Philosophy of Science (2026), 1–20
doi:10.1017/psa.2026.10282

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

ARTICLE

Why No One Has Implemented the Turing Test: Identifying Turing's Design and Putting it into Practice

Diane Proudfoot[1], Sharon Temtsin[2] and Christoph Bartneck[2]

[1]Philosophy, University of Canterbury, Christchurch, New Zealand and [2]Computer Science and Software Engineering, University of Canterbury, Christchurch, New Zealand
Corresponding author: Diane Proudfoot; Email: diane.proudfoot@canterbury.ac.nz

(Received 23 January 2026 UTC; revised 09 August 2026 UTC; accepted 10 August 2026 UTC)

Abstract

In October 1950, Alan Turing proposed that a machine's success in an imitation game provides a criterion for thinking. Turing's test became one of the philosophical foundations of AI, but three-quarters of a century later, it is widely claimed that generative AI models pass the test, rendering it obsolete. Our aim here is to establish definitively Turing's conceptual design for his game. We argue that influential interpretations of Turing's design are mistaken, with the consequence that his test has *not* been passed. We also present the results of carrying out, for the first time, Turing's experiment as he proposed it.

1. Introduction

In his famous October 1950 paper in *Mind*, "Computing Machinery and Intelligence", Turing described an imitation game involving a "question and answer method" and proposed that success in this game provided a "criterion for 'thinking'" in machines (1950, 435, 436).[1] Turing's paper was a dramatic contrast to other articles in the journal in 1950 and the years immediately prior; the few papers that might be described as on philosophy of mind typically discussed the epistemology of sense data, selves as particulars or substances, mental acts as dispositions, psycho-physical parallelism, and souls—nothing on learning machines or arguments for or against the possibility of artificial intelligence (AI). Perhaps for this reason, some early replies to Turing's paper did not take his proposal very seriously. Only a few months later, Leonard Pinsky whimsically proposed an alternative test for intelligence in machines. He suggested that one of Turing's computers be sent, as a "Christmas gift", a

[1] There is no evidence that, in discussing his test, Turing made a distinction between testing for "intelligence" and testing for "thinking"; in this paper, we follow Turing's practice. Likewise, we make no distinction between AI and AGI ("artificial general intelligence").

© The Author(s), 2026. Published by Cambridge University Press on behalf of Philosophy of Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

"subscription to MIND, retroactive to October, 1950" (Pinsky 1951, 397–98). The computer, on reading Turing's article, might realize that "it is thinking about the possibility of machines thinking" and consequently suffer "a nervous breakdown"—if the computer could be cured by the reassurance that it was "only tussling with a pseudo-problem", this really would demonstrate that a machine thinks, Pinsky declared (1951, 398).

Sober treatments of Turing's proposed test appeared quite quickly. On the one hand, in April 1952, Wolfe Mays claimed that Turing was a "mechanical necromancer", in whose account "the meaning of the word 'thinking' has changed to such an extent that it has little in common with what we normally mean by it"; Turing's "behaviouristic" approach failed to see that what is important is not "what" the machine does, "but how it does it", Mays claimed (1952, 150–53, 158). Several early discussions pressed Mays' first point; for example, Michael Scriven claimed in 1953 that the expression "intelligent" was "an improper adjective for the noun 'computer'", and so talk of intelligent computers implied "a change of meaning in one of these terms" (1953, 236). On the other hand, in 1951 Donald MacKay said that there is no distinction in principle between the "observable behaviour of a human brain" and the "behaviour possible in a suitably designed artefact", and in 1952 Henry Cohen stated that arguments against the possibility of thinking machines "lead inevitably, as Turing has shown, to a naïve solipsism" (MacKay 1951, 120; Cohen 1952, 206). Subsequently, F. H. George claimed that Mays had made "no impression whatsoever on Turing's position"—and, echoing Turing's sentiments in *Mind*, that "much dispute" over the capabilities of machines was due merely to "feelings of an emotional kind" (1956, 248; 1958, 59).[2]

Three-quarters of a century later, the verdict on Turing's test is often cavalier. For example, an article in October 2025 in *Nature* claimed that large language models (LLMs) "killed the Turing test": "[c]hatbots now ace the mathematician's famous imitation game", it proclaimed (Gibney 2025). The evidence cited was a 2025 study stating that two LLMs "pass the Turing test" (Jones and Bergen 2025).[3] Several recent studies making the same claim have prompted commentators to declare that the test is obsolete—either because we no longer need an existence proof of the possibility of AI, or (assuming these LLMs do *not* think) because their success in the test counterexamples Turing's claim to have provided a criterion for thinking in machines. The vital question here, however, is whether these machines have passed *Turing's* test, or merely *some* imitation test (Copeland 2000, 2004a).

Our first aim in the present paper is, taking Turing's descriptions of his test, to establish definitively his conceptual design for the test. We argue that influential interpretations of his design are incompatible with Turing's own words; these readings also conflict with the early perception and historical context of his ideas.[4] A consequence is that any claim that *Turing's* test has been passed by unintelligent

[2] In Mind, Turing criticized people who oppose the idea of AI as "too dreadful", who "hope" that AI does not arrive, or who fear that humans will lose their "commanding position" (1950, 444).

[3] Specifically, GPT-4.5 and LLaMa-3.1-405B, "when they are given prompts to adopt a humanlike persona" (Jones and Bergen 2025). For additional studies making similar claims, see Temtsin et al. (2025).

[4] Numerous theorists have contributed to the voluminous literature on Turing's test: in this paper, we discuss only the first, best-known, or most influential proponents of the relevant interpretations.

LLMs, and his criterion for thinking thereby rebutted, is false. Our second aim is to present the results of implementing Turing's conceptual design for the first time (full technical specifications of our study, data, and analyses are available in Temtsin et al. 2025). The results of our study do *not* support a "general consensus", announced at the October 2025 Royal Society celebration of the Turing test's 75th anniversary, that AI has passed the test.[5] When we tested a generative pre-trained transformer (GPT) machine, the machine's performance was extremely poor.

