The invention and development of the computer will un-
doubtedly rank as one of the twentieth century’s most far-
reaching achievements that will ultimately rival or even sur-
developed a simple mathematical model for a universal com-
puting machine (today referred to as a Turing Machine). The
puting machine (today referred to as a Turing Machine). The

The Turing Test: the first 50 years

Robert M. French

The Turing Test, originally proposed as a simple operational definition of intelligence, has now been with us for exactly half a century. It is safe to say that no other single article in computer science, and few other articles in science in general, have generated so much discussion. The present article chronicles the comments and controversy surrounding Turing’s classic article from its publication to the present. The changing perception of the Turing Test over the last 50 years has paralleled the changing attitudes in the scientific community towards artificial intelligence: from the unbridled optimism of 1960s to the current realization of the immense difficulties that still lie ahead. I conclude with the prediction that the Turing Test will remain important, not only as a landmark in the history of the development of intelligent machines, but also with real relevance to future generations of people living in a world in which the cognitive capacities of machines will be vastly greater than they are now.
one of the very first electronic, programmable, digital computers. Finally, his third contribution was philosophical: he provided an elegant operational definition of thinking that, in many ways, set the entire field of artificial intelligence (AI) in motion. In this article, I will focus only on this final contribution, the Imitation Game, proposed in his classic article in Mind in 1950 (Ref. 1).

The Imitation Game
Before reviewing the various comments on Turing’s article, I will briefly describe what Turing called the Imitation Game (called the Turing Test today). He began by describing a parlour game. Imagine, he says, that a man and a woman are in separate rooms and communicate with an interrogator only by means of a teletype – the 1950s equivalent of today’s electronic ‘chat’. The interrogator must correctly identify the man and the woman and, in order to do so, he may ask any question capable of being transmitted by teletype. The man tries to convince the interrogator that he is the woman, while the woman tries to communicate her real identity. At some point during the game the man is replaced by a machine. If the interrogator remains incapable of distinguishing the machine from the woman, the machine will be said to have passed the Test and we will say that the machine is intelligent. (We see here why Turing chose communication by teletype – namely, so that the lack of physical features which Turing felt were not essential for cognition, would not count against the machine.)

The Turing Test, as it rapidly came to be described in the literature and as it is generally described today, replaces the woman with a person of either gender. It is also frequently described in terms of a single room containing either a person or a machine and the interrogator must determine whether he is communicating with a real person or a machine. These variations do, indeed, differ somewhat from Turing’s original formulation of his imitation game. In the original test the man playing against the woman, as well as the computer that replaces him, are both ‘playing out of character’ (i.e. they are both relying on a theory of what women are like). The modern description of the Test simply pits a machine in one room against a person in another. It is generally agreed that this variation does not change the essence of Turing’s operational definition of intelligence, although it almost certainly makes the Test more difficult for the machine to pass. One significant point about the Turing Test that is often misunderstood is that finding it proves nothing. Many people would understandably fail if they were put in the role of the computer, but this certainly does not prove that they are not intelligent! The Turing Test was intended only to provide a sufficient condition for intelligence.

To reiterate, Turing’s central claim is that there would be no reason to deny intelligence to a machine that could flawlessly imitate a human’s unrestricted conversation. Turing’s article has unquestionably generated more commentary and controversy than any other article in the field of artificial intelligence, and few papers in any field have created such an enduring reaction. At least 13 years after Turing’s article appeared, Anderson had already counted over 1000 published papers on whether machines could think. For half a century, references to the Turing Test have appeared regularly in artificial intelligence journals, philosophy journals, technical treatises, novels and the popular press. Type ‘Turing Test’ into any Web browser and you will have thousands of hits. Perhaps the reason for this high profile is partly our drive to build mechanical devices that imitate what humans do. However, there seems to be a particular fascination with mechanising our ability to think. The idea of mechanised thinking goes back at least to the 17th century with the Characteristica Universalis of Leibnitz and extends through the work of La Mettrie to the writings of Hobbes, Pascal, Boole, Babbage and others. The advent of the computer meant that, for the first time, there was a realistic chance actually achieving the goal of mechanized thought. It is this on-going fascination with mechanised thought that has kept the Turing Test in the forefront of discussions about AI for the past half century.

