is more interesting, however, and what Turing apparently did not foresee, is the emergence of a kind of linguistic creepage, where the boundaries of "real thinking" are redrawn so as to exclude whatever it is that machines become able to do. When electronic computers were new, the ability to perform mental arithmetic rapidly and accurately was widely admired as a human mental ability; now it is "merely" mechanical. Now that a computer has beaten the world chess champion, skill at chess is becoming perceived as "merely" mechanical. This gradual but irresistible cultural shift in meaning also bears on the utility of the imitation game: One has to ask, to which generation does the judge belong? Behaviour that someone of Turing's generation would have found convincing may completely fail to impress someone who grew up with talking teddy bears.

♣ HARNAD: Yes, but only at a cost of demoting "thinking" to meaning only "information processing" rather than what you or I do when we think, and what that feels-like.

♠ FORD, GLYMOUR, AND HAYES: One can make out a reasonable case that this paper, and its bold conjectures, played a central role in the emergence of AI and cognitive science in the 1960s and 1970s.

♣ HARNAD: This is mistaken. Yes, science proceeds by a series of better approximations, from empirical theory to theory. But the theory here would be the actual design of a successful Turing Test candidate, not the conjecture that computation (or anything else) will eventually do the trick. Turing is confusing formal conjectures (such as that the Turing machine and its equivalents capture all future notions and instances of what we mean by "computation" – the "Church/Turing Thesis") and empirical hypotheses, such as that thinking is just computation. Surely the Turing Test is not a license for saying that we are explaining thinking better and better as our candidates fool more and more people longer and longer. On the other hand, something else that sounds superficially similar to this could be said about scaling up the Turing Test empirically by designing a candidate that can do more and more of what we can do. And Turing testing certainly provides a methodology for such cumulative theory-building and theory-testing in cognitive science.

♣ HARNAD: The real theological objection is not so much that the soul is immortal but that it is immaterial. This view also has non-theological support from the mind/body problem: no one – theologian, philosopher, or scientist – has even the faintest hint of an idea of how mental states

I am unable to accept any part of this, but will attempt to reply in theological terms. I should find the argument more convincing if animals were classed with men, for there is a greater difference, to my mind, between the typical animate and the inanimate than there is between man and the other animals.♣ The arbitrary character of the orthodox view becomes clearer if we consider how it might appear to a member of some other religious community. How do Christians regard the Moslem view that women have no souls?♠ But let us leave this point aside and return to the main argument. It appears to me that the argument quoted above implies a serious restriction of the omnipotence of the Almighty. It is admitted that there are certain things that He cannot do such as making one equal to two, but should we not believe that He has freedom to confer a soul on an elephant if He sees fit? We might expect that He would only exercise this power in conjunction with a mutation which provided the elephant with an appropriately improved brain to minister to the needs of this soul. An argument of exactly similar form may be made for the case of machines. It may seem different because it is more difficult to "swallow". But this really only means that we think it would be less likely that He would consider the circumstances suitable for conferring a soul. The circumstances in question are discussed in the rest of this paper. In attempting to construct such machines we should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will be providing mansions for the souls that He creates.

However, this is mere speculation. I am not very impressed with theological arguments whatever they may be used to support. Such arguments have often been found unsatisfactory in the past. In the time of Galileo it was argued that the texts, "And the sun stood still… and hasted not to go down about a whole day" (Joshua x. 13) and "He laid the foundations of the earth, that it should not move at any time" (Psalm cv. 5) were an adequate refutation of the Copernican theory. With our present knowledge such an argument appears futile. When that knowledge was not available it made a quite different impression.♠

---

can be material states (or, as I prefer to put it, how functional states can be felt states). This problem has been dubbed "hard" (Chalmers in Shear, 1997). It may be even worse: it may be insoluble (Harnad, 2001). But this is no objection to Turing Testing which, even if it will not explain how thinkers can feel, does explain how they can do what they can do.

♣ HARNAD: Yes, and this is why the other-minds problem comes into its own in doing Turing testing of machines rather than in doing mind reading of our own species and other animals. ("Animate" is a weasel word, though, for vitalists are probably also animists; Harnad, 1994a.)

♠ FORD, GLYMOUR, AND HAYES: Turing's source for this view is unknown. The contrary opinion is given in the Qu'ran.

♠ FORD, GLYMOUR, AND HAYES: The last three sentences are a bit odd. We only acquired our present knowledge because many people (Galileo himself, Bruno before him, Kepler, Newton, etc.) already found the argument futile.

3.6.2 The "Heads in the Sand" Objection

"The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so."