The paper is in two parts. The first part (sections 2–4) aims to identify Turing's conceptual design for his test on the basis of exegetical, historical, and experimental considerations.[6] From Turing's descriptions, we extract informal specifications of the structure, scoring method, and essential features of his imitation game. The second part (section 5) presents (informally) our new empirical study implementing Turing's design and the philosophical lessons to be derived from it. The paper, we argue, generates a challenge for AI experimenters: to earn the cachet of passing *Turing's* test, their machine must do well in the imitation game presented here.

1.1 Two preliminary points

Discussions of the Turing test often begin with some claim concerning Turing's philosophical commitments—roughly, either to behaviourism, some classical or non-classical version of computationalism, or (more recently) response-dependence concerning judgements of mind. Our aim is to identify Turing's conceptual design *without* tying it to any specific theory of mind, in order that his blueprint may be useful to theorists, regardless of their interpretations of Turing's philosophy. (We turn briefly to Turing's philosophy of mind at the beginning of section 5.) For the same reason, we make no assumptions regarding the nature of the concept of intelligence, or of the property of being intelligent, for which success in Turing's game is to suffice.

In addition, many discussions deny that Turing intended his test as an actual experiment to be performed. For example, at the Royal Society celebration, it was stated that Turing was only "kidding" in his 1950 paper, and that we should not take the paper "too seriously".[7] It has been claimed that Turing intended his test merely as a "joke" (Minsky 2013), a "philosophical conversation-stopper" (Dennett 1985), or as a rhetorical device in order to counter popular objections to the possibility of AI (Sloman 2013). However, these claims neglect the fact that Turing discussed his test, not only in the *Mind* paper, but also in a 1952 radio symposium on machine learning, "Can Automatic Calculating Machines Be Said to Think?", and he anticipated his computer-imitates-human game in "Intelligent Machinery", an official 1948 report for the National Physical Laboratory (NPL) (Turing et al. 1952; Turing 1948; Copeland 2004a). In addition, these claims overlook the historical context: in the late 1940s and early 1950s, the emergence of the world's first electronic stored-program computers in the UK and the US had generated vigorous public and scientific debate on the

[5] Wendy Hall, at "Celebrating the 75th Anniversary of the Turing Test", The Royal Society and the Web Science Institute, October 2, 2025. https://www.youtube.com/watch?v=w6LrZu5ku_o.

[6] We confine discussion of Turing's sources and of the many early commentaries on his 1950 paper to those that are directly relevant to determining the conceptual design for his test.

[7] Sarah Dillon and Peter Millican at "Celebrating the 75th Anniversary of the Turing Test".

possibility of AI—Turing was aware that a criterion of intelligence in machines was badly needed.[8]

Some commentators claim that Turing intended his test only as a thought experiment (e.g., Mitchell 2024). However, depicting Turing as solely an armchair philosopher would be almost as inaccurate as characterizing him as solely a pure mathematician. Turing was interested in turning his abstract universal computing machine into concrete machines; at the NPL, for example, Turing and his engineering team were designing his Automatic Computing Engine (Copeland 2004b). In addition, he was interested in actual human reactions to computers. In his NPL report, he described a "little experiment" in which a human (C) plays chess with another human or with a "paper machine"—which is to say, a human being simulating a computer by following a set of rules, using pencil and paper, the only sort of programmable computer then available (1948, 431). C, Turing said, "may find it quite difficult to tell which he is playing" (1948, 431). This experiment anticipated in important respects the imitation game he announced two years later, and Turing ended his description of the chess game with the words, "This is a rather idealized form of an experiment I have actually done" (1948, 431).

2. Turing's man-imitates-woman game: Rival interpretations

In the second paragraph of his 1950 paper, Turing introduces "a game which we call the 'imitation game'", as follows:

> It is played with three people, a man (A), a woman (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". (1950, 433)[9]

After more discussion of this man-imitates-woman game, Turing said:

> 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 played like this as he does when the game is played between a man and a woman? (1950, 434)

[8] It is often assumed that this debate arose only as a result of the 1956 Dartmouth Summer Research Project on Artificial Intelligence, and that philosophical discussion of AI is a recent phenomenon. On the contrary, Turing was probably the first to discuss seriously the possibility of intelligence or thinking in machines, at Bletchley Park during World War II, and several of his writings refer to intelligent machinery or thinking machines. The first public lecture on developments in artificial intelligence (as we would now use the term), so far as we are aware, was given in the UK by Viscount Mountbatten at the end of October 1946 (Mountbatten 1946). As a consequence of this lecture, a vigorous professional and public debate began over the possibility and desirability of thinking machines (or "electronic brains"), not only in the UK, but also in the US, Europe, and beyond.

[9] In describing the imitation game, Turing made no distinction between sex and gender; we follow his practice.

Later in the paper Turing suggests that "the question, 'Can machines think?' should be replaced by 'Are there imaginable digital computers which would do well in the imitation game?'" (1950, 442). When a computer is a player, Turing says, the computer takes "the part of A in the imitation game, the part of B being taken by a man" (1950, 442).

A persistent and sometimes torturous debate has grown up around these remarks. In moving from a man-imitates-woman game to a new game involving a computer, is Turing proposing that, in the latter, (1) the interviewees are a computer and a woman, and the computer's task is to imitate a woman? Or is it that (2) the interviewees are a computer and a man (i.e., male or masculine human being), and the computer's task is to imitate a man? Or is it that (3) the interviewees are a computer and a human being (of any gender), and the computer's task is to imitate a human being? All three interpretations can be found in the recent scholarly literature as well as in popular discussions of Turing's test.