The value and the validity of the Turing Test
Opinions on the validity and, especially, the value of the Turing Test as a real guide for research vary widely. Some authors have maintained that it was precisely the operational definition of intelligence that was needed to sidestep the philosophical quagmire of attempting to define rigorously what was meant by ‘thinking’ and ‘intelligence’ (see Refs 4–7). At the other extreme, there are authors who believe that the Turing Test is, at best, ‘poor’ and, at worst, a real impediment to progress in the field of artificial intelligence9,10. Hayes and Ford claim that abandoning the Turing Test as an ultimate goal is ‘almost a requirement for any rational research program which declares itself interested in any particular part of cognition or mental activity’. Their non-reasonable view is that research time is better spent developing what they call ‘a general science of cognition’ that would focus on more restricted areas of cognition, such as analogy-making, vision, generalization and categorization abilities. They add, from a practical perspective, why would anyone want to build machines that could pass the Turing Test? Human cognition, even high-quality human cognition, is not in short supply. What extra functionality would such a machine provide?


I am not sure exactly what Whaley means by ‘consigned to history’, but if he means ‘forgotten’, I personally doubt that this will be the case. I believe that in 300 years’ time people will be discussing the arguments raised by Turing in his paper. It could even be argued that the Turing Test will take on an even greater significance several centuries in the future when it might serve as a moral yardstick in a world where machines will move around much as we do, will use natural language, and will interact with humans in ways that are almost inconceivable today. In short, one of the questions facing future generations may well be: ‘To what extent do machines have to act like humans before it becomes immoral to damage
or destroy them? And the very essence of the Turing Test is our judgment of how well machines act like humans.

Shift in perception of the Turing Test

It is easy to forget just how high the optimism once ran for the rapid achievement of artificial intelligence. In 1958, a mere eight years after the appearance of Turing’s article, when computers were still in their infancy and even high-level programming languages had only just been invented, Simon and Newell1, two of the founders of the field of artificial intelligence, wrote, ‘...there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until – in a visible future – the range of problems they can handle will be co-extensive with the range to which the human mind has been applied’. Minsky, head of the MIT AI Laboratory, wrote in 1967, ‘Within a generation the problem of creating “artificial intelligence” will be substantially solved’.

During this period of initial optimism, most of the authors writing about the Turing Test shared with the founders of AI the belief that a machine could actually be built that would be able to pass the Test in the foreseeable future. The debate, therefore, centered almost exclusively around Turing’s operational definition of disembodied intelligence – namely, did passing the Turing Test constitute a sufficient condition for intelligence or not? As it gradually dawned on AI researchers just how difficult it was going to be to produce artificial intelligence, the focus of the debate on the Turing Test shifted. By 1982, Minsky’s position regarding artificial intelligence had undergone a radical shift from one of unbounded optimism 15 years earlier to a far more sober assessment: ‘Defenders of the computing machine analogy seem implicitly to assume that the whole of intelligence and thought can be built up summatively from the warp and woof of formal systems and, indeed, may also have a limit to their truths they can recognize. The second objection is the “argument from consciousness” or the “problem of other minds”. The only way to know if anything is thinking is to be that thing, so we cannot know if anything else really thinks. Turing’s reply was that if we adopt this solipsistic position for a machine, we must also adopt it for other people, and few people would be willing to do that.

Finally, the most important objection that Turing raised was what he calls ‘Lady Lovelace’s objection’. The name of this objection comes from a remark by Lady Lovelace concerning Charles Babbage’s ‘Analytical Engine’, and was paraphrased by Turing as ‘the machine can only do what we know how to order it to do’. In other words, machines, unlike humans, are incapable of creative acts because they are only following the programmer’s instructions. His answer is, in essence, that although we may program the basics, a computer, especially a computer capable of autonomous learning (see section 2 of Turing’s article), might well do things that could not have been anticipated by its programmer.

A brief chronicle of early comments on the Turing Test

Mays wrote one of the earliest replies to Turing, questioning the fact that a machine designed to perform logical operations could actually capture ‘our intuitive, often vague and imprecise, thought processes’10. Importantly, this paper contained a first reference to a problem that would take center stage in the artificial intelligence community three decades later: ‘Defenders of the computing machine analogy seem implicitly to assume that the whole of intelligence and thought can be built up summatively from the warp and woof of formal systems’. Gunderson clearly believed that passing the Turing Test would require ‘a very large range of other properties’20.