This argument is seldom expressed quite so openly as in the form above. But it affects most of us who think about it at all. We like to believe that Man is in some subtle way superior to the rest of creation. It is best if he can be shown to be *necessarily* superior, for then there is no danger of him losing his commanding position. The popularity of the theological argument is clearly connected with this feeling. It is likely to be quite strong in intellectual people, since they value the power of thinking more highly than others, and are more inclined to base their belief in the superiority of Man on this power.

I do not think that this argument is sufficiently substantial to require refutation. Consolation would be more appropriate: perhaps this should be sought in the transmigration of souls.♠

3.6.3 The Mathematical Objection

There are a number of results of mathematical logic which can be used to show that there are limitations to the powers of discrete-state machines. The best known of these results is known as Gödel's theorem (1931), and shows that in any sufficiently powerful logical system statements can be formulated which can neither be proved nor disproved within the system, unless possibly the system itself is inconsistent. There are other, in some respects similar, results due to Church (1936), Kleene (1935), Rosser, and Turing (1937). The latter result is the most convenient to

♠ Ford, Glymour, and Hayes: Turing raises an issue still central in our own time: the limitation of scientific inquiry by religious dogma, and in particular by the doctrine of souls. The fundamental religious objection to embryonic stem cell research is that when a human sperm cell and an ovum form an embryonic cell, the cell is "ensouled," it supernaturally acquires a soul. In this subsection, irritation or exasperation seems to have overwhelmed Turing's usual ingenuity in argument. While many current advocates of "Heads in the Sand" may be utterly thoughtless, there is a history of arguments for the position, all of which Turing ignores. William James, in *The Will to Believe*, argued roughly as follows: we should not believe that human intelligence has a purely biological, chemical, and physical explanation, for if we did so believe, we would conclude there is no basis for moral assessment; the world would be a worse place if we believed there is no basis for moral assessment, and it is rational not to act to bring about the worse case. The argument is in the spirit of Pascal's Wager. Pascal, the Turing of the 17th century, argued that one should act so as to cause oneself to believe in God, because the expected payoff of believing is infinitely greater than the expected payoff of not believing. We think neither James' argument nor Pascal's is sound, but the arguments deserve at least as much consideration as others Turing does respond to.

consider, since it refers directly to machines, whereas the others can only be used in a comparatively indirect argument: for instance, if Gödel's theorem is to be used we need in addition to have some means of describing logical systems in terms of machines, and machines in terms of logical systems. The result in question refers to a type of machine which is essentially a digital computer with an infinite capacity. It states that there are certain things that such a machine cannot do. If it is rigged up to give answers to questions as in the imitation game, there will be some questions to which it will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply. There may, of course, be many such questions, and questions which cannot be answered by one machine may be satisfactorily answered by another. We are of course, supposing for the present that the questions are of the kind to which an answer "Yes" or "No" is appropriate, rather than questions such as "What do you think of Picasso?" The questions that we know the machines must fail on are of this type, "Consider the machine specified as follows.… Will this machine ever answer "Yes" to any question?" The dots are to be replaced by a description of some machine in a standard form, which could be something like that used in §5. When the machine described bears a certain comparatively simple relation to the machine which is under interrogation, it can be shown that the answer is either wrong or not forthcoming. This is the mathematical result: it is argued that it proves a disability of machines to which the human intellect is not subject.

The short answer to this argument is that although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intellect.♣♥ But I do not think this view can be dismissed quite so lightly. Whenever one of these machines is asked the appropriate critical question, and gives a definite answer, we know that this answer must be wrong, and this gives us a certain feeling of superiority. Is this feeling illusory? It is no doubt quite genuine, but I do not think too much importance should be attached to it. We too often give wrong answers to questions ourselves to be justified in being very pleased at such evidence of fallibility on the part of the machines. Further, our superiority can only be felt on such an occasion in relation to the one machine over which we have scored our petty triumph. There would be no question of triumphing simultaneously over *all* machines. In short, then, there might be men cleverer than any

---

♣ HARNAD: Gödel's theorem shows that there are statements in arithmetic that are true, and we know are true, but their truth cannot be computed. Some have interpreted this as implying that "knowing" (which is just a species of "thinking") cannot be just computation. Turing replies that maybe the human mind has similar limits, but it seems to me it would have been enough to point out that "knowing" is not the same as "proving". Gödel shows the truth is unprovable, not that it is unknowable. There are far better reasons for believing that thinking is not computation.

♥ SAYGIN: I do not see this issue discussed much anywhere, but I think it is a profound idea. How do we know that the human brain-computer would "halt" given the description of another human brain-computer and asked what it would reply to a given input?