In an influential paper criticizing Turing's test as "actively harmful", Patrick Hayes and Kenneth Ford propose (1)—the computer-imitates-*woman* interpretation.[10] They say, "If we take [Turing] at his word, the test is rather clever":

> It has a woman and a machine each trying to convince the judge that they are a woman, and the judge's task is still to decide which is the woman and which, therefore, is not. But this judge is not thinking about the differences between women and machines, but between women and men. 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. We will call this the gender test. (Hayes and Ford 1995, 972)

The conception of Turing's test as a "gender test" is still prevalent (e.g., Sterrett 2020; Gonçalves 2024).[11] However, Turing nowhere explicitly mentioned a computer's imitating a woman. Nor did any of his 1952 co-symposiasts or the early 1950s' theorists responding to the *Mind* paper—the philosophers and scientists who could be expected to have most familiarity with Turing's ideas.

An influential proponent of (2), the computer-imitates-*man* interpretation, is Judith Genova. According to Genova, Turing held that "being a man is a matter of simulating (imitating, voguing) masculinity" and he "wanted to show the similarity of the two issues", gender and intelligence (1994, 317). Genova claims that "[t]extual considerations alone" argue for this reading (1994, 314).

Certainly, as quoted above, when a computer takes the role of A, the role of B is to be taken by "a man". However, at issue is whether Turing's use here of the word "man" is the "masculine" or the "generic" use; the latter use of the terms "man" and "he", to indicate a human being irrespective of sex, was common in Turing's time and

[10] Interpretation (1) is also implicit in Gunderson (1964).

[11] Susan Sterrett (2000, 2020) is the theorist most strongly associated today with this conception. In her view, regardless of Turing's intention, we can find in the *Mind* paper two distinct imitation games involving a computer, corresponding to interpretations (1) and (2) here—and as a corollary, two different tests of intelligence in machines. See further section 5.4.

did not come under sustained attack until the early 1970s (Curzan 2003; Earp 2012). It is evident that Turing used masculine terms generically in the *Mind* paper: after stating that the interrogator "may be of either sex", he used throughout the pronoun "he" to refer to the interrogator. In *Mind* he also remarks, for example: "Thinking is a function of man's immortal soul"; "God cannot make a man to have no soul"; and "the only way to know that a *man* thinks is to be that particular man" (1950, 443, 446).[12] Turing could expect his readers to have no difficulty in identifying the generic use, since they too commonly employed it. For example, Geoffrey Jefferson's Lister Oration, to which Turing responds in his 1950 paper, was titled "The Mind of Mechanical Man"; in this paper, Jefferson discussed "man's characteristic desire for knowledge", "the qualities of man's mind", and "the difference between man and beasts" (1949, 1105, 1106). The 1950s' commentaries on Turing's paper also used "man" as a generic term, and Mays illustrated the generic use when he remarked:

> From the point of view of logic there is little to choose (except in precision) between the performance of a man and a machine. A human being might well be looked at as a machine for producing logical conclusions. (1952, 158)

In Turing's paper, textual considerations argue against, rather than for, the computer-imitates-man interpretation. Most uses of the term "man" in the paper are generic. Turing's few masculine uses are signalled by a contrast with "woman" (e.g., "God has given an immortal soul to every man and woman" [1950, 443]). When Turing said that the part of B is to be taken by "a man", the contrast is with "digital computer"—this offers no reason to read the use of "man" as masculine.

So, Turing's own words, along with those of his milieu and readership, point away from both (1) and (2) and toward (3), the computer-imitates-*human* interpretation. This fits, as Copeland points out, with Turing's claim, in his 1951 radio lecture "Can Digital Computers Think?", that the point of the test was to determine if the computer can "imitate a brain" (Turing 1951, 485; Copeland 2004, 436). Also, it is supported by early 1950s' commentaries; for example, Pinsky said that the object of the imitation game is for the interrogator to determine "which of the participants is a human being and which a machine" (1951, 397).

Yet, it might be inquired: if the computer-imitates-human reading is correct, *why* did Turing introduce the man-imitates-woman game? Turing's biographer, Andrew Hodges, offers one answer:

> The whole point of this game was that a successful imitation of a woman's responses by a man would *not* prove anything. Gender depended on facts which were *not* reducible to sequences of symbols. In contrast, he [Turing] wished to argue that such an imitation principle did apply to "thinking" or "intelligence". If a computer, on the basis of its written replies to questions, could not be distinguished from a human respondent, then "fair play" would oblige one to say that it must be "thinking". (2014, 523)

[12] Similarly, in his 1948 report Turing said, for example: "it is possible to make machinery to imitate any small part of a man"; "in so far as a man is a machine"; and "a man when concentrating may behave like a machine" (1948, 420, 421).

Hodges' account, however, faces difficulties. Why would Turing choose a game where successful performance fails as a criterion as a model for a game where successful performance is to succeed as a criterion? Hodges provides no explanation. Nor does he provide any reason why Turing would make this choice without saying anything to make the "whole point" clear. On this, Hodges remarks only that the beginning of Turing's *Mind* paper was "one of the few passages of the paper that was not expressed with perfect lucidity" (2014, 523).

Amy Kind offers another explanation. In her view, it is "critical" that Turing begins with "a sex analogy" (2022, 63). According to Kind, Turing had a "mid-twentieth-century conception" of sex, such that it is "both binary and fixed", and he thought of intelligence too as an "all-or-nothing property", which encourages a "humancentric prejudice" (2022, 60, 63). However, Turing did not deny that intelligence may exist on a continuum; for example, he wrote of the "real brains" of non-human animals (1951, 483). Nor does the fact that the man-imitates-woman game presents a binary choice imply that (or that Turing believed that) sexual identity is binary—any more than a red/green colour blindness test implies that being red is a mind-independent, discrete property.