In a paper written in the early 1970s, Lucas wrote that intelligence, written a decade later when the difficulties with AI research had become clearer, criticized Gunderson’s single-skill objection, insisting that to play the game would require ‘a very large range of other properties’.

In articles written in the early 1970s we see the first shift away from the acceptance that it might be possible for a machine to pass the Turing Test. Even though Puttill’s basic objection16 to the Turing Test was essentially the Lady Lovelace objection (i.e. that any output is determined by what the programmer explicitly put into the machine, and therefore can be explained in this manner), he suggested that thinking is a very broad concept and that a machine passing the Imitation Game is merely ‘imitating human behaviour’.

In 1972, Minsky gave a talk at the IBM Watson Research Center in Yorktown Heights, New York, titled ‘The Imitation Game’, in which he made the bold claim that machines could pass the Imitation Game if we were to consider a supercomputer as to make it wholly unreasonable to deny that it had feelings.

Gunderson clearly believed that passing the Turing Test would not necessarily be a proof of real machine intelligence. In articles written in the early 1970s we see the first shift away from the acceptance that it might be possible for a machine to pass the Turing Test. Even though Puttill’s basic objection16 to the Turing Test was essentially the Lady Lovelace objection (i.e. that any output is determined by what the programmer explicitly put into the machine, and therefore can be explained in this manner), he concluded his paper in a particularly profound manner, thus: ‘...if a
computer could play the complete, “any question” imitation game it might indeed cause us to consider that perhaps the computer was capable of thought. But that any computer might be able to play such a game in the foreseeable future is so immensely improbable as to make the whole question academic.34 Sampson replied that low-level determinism (i.e. the program and its inputs) does not imply predictable high-level behaviour.35 Two years later, Millar presented the first explicit discussion of the Turing Test’s anthropomorphism: ‘Turing’s test forces us to ascribe typical human objectives and human cultural background to the machine, but if we are to be serious in contemplating the use of such a term [intelligence] we should be open-minded enough to allow computing machinery or Martians to display their intelligence by means of behaviour which is well-adapted for achieving their own specific aims’.36 More agreed that passing the test would constitute a sufficient proof of intelligence.37 He viewed the Test as a potential source of good inductive evidence for the hypothesis that machines can think, rather than as a purely operational definition of intelligence. However, he suggested that it is of little value in guiding real research on artificial intelligence. Stalker replied that an explanation of how a computer passes the Turing Test would require an appeal to mental, not purely mechanical notions.38 More recently it has been demonstrated that these two explanations are not necessarily competitors.39

Comments from the 1980s

Numerous papers on the Turing Test appeared at the beginning of the 1980s, among them one by Hofstadter.40 This paper covers a wide range of issues and includes a particularly interesting discussion of the ways in which a computer simulation of a hurricane differs or does not differ from a real hurricane. (For a further discussion of this point, see Ref. 28.) The two most often cited papers from this period were by Block41 and Searle.42 Instead of following up the lines of inquiry opened by Putnam43 and Millar,44 these authors continued the standard line of attack on the Turing Test, arguing that even if a machine passed the Turing Test, it still might not be intelligent. The explicit assumption was, in both cases, that it was, in principle, possible for machines to pass the Test.
Many replies have been made to this argument and I will believe that the person inside the room understands Chinese. No Chinese whatsoever. The Chinese person would therefore provide a perfect response, even though the English-speaker understood the Chinese person on the outside of the room would see a string of symbols that constitutes an answer to the question. He copies this answer on a piece of paper and sends it into the room. The room is full of a native Chinese person writes a question in Chinese on a piece of paper and sends it into the room. The room is full of symbolic rules specifying inputs and outputs. The English-speaker then matches the symbols in the question with symbols in the rule base. This does not have to be a direct table matching of the string of symbols in the question with symbols in the rule base, but can include any type of look-up program, regardless of its structural complexity. The English-speaker is blindly led through the maze of rules to a string of symbols that constitutes an answer to the question. He copies this answer on a piece of paper and sends it out of the room. The Chinese person on the outside of the room would see a perfect response, even though the English-speaker understood no Chinese whatsoever. The Chinese person would therefore believe that the person inside the room understands Chinese. Many replies have been made to this argument and I will not include them here. One simple refutation would be to ask how the room could possibly contain answers to questions that containedcaricatured distorted characters. So, for example, assume the last character in a question had been distorted in a very phallic manner (but the character is still clearly recognizable to a native Chinese person). The question sent into the room is: "Would the last character in this question be likely to embarrass a very shy young woman?"