A simpler explanation of Turing's man-imitates-woman game is that it was part of his response to the general public's and scientific community's antipathy to the possibility of AI.[13] If Turing were to have any hope of persuading his audience of the reasonableness of his test, he needed to smooth its path. A man's imitating a woman was a familiar sight in mid-twentieth century Britain, in theatres and on BBC radio and television. According to *The Times*, female impersonation held "a deep and irresistible appeal" and was "a universal theme" bringing "multitudes" to watch.[14] The Cambridge Footlights and the Cambridge Amateur Dramatic Club regularly had female impersonators; the "very presence of men in female parts" in these productions, *The Times* said, generated a "feeling of lightheartedness".[15] In the late 1940s and early 1950s, an analogy with a man-imitates-woman game was a perfect vehicle to break the ice for Turing's radical proposal.

3. Scoring the computer-imitates-human game: Three competing frameworks

In Turing's *Mind* paper, "do[ing] well" in the computer-imitates-human game is a sufficient condition of intelligence or thinking in machines (1950, 436)—but what constitutes doing well? Three influential answers are given.

Numerous commentaries suppose that Turing's framework for scoring the game is to be found in this remark (P):

> 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

[13] We found no evidence for the frequent claim that Turing's imitation game originated in an actual parlour game. According to Mays, Turing's game "seems to be a variation of the radio game of Twenty Questions", a BBC Home Service programme beginning in 1947 (1952, 148).

[14] The Times. "Christmas in the Theatre." December 27, 1947, 6.

[15] The Times. "Footlights Dramatic Club." June 10, 1936, 14. At this time, University of Cambridge regulations required that one in every six CADC productions have an all-male cast (as reported in The Times, "The Theatres," November 9, 1936, 10).

> 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. (Turing 1950, 442)

Using this framework, a computer passes the test if it attains a "30 per cent or more deception rate" in "a five-minute interaction" between the interrogator and the two interviewees, the interrogator's time shared roughly equally between the interviewees (Shah and Warwick 2010, 452). These are the "strict Turing rules", advocates of this framework claim (Warwick and Shah 2016a, 990).

However, (P) plainly expresses a *prediction* (Copeland 2000; Moor 2001). Immediately before (P), Turing says that he will "explain first my own beliefs in the matter"; following (P), he states additional beliefs about the future of AI, ending with the words "I believe further that no useful purpose is served by concealing these beliefs" (1950, 442). A prediction may stipulate a scoring protocol, but it need not: for example, my prediction that a student will score no higher than 30% in their philosophy exam does not imply that 30% is the pass mark. Is there reason to think that (P) states *both* a prediction and the framework for scoring the game?[16] Not, at least, in the early responses to Turing's paper, which did not read (P) in this way. Of this part of Turing's text, Mays said only: "Turing is firmly confident that in fifty years time it will be possible to programme computors to play the imitation game" (1952, 148). In addition, concerning Turing's scoring framework, Pinsky said only: "If the interrogator cannot distinguish the machine from the human player in a high percentage of the attempts, presumably machines do think" (1951, 397).

Moreover, there are reasons to think that (P) does *not* state Turing's framework for scoring the game. In his 1952 broadcast, Turing appeared to set the bar high for any machine, saying that in his test a "considerable proportion" of interrogators must be fooled—and even the advocates of the "strict Turing rules" acknowledge that a 30% pass mark is "weak" (Turing et al. 1952, 495; Shah and Warwick 2010, 452). This aside, treating (P) as stating Turing's scoring framework generates a discrepancy between his words in the 1950 paper and those in his 1952 broadcast. In the latter, when Max Newman suggested that "it will be a long time from now" before a machine in Turing's test "is to stand any chance with no questions barred", Turing replied, "Oh yes, at least 100 years, I should say" (Turing et al. 1952, 495). If (P) states Turing's authorized scoring framework, there is a prima facie inconsistency between the "about fifty years" prediction and the "at least 100 years" prediction. To resolve this inconsistency, advocates of the "strict Turing rules" declare that it is "unlikely" that Turing "truly felt" that a machine would pass the test (i.e., would attain a 30% deception rate in five minutes) by the end of the

[16] Later in the 1950 paper, Turing said that "the only really satisfactory support" that can be given for his prediction (P) "will be that provided by waiting for the end of the century and then doing the experiment described" (1950, 454). (We are grateful to an anonymous *Philosophy of Science* reviewer for emphasizing this remark.) The "experiment" here is that referred to in (P)—so, does this additional remark add support to the claim that Turing's rule for scoring his test is 30% success after five minutes? Even if we disregard the reasons against this interpretation (given in the text), the answer is still No: I might say that the only way to know if my student will score 30% or less in the exam is to conduct the exam and see what happens—my saying this would provide no additional reason to think that 30% is the pass mark.

twentieth century (Warwick and Shah 2016b, 39). However, this is to claim, implausibly, that Turing did not mean what he very clearly said.

The second scoring framework proposed as Turing's own (and often used in recent LLM studies) centres on chance. For example, James Moor holds that a computer does well if, after the game is played by "many different interrogators", each with "ample opportunity to ask many questions", the "probability of the average interrogator correctly identifying the machine is not significantly greater than 50 percent" (1976, 249–50). Other accounts of Turing's test require that the computer must do better than chance—that is, the probability of the average interrogator's judging correctly must be less than 50%. However, Turing does not indicate either formulation in the *Mind* paper or in his other comments on his test. An additional problem with several of the proposed chance frameworks is that they also stipulate that the game is to last only five minutes. However, except for (P), Turing did not mention a duration for the imitation game, or even state that the game *has* a set duration, in his 1950 paper or anywhere else.

A third well-known interpretation of Turing's scoring framework is due to Jack Copeland (Copeland 2004a, 2023). In Copeland's view, Turing's remark "Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" indicates "the protocol for scoring the test": a computer passes the test if the interrogator in the computer-imitates-human game does no better than the interrogator in the man-imitates-woman game (Copeland 2004a, 436). Using this framework, Turing's discussion of the man-imitates-woman game does not lack in "lucidity", as Hodges claimed; on the contrary, Turing meant exactly what he said.

Copeland does not address the question of why Turing might have chosen this framework. It might seem odd: although comparing the performances of different test-takers on the same test is common scientific practice, Copeland's interpretation appears to compare the performances of different test-takers (the interrogators) on *different* tests (the computer-imitates-human game and the man-imitates-woman game). This oddity may disappear, though, if both games are regarded simply as two tokens of one type of test, an imitation test—Turing speaks of the machine "being put through one of my imitation tests" (Turing et al. 1952, 503). Different test-takers are then being compared on the same type of test. Certainly, this only invites another question: why select performance in an "imitation test" as a criterion of thinking in machines? This question, however, concerns Turing's philosophical justification for his test, rather than his design, and so arises for any proposed scoring framework.

In sum, of the three prominent scoring frameworks for the computer-imitates-human game proposed as Turing's pick, his own words exclude the first (30% deception in five minutes), neither support nor exclude the second (chance)—so long as it does not incorporate a five-minute limit—and are in line with the third (comparison with the man-imitates-woman game). Juxtaposing Moor's proposed framework with Copeland's, the former appears to have the greater explanatory burden. It has to explain why Turing failed to mention a chance benchmark and also to provide an alternative construal of Turing's words, "Will the interrogator decide wrongly . . . ".

4. Essential features of Turing's design

Additional key aspects of Turing's design are commonly overlooked or misunderstood.

4.1 Controlling for anthropomorphism

Frequently, including in many online "Turing tests", Turing's game is presented as a *two*-player game, where the interrogator must decide whether a single hidden interviewee is a computer or a human being. Yet, as quoted above, in *Mind* Turing stated that "a game which we call the 'imitation game'" is "played with three people". Where one interviewee is a computer, the three-player setup also has theoretical advantages.

Turing himself pointed out a flaw in the two-player game. In his 1952 broadcast, he described what appears to be a two-player imitation game, saying that interrogators should be required to judge "quite a number of times", where "sometimes they really are dealing with a man and not a machine"—this is to "prevent them saying 'It must be a machine' every time without proper consideration" (Turing et al. 1952, 495). Turing's observation is astute: a defect in the two-player game is that judges, to avoid the embarrassment of being fooled by a machine, might always or mostly declare the interviewee to be a computer. Such behaviour has been observed in two-player imitation games held as part of the Loebner Prize Contests in Artificial Intelligence. This flaw does not exist in the three-player imitation game, since it involves a forced choice (Copeland 2000; Moor 2001).

In addition, the three-player game turns an interrogator's natural disposition to anthropomorphize into a controlled variable, thereby preventing it from compromising the test. In his 1948 report, Turing remarked that interacting with even a "paper machine" produces "a definite feeling that one is pitting one's wits against something alive" (1948, 412). This is another astute observation: humans frequently treat chatbots as animate, in particular those designed as "AI companions" with "personality" attributes (Østergaard 2025). Several current researchers posit an "AI psychosis", where users of agentic AI come to believe that these systems are conscious beings (Morrin et al. 2025). The result of humans' disposition to anthropomorphize is that no two-player imitation game can constitute a reliable test: the interrogator might simply be fantasizing. The three-player setup, however, is not vulnerable in the same way: we may presume that the interrogator's inclination to anthropomorphize is present in interactions with the machine and with the human (Proudfoot 2011).

4.2 Mitigating the effect of irrelevant attributes

In *Mind*, Turing said that the computer can "deliberately introduce mistakes in a manner calculated to confuse the interrogator" (1950, 448). The machine is "permitted all sorts of tricks so as to appear more man-like, such as waiting a bit before giving the answer, or making spelling mistakes" (Turing et al. 1952, 495). This aspect of Turing's design—allowing the programmer to hide the machine's perfect spelling or fast response time—is often regarded as a defect: his test, critics claim, merely encourages "artificial stupidity" in the machine.[17] However, incorporating such "tricks" is necessary and proper.

[17] The Economist. "Artificial stupidity." 324(7770). August 1, 1992, 14.

In the late 1940s and early 1950s, newspapers commented, often with photographs, on the "somewhat unprepossessing" appearance of the Manchester machine (the successor to the Manchester "Baby", the world's first electronic stored-program digital computer).[18] This computer, they said, is "a mass of untidy wires, valves, chassis, and display tubes".[19] The Cambridge EDSAC (the second electronic stored-program computer to function) was also described in unflattering terms—as a "vast number of valves set in grey-painted metal racks", which could be taken for "a valve-maker's store or part of a spares room at the B.B.C.".[20] These depictions of the early computers made them seem more like primitive mechanical calculators and less like thinking machines.[21]

It is in the context of such media coverage[22] that Turing remarked, "We do not wish to penalise the machine for its inability to shine in beauty competitions" (1950, 435). This inability, he said, is an "irrelevant" property—that is to say, a machine's "unprepossessing" appearance does not defeat the claim that it can think (1950, 435). It follows that, when testing a machine for intelligence, its appearance, along with all its other irrelevant properties (such as perfect spelling), should be disguised—if, as Turing entreated, "fair play" is to be given to the machine (1947, 394).[23]

4.3 Identifying real understanding

Turing described his imitation game as "something like a viva-voce examination" (1951, 484). Several recent theorists claim that a viva is "just one way" of structuring the game—and also that this format is undesirable, since it puts "too much emphasis" on the computer's ability to "articulate one's reasons for one's actions" (Bohn 2024, 8; Allen et al. 2000, 254). Increasingly, experiments described as Turing tests simply provide an interrogator with (say) transcripts of text previously produced by computers and humans; the interrogator is to identify, based only on the transcripts, which authors are machines (e.g., Aharoni et al. 2024; Dillion et al. 2025; Mei et al. 2024).

Although these tests are logistically easier, losing the opportunity to challenge the machine is a major disadvantage (Proudfoot 2024). This is because, as Turing said, the viva format can be used "to discover whether some one really understands something or has 'learnt it parrot fashion'"—that is, merely by rote memorization (1950, 446). It is well known that an LLM may perform well due only to "spurious correlations"—where, for example, a network with added image processing is being trained to identify images of boats but mistakenly takes a harbour in the background of several

[18] The Times. "The Mechanical Brain: Answer Found to 300-Year-Old Sum." June 11, 1949, 4.