Now, to answer this question, all possible inputs, including all possible distortions of those inputs, would have to be contained in the rules in the room. Combinatorial explosion, once again, brings down this line of argument.

Could any machine ever pass the Turing Test? In the 1980s, Hauser suggested the extreme difficulty of achieving the Turing Test6. He accepted it as a sufficient condition for intelligence, but wrote that, 'A failure to think imaginatively about the test actually proposed by Turing has led many to underestimate its severity...'. He suggests that the Turing Test, when we think of just how hard it would be to pass, also shows why AI has turned out to be so hard.

At the 1980s ended, a new type of discussion about the Turing Test appeared, one that reflected not only the difficulties of traditional, symbolic AI but also the range of interest in sub-symbolic AI fueled by the ideas of connectionism33,34. These new ideas were the basis of work by French35,36 that sought to show, by means of a technique based on sub-cognitive questions (see Box 1), that only a computer that had acquired adult human intelligence by experiencing the world as we have could pass the Turing Test37. Further, he argued that any attempt to fix the Turing Test so that it could test for intelligence in general and not just human intelligence is doomed to failure because of the completely interwoven and interdependent nature of the human physical, subcognitive, and cognitive levels36. French also emphasized the fact that the Turing Test, when rigorously administered, probes deep levels of the associative concept networks of the candidates and that these networks are the product of a lifetime of interaction with the world which necessarily crosses human sense organs, their location on the body, their sensitivity to various stimuli, etc.38. A similar conclusion was reached by Davidson, who wrote, 'Turing wanted his Test to draw a fairly sharp line between the physical and the intellectual capacities of man.' There is no such line39.

In the past decade, Hamad has been one of the most prolific writers on the Turing Test40. Most importantly he has proposed a Total Turing Test' (TTT) in which the screen provided by the teletype link between the candidates and the interrogator is removed41. This is an explicit recognition of the importance of bodies in an entity's interaction with the environment. The heart of Hamad's argument is that mental semantics must be 'grounded', in other words, the meanings of internal symbols must derive, at least partly, from interactions with the external environment42. Shanon also recognized the necessity of an interaction with the environment43. However, Hauser argued that the switch from the normal Turing Test to the TTT is unwarranted44. In later papers, Hamad extended this notion by defining a hierarchy of Turing Tests (see Box 2) of which the second (T2: the symbols-in/symbols-out Turing Test) corresponds to the standard Turing Test. T3 (the Total Turing Test) is the Robotic Turing Test in which the interrogator directly, visually, tactically, addresses the two candidates – the teletype 'screening' mechanism is eliminated. But we might still be able to detect some internal differences, even if the machine passed T3. Therefore, Hamad proposes T4: Internal Microfunctional Indistinguishability. And finally, T5: Grand Unified Theories of Everything, where the two candidates would be microfunctionally equivalent by every test relevant to a neurologist, neurophysiologist, and neurobiophysicist (for example, both fully obey the Hodgkin-Huxley equations governing neuronal firing) but would nonetheless be distinguishable to a physical chemist.

Hamad clearly recognizes the extreme difficulty of achieving even T2 and stresses the impossibility of implementing disembodied cognition. Schweizer wishes to improve the Robotic Turing Test (T3) by proposing a Truly Total Turing Test in which a long-term temporal dimension is added to
**Box 2. The Turing Test hierarchy**

Stewart Harnad has proposed a five-level Turing Test (TT) hierarchy (Ref. 4). This hierarchy attempts to encompass various levels of difficulty in playing an Imitation Game. The levels are T1, T2, T3, T4, and T5. The Harnad hierarchy works as follows:

**Level T1**
The ‘toy model’ level. There are modules (‘toy’s) that only handle a fragment of our cognitive capacity. So, for example, Colby’s program designed to imitate a parrotal aphonic would fall into this category, because ‘the TT is predicated on total functional indistinguishability, and toys are mostly distinguishable from the real thing’.