[19] The Times. "The Mechanical Brain: Answer Found to 300-Year-Old Sum." June 11, 1949, 4.

[20] Evening Telegraph. "How the 'Mechanical Brain' Works." October 19, 1949, 5.

[21] For example, Turing's ACE was described as merely "the big brother of the sliderule and the adding machine. . . . there is nothing more 'mental' about it" (Langdon-Davies 1946, 2).

[22] The pioneers in the development of the early computers in the UK, along with journalists and the general public, discussed fundamental philosophical points in newspaper articles, magazines, Letters to the Editor, and on radio.

[23] According to Warwick and Shah, Turing's licence for "tricks" permitted the Eugene Goostman chatbot to identify itself as a 13-year-old Ukrainian boy, in order to explain away its limited English language skills (2016b, 190). However, Turing did not suggest that a programmer can use tricks to disguise a machine's *relevant* inabilities.

images as the salient feature (Hosseini et al. 2025). In this case, even though the model might perform well on some boat image tests, it does not "really understand" what a boat looks like and may go on to mislabel any image with a harbour background as an image of a boat. In Turing's viva, the interrogator can require a machine to explain and justify its behaviour, thereby reducing the chances of its doing well due solely to spurious associations.

The importance of the viva format also has implications for the choice of interrogators—who, Turing said, "should not be expert" about machines (Turing et al. 1952, 495). What might Turing have meant by "expert"? Since he was speaking to the Third Programme's diverse audience, we might presume that that he used "expert" as a vague term—and similarly that, when predicting in *Mind* how an average interrogator might do, he did not intend "average" as a precise statistical measure. Turing sought primarily to exclude interrogators with knowledge of a computer's abilities and architecture, rather than to include a certain sort of interrogator.[24] This exclusion is usually ignored; for example, in recent online imitation games, typically interrogators were selected without controlling for expertise, and in the much-publicized 2014 Royal Society Turing Test (where the Eugene Goostman chatbot was declared to pass the test), several were computer scientists.

Indeed, Turing's restriction may seem a fault in his test, since one might expect computer scientists to be the most perceptive interrogators. However, the situation is more complicated. In the Loebner Prize Contests, for example, knowledgeable interrogators probed specific weaknesses of AI systems by replacing ordinary conversation with gibberish, peculiar "common knowledge" questions, and syntactically tricky sentences quite unlike everyday language. These strategies deny a machine the opportunity to give, as Turing required, a "satisfactory and sustained" viva performance (1950, 447). Moreover, programmers reacted to these strategies by incorporating canned replies to such odd inputs, typically non sequiturs (e.g., Eugene Goostman, when unable to answer a question, frequently asked the interrogator for their address or occupation).[25] Expert interrogators can, then, hinder the practical role of Turing's game in improving AI capabilities.

5. Experimental implementation and results: An informal description

For any thought experiment or actual implementation to be *Turing's* test, we have argued, it should possess the design features established in sections 2–4. Yet one might ask: why should scientists pay attention to Turing's test?[26] Benchmarks (e.g., Guo et al. 2025) that are currently used for "reasoning", "common sense", "natural language processing", "logical thinking", "theory of mind", and suchlike in LLMs

[24] It might be thought that today, given widespread familiarity with LLMs, very many people are experts. In fact, it is unclear how accurate people are in detecting AI-generated content, including text; some studies report that humans who frequently use LLMs for writing tasks excel at identifying AI-generated text, outperforming machine detectors (e.g., Russell et al. 2025), while others conclude that human performance is only at chance level or a little above (e.g., Fiedler and Döpke 2025; Cooke et al. 2025). It may be that we overestimate our ability to identify AI-generated text (over-confidence has been reported in the case of detecting deepfake videos [Köbis et al. 2021]).

[25] Transcripts in Warwick and Shah (2016a).

[26] For the purposes of this question, we assume that the numerous objections in the literature to the Turing test can be answered satisfactorily.

appear to provide criteria of intelligence that are much simpler, logistically and conceptually, than Turing's test. In addition, neuroscientists (e.g., Kazemian et al. 2025) offer the promise of building artificial neural networks with architectural similarities to human brains, and consequently of testing machines for those similarities. Turing's test has an important role as one of the philosophical foundations of AI. Apart from this, however, what reason could a scientist have for running Turing's test?[27]

Turing's philosophy of mind may provide a reason why his test is not redundant, even assuming vastly improved versions of current modes of testing generative AI models. (It seems unlikely that Turing simply did not envisage automated testing.) In his 1948 paper, Turing said that the concept of intelligence is an "emotional concept" and the idea of "intelligence" is "itself emotional rather than mathematical" (1948, 431, 411). On a recent interpretation of this and other remarks, the philosophical justification for his test is (what we would now call) a response-dependence approach to the mind: whether an entity is intelligent or thinks is in part determined by how human beings respond to the entity, including in the conditions of the imitation game (Proudfoot 2013).[28] Neither standard LLM benchmarking nor neuroscience-inspired AI involves responses in the way suggested by Turing's remarks—so, neither could replace his test as a test for "intelligence" or "thinking" in the response-dependence sense.[29]

The other well-known interpretations of Turing's philosophy of mind—behaviourism and (classical or non-classical) computationalism—appear to take intelligence or cognition as response-independent. On those readings, improved automated testing or neuroscience might properly replace the imitation game. Ultimately, then, motivating the implementing of Turing's test might require deciding fundamental questions in the philosophy of mind regarding the concept of intelligence or thinking.

For now, however, we may have good reason to run Turing's test. Perhaps (as suggested in section 4.3) the imitation game can check that LLMs that do well on "reasoning" benchmarks are in fact capable of reasoning (see, e.g., Mondorf and Plank 2024 on LLM failures). Or perhaps the game will prove an efficient means of testing for several cognitive abilities at once. Even the scientist who is sceptical of the test's traditional role might regard Turing's test as a useful stopgap until artificial networks become significantly more brain-like.

5.1 Testing an LLM

In our study, we carried out a computer-imitates-human game (CIHG), with GPT-4-Turbo as the computer. (As noted in the introduction, here we give only a brief

[27] We are indebted to an anonymous Philosophy of Science reviewer for pressing this issue, and also for suggesting that the response-dependence interpretation may have a role.

[28] This approach is a theoretical claim about the concept of thinking or intelligence, namely that it is dependent on the concept of our responses in certain conditions. This claim is distinct from any empirical hypothesis concerning a correlation between perceived humanness and perceived intelligence.

[29] Turing said that "[t]he extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration." (1948, 431). This may allow a role for the Turing test *and* current LLM testing, the former to investigate response-dependent properties and the latter response-independent properties of the machine.

informal report of our study; full technical details are available in Temtsin et al. 2025.) Our CIHG had three players, and the interrogator could initiate conversation on any topic. The game was open-ended; the length of each trial was determined by the interrogator, who continued the interaction until confident of identifying the computer. We excluded interrogators and interviewees who might reasonably be expected to have some AI (or philosophy of AI) expertise.[30]

In his computer-imitates-human game, Turing said, the machine can "brag" about its "charms, strength or heroism"—just as the (mid-twentieth century English) man in the man-imitates-woman game can say falsely that his hair "is shingled, and the longest strands are about nine inches long" (1950, 435, 433). Analogously, GPT-4-Turbo was prompted to behave as an English-speaking university student, using colloquial language and idioms, and to deny that it was a machine. It was also prompted to delay its responses.[31]

In addition, we implemented Turing's man-imitates-woman game (MIWG). Doing so enabled us to score the CIHG using each of the two frameworks (discussed in section 3) that are not ruled out by Turing's own words. Our MIWG was another three-player game, on any conversational topic, with the duration determined in the same way by the interrogator. For consistency with the CIHG, we excluded participants from the MIWG who could reasonably be expected to have knowledge of gender differences in interaction or communication style.

All other aspects of the study design conformed to guidelines for good practice.

To the best of our knowledge, no previous empirical study conforms to Turing's conceptual design. For example, although the latest study claiming success in the Turing test (a Prolific replication study reported in Jones and Bergen 2026) implemented a three-player imitation game and increased the time limit to 15 minutes, it did not exclude all participants who might be presumed to be aware of AI debates or the test,[32] and it retained a set duration for the game. In addition, this study did not implement a man-imitates-woman game (in the study, a machine passed the test if interrogators did not do significantly better than chance).

5.2 The duration issue

In our study, 33 out of 37 interrogators—89%—in the CIHG took longer than five minutes to make a decision. The average (mean) duration of CIHG trials was 14 minutes, and the longest trial lasted 46 minutes. Typically the longer a task takes, the greater the cognitive resources required for that task (Barrouillet et al. 2007).[33]

[30] In addition to excluding computer science students and faculty, we excluded philosophy students and faculty, since it was likely that they would be aware of debates about AI, distinctive features of machines, or the Turing test itself.

[31] The technical report of our study includes our prompt in full, so other experimenters may try to replicate our results or try out different prompts.

[32] The Prolific study excluded participants who reported familiarity with the experiment, but did not exclude participants due simply to education or profession (Appendix 01 to Jones and Bergen 2026, https://www.pnas.org/doi/10.1073/pnas.2524472123#supplementary-materials).

[33] Thanks to David Kaber for directing us to this study.

It would appear, then, that most CIHG interrogators need, or believe that they need, considerably longer than five minutes in order to come to a decision.[34]

This result supports individual interrogators' critical comments regarding earlier "Turing tests". For example, Barry Cooper said of the 2014 Royal Society Turing test that he felt "confusion skewed the outcome", remarking that "of course, the 5 minutes was for 2 conversations, so one had little more than 2 minutes with the machine" (Cooper, personal communication with Proudfoot, 2014). Similarly, Dan Falk said about the 2012 Bletchley Park Turing Test: "five minutes goes by pretty fast. And the shorter the conversation, the greater the computer's advantage" (Falk 2014).[35]

Even setting aside Turing's intention, then, our study suggests that the practice of setting a five-minute (and even a 15-minute) limit to his imitation game may hamper interrogators, potentially invalidating the results of any game.

5.3 The machine's performance (poor)

In the CIHG, 36 out of 37 interrogators—97%—correctly identified the computer. So, using Moor's scoring framework (namely, that the average interrogator should not do significantly better than chance), GPT-4-Turbo did not pass the test.[36] In addition, 16 out of 37 interrogators in the MIWG—43%—correctly identified the man. Consequently, the performance of the average CIHG interrogator was very much better than the performance of the average MIWG interrogator; so, using Copeland's scoring framework (namely, that the CIHG interrogator not do better than the MIWG interrogator), again GPT-4-Turbo did not pass the test.

Of course, like some recent Turing test studies (Rahimov et al. 2025; Restrepo Echavarría 2025), we tested only one model. Also like some studies (Jones and Bergen 2024; Rahimov et al. 2025;[37] Restrepo Echavarría 2025), we used only one prompt (per model). Different models or prompts might produce a better result for the machine. New generative AI models are released frequently (at the date of writing, the latest available GPT model is the GPT-5.6 family). What might the results of our study have been, had we used GPT-5.6 Sol, or the most recent Claude or Gemini iteration—let alone an LLM under development now (such as Open AI's Astra) or in the future? It is impossible to say: the literature on improvements in LLMs is complicated, and it appears difficult systematically to compare the performance of different models (Afridi et al. 2026). How well either a different LLM already heralded as a Turing-test

[34] The four CIHG interrogators who decided within five minutes all accurately identified the computer (see section 5.3)—it might be objected, therefore, that in fact our study demonstrates that five minutes is sufficient. This objection would fail, however, since there is an alternative explanation of why these interrogators did no worse than those taking longer: namely, GPT-4-Turbo's extremely poor performance overshadowed any effect arising from longer interactions.