**Level T2**
This is the level described in Turing’s original article. Harnad refers to it as the ‘pen-pal version’ of the Turing Test, because all exchanges are guaranteed by the teletype link to occur in a symbol-in/symbol-out manner. Thus, T2 calls for a system that is indistinguishable from us in its symbolic (i.e. linguistic) capacities. This is also the level for which Searle’s Chinese Room experiment is written. One central question is to what extent questions at this level can be answered successfully, but indirectly, to probe the deep levels of cognitive, or even physical structure of the candidates.

**Level T3**
The ‘Total Turing Test’ (or the robotic Turing Test) At this level the ‘teletype screen’ is removed. T3 calls for a system that is not only indistinguishable from us in its symbolic capacities, but it further requires indistinguishability in all of our mental capacities: in other words, indistinguishability in ontological (i.e. behavioral) function. At this level, physical appearance and explicitly observable behaviors matter.

**Level T4**
‘Microfunctional Indistinguishability’
This level would call for functional indistinguishability, right down to the last neuron and neurotransmitter. This could be synthetic neurons, of course, but they would have to be functionally indistinguishable from real ones.

**Level T5**
‘Grand Unified Theory of Everything (GUTE)’
At this level the candidates are ‘empirically identical in kind, right down to the last electron, but there remain unobservable-in-principle differences at the level of their designers’ GUTEs.

Harnad feels that T3 is the right level for true cognitive modeling. He writes, ‘My own guess is that if ungrounded T2 systems are undetermined and hence open to over-interpretation, T3 systems are overdetermined and hence include physical and functional properties that may be indiscernible to cognition. I think T5 is just the right empirically filtered for mind-modeling.’

**References**

Many of these objections concerning the difficulty of making an actual machine that could pass the Turing Test are also voiced by Crockett in his discussion of the relationship of the Turing Test to the famous frame problem in AI (i.e. the problem of determining exactly what information must remain undetected at a representational level within a system after the system has performed some action that affects its environment)24. In essence, Crockett claims that passing the Turing Test is essentially equivalent to solving the frame problem (see also Ref. 49). Crockett arrives at essentially the same conclusion as French: ‘I think it is unlikely that a computer will pass the test…because I am particularly impressed with the test’s difficulty which is more difficult and anthropocentric than even Turing fully appreciated’.

Mitchie introduced the notion of ‘superarticulacy’ into the debate25. He claims that for certain types of phenomena that we view as purely intuitive, there are, in fact, rules that can explain our behavior, even if we are not consciously aware of them. We could understand the computer in a Turing Test because, if we gave the machine rules to answer certain types of subcognitive questions—e.g. questions that have no general formulation, but only particular answers—’the machine would be able to explain the how it gave these answers, but we humans could not, or at least our explanations would not be the one given by the computer. In this way we
could catch the computer out and it would fail the Turing Test. The notion of superarticulacy is particularly relevant to current cognitive science research. Our human ability to know something without being able to articulate that knowledge, or to learn something (as demonstrated by an ability to perform a particular task) without being aware that we have learned it, is at present a very active line of research in cognitive science.

In a recent and significant comment on the Turing Test, Watt proposed the Inverted Turing Test (ITT) based on considerations from “naive psychology”18—our human tendency and ability to ascribe mental states to others and to ourselves. In the ITT, the machine must show that its tendency to ascribe mental states to others (in a manner roughly comparable to that of a real human. A machine will be said to pass the ITT if it is “unable to distinguish between two humans, or between a human and a machine that can pass the normal TT, but which can discriminate between a human and a machine that can be told apart by a normal TT with a human observer”19. There are numerous replies to this proposal20-22. It can be shown, however, that the ITT can be simulated by the standard Turing Test18. French used the technique of a ‘Human Subcognitive Profile’ (i.e. a list of subcognitive questions whose answers have been gathered from people in the large population, see Box 1) to show that a mindless program using the Profile could pass this variant of the Turing Test23. Ford and Hayes24 renewed their appeal to reject this type of test as any kind of meaningful yardstick for AI. Collins suggested his own type of test, the Editing Test25, based on ‘the skilled way in which humans “repair” deficiencies in speech, written texts, handwriting, etc., and the failure of computers to achieve the same interpretative competence’26.

Loebner Prize

An overview of the Turing Test would not be complete without briefly mentioning the Loebner Prize26—27, which originated in 1991. The competition stipulates that the first program to pass an unrestricted Turing Test will receive $100,000. For the Loebner Prize, both humans and machines answer questions by the judges. The competition, however, is among the various machines, each of which attempts to fool the judges into believing that it is a human. The machine that best plays the role of a human wins the competition.

Initially, restrictions were placed on the form and content of the questions that could be asked. For example, questions were restricted to specific topics, judges who were computer scientists were disallowed, and ‘trick questions’ were not permitted.

There have been numerous attempts at ‘restricted’ simulations of human behaviour over the years, the best known probably being Colby’s PARRY28,29, a program that simulates a paranoid schizophrenic by means of a large number of canned routines, and Weizenbaum’s ELIZA30, which simulates a psychiatrist’s discussion with patients.

Aside from the fact that restricting the domain of allowable questions violates the spirit of Turing’s original anything-goes’ Turing Test, there are at least two major problems with domain restrictions in a Turing Test. First, there is the virtual impossibility of clearly defining what does and does not count as being part of a particular real-world domain. For example, if the domain were International Politics, a question like, ‘Did Ronald Reagan wear a shirt when he met with Mikhail Gorbachev?’ would seem to qualify as a ‘trick question’, being pretty obviously outside of the specified domain. But now change the question to, ‘Did Mahatma Gandhi wear a shirt when he met with Winston Churchill?’ Unlike the first, the latter question is squarely within the domain of international politics because it was Gandhi’s practice, in order to make a political/cultural statement, to be shirtless when meeting with British statesmen. But how can we differentiate these two questions a priori, accepting one as within the domain of international politics, while rejecting the other as outside of it? Further, even if it were somehow possible to clearly delineate a real domain, what would determine whether a domain were too restricted? In a tongue-in-cheek response to Colby’s claims that PARRY had passed something that could rightfully be called a legitimate Turing Test, Weizenbaum claimed to have written a program for another restricted domain: infant autism31. His program, moreover, did not even require a computer to run on; it could be implemented on an electric typewriter. Regardless of the question typed into it, the typewriter would just sit there and hum. In terms of the domain-restricted Turing Test, the program was indistinguishable from a real autistic infant. The deep point of this example is the problem with domain restrictions in a Turing Test.

To date, nothing has come remotely close to passing an unrestricted Turing Test and, as Dennett, who agreed to chair the Loebner Prize event for its first few years, said, “...passing the Turing Test is not a sensible research and development goal for serious AI32. Few serious scholars of the Turing Test, myself included, take this competition seriously and Minsky has even publicly offered $100 for anyone who can convince Loebner to put an end to the competition33. (For those who wish to know more about the Loebner Competition, refer to Ref. 57.)

There are numerous other commentaries on the Turing Test. Two particularly interesting comments on actually building truly intelligent machines can be found in Dennett34 and Waltz35.

Conclusions

For 50 years the Turing Test has been the object of debate and controversy. From its inception, the Test has come under fire as being either too strong, too weak, too anthropocentric, too broad, too narrow, or too coarse. One thing, however, is certain: gradually, indelibly, we are moving into a world where machines will participate in all of the activities that have hitherto been the sole province of humans. While it is unlikely that robots will ever perfectly simulate human beings, one day in the far future they might indeed have sufficient cognitive capacities to pose certain ethical dilemmas for us, especially regarding their destruction or exploitation. To resolve these issues, we will be called upon to consider the question: ‘how much are these machines really like us?’ and I predict that the yardstick that will be used measure this similarity will look very much like the test that Alan Turing invented at the dawn of the computer age.
Acknowledgements

The present paper was supported in part by research grant EU/PA/91 from the Belgian government. I am grateful to Dan Dennett and Steven Harnad for their particularly helpful comments on an earlier draft of this review.

References


Do you want to reproduce material from Trends in Cognitive Sciences?

This publication and the individual contributions within it are protected by the copyright of Elsevier Science Ltd. Except as outlined in the copyright statement (see p. ix), no part of Trends in Cognitive Sciences may be reproduced, either in print or electronic form, without written permission from Elsevier Science Ltd. Please address any permission requests to:

Elsevier Science Global Rights Department, PO Box 800, Oxford, UK. OX5 1DX