[35] The "trick" of slowing the computer's responses may increase this advantage. For example, the Julia chatbot's responses were slowed to match a human's typing speed, with (the experimenter said) the "side benefit" that the interrogator would have less time to ask questions, giving Julia "fewer chances to make a fatal mistake" (Mauldin 2009, 425). We presume that this advantage disappears in an imitation game with no set duration.

[36] The result is the same if the interrogator must perform worse than chance.

[37] Rahimov et al. used a single prompt and compared the model's performance with and without prompt.

passer or a model later than GPT-4-Turbo would do in our CIHG is an entirely open question, to be answered experimentally.

GPT-4-Turbo's poor performance does not indicate any inherent difficulty in an LLM's passing Turing's test. Several recent discussions claim that no LLM genuinely cognizes—cognition is said to require, for example, consciousness (e.g., Hadid et al. 2026), sensory grounding or embodiment (e.g., Bender and Koller 2020; Gahrn-Andersen 2025), or a hybrid architecture incorporating symbolic AI (e.g., Marcus and Davis 2019). We take no position here with respect to these larger debates. Rather, we provide experimenters with a modest challenge: to earn the cachet of passing *Turing's* test, their LLM must do well in the CIHG.

5.4 The gender test revisited

The average (mean) duration of the MIWG trials was 24 minutes—ten minutes longer than the average for CIHG interrogators. The longest trial lasted 49 minutes.

Several scholars claim that, irrespective of Turing's intention, a computer-imitates-woman test is "vastly superior" to a computer-imitates-human test (Sterrett 2000, 542; see also Sterrett 2020). The former, it is said, is "more stringent" than the latter, which is "a weaker game for the computer to play" (Traiger 2000, 570). A reason given for the superiority of the gender test is that female impersonation is (it is claimed) especially challenging for the impersonator, *whether a man or a machine*: it demands "more of what is relevant to thinking", including "resourcefulness in dealing with unfamiliar tasks" and "a critical attitude to one's habitual responses" (Sterrett 2000, 547, 542, 541).

Our study offers no empirical support for these theoretical claims. As noted above, longer response times typically indicate a more demanding task. So, with respect to the demand our female impersonation test (the MIWG) placed upon a player's cognitive resources, our study points to a demand only on the interrogator (as compared with the CIHG interrogator), and not on the impersonator. In addition, if a female impersonation test were especially demanding, then, all things being equal, we might expect more MIWG interrogators than CIHG interrogators to make correct identifications. In our study, in contrast, CIHG interrogators were markedly more successful.

6. Conclusion

Numerous accounts profess to analyse or implement Turing's test. In this paper we aimed to put to bed some very well-trodden interpretative debates and to provide a conceptual blueprint for future implementations. Only a test that instantiates the design identified here, we have argued, can claim to be *Turing's* test.

A fortiori, only an unintelligent (unthinking) machine that passes such a test provides a counterexample to Turing's criterion for thinking. In our view, no counterexample to the test has been demonstrated. Of the numerous proffered counterexamples, many are imaginary machines; for example, Ned Block famously claimed that his hypothetical gigantic lookup table machine, possessing the intelligence merely of a "toaster", could pass the test (1981, 21). Turing made clear, however, that his proposal concerned real-world machines performing in real time—with the result that merely logically possible machines are not counterexamples

(Turing et al. 1952, 503; Copeland 2000; Proudfoot 2013). As to actual machines, recent generative AI models have been said finally to "sail through" the test (Gibney 2025);[38] and many theorists state that these machines "do not think"—they have "great memories" but produce only "inane fabrications" (Johnson-Laird and Ragni 2023; Berghel 2025, 141). However, to the best of our knowledge, no LLM study conforms to Turing's design, and so none provides a counterexample. In sum, studies to date provide no empirical evidence that a new test of intelligence in machines is, as is increasingly often claimed, "urgently required" (Chaudhri 2025).

As to our own study, the results of implementing for the first time Turing's test as he proposed it are consistent with his 1952 prediction that it would be at least a century before a machine passed the test. Nevertheless, our study implemented Turing's conceptual blueprint using a specific LLM, prompt, and participants. Like all empirical studies, this design must be repeated with different models, different prompts, different subjects, and other relevant variations.

Acknowledgements. We would like to thank three anonymous *Philosophy of Science* reviewers, whose suggestions substantially improved the paper.

Funding statement. None to declare.

Declaration. None to declare.

References

Afridi, Hina, Habib Ullah, Sultan Daud Khan, and Mohib Ullah. 2026. "From GPT-3 to GPT-5: Mapping their Capabilities, Scope, Limitations, and Consequences." Preprint, arXiv, April 11, 2026. https://doi.org/10.48550/ARXIV.2604.10332.

Aharoni, Eyal, Sharlene Fernandes, Daniel J. Brady, Caelan Alexander, Michael Criner, Kara Queen, Javier Rando, Eddy Nahmias, and Victor Crespo. 2024. "Attributions toward Artificial Agents in a Modified Moral Turing Test." *Scientific Reports* 14 (1):8458. https://doi.org/10.1038/s41598-024-58087-7.

Allen, Colin, Gary Varner, and Jason Zinser. 2000. "Prolegomena to any Future Artificial Moral Agent." *Journal of Experimental & Theoretical Artificial Intelligence* 12 (3):251–61. https://doi.org/10.1080/09528130050111428.

Barrouillet, Pierre, Sophie Bernardin, Sophie Portrat, Evie Vergauwe, and Valérie Camos. 2007. "Time and Cognitive Load in Working Memory." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33 (3):570–85. https://doi.org/10.1037/0278-7393.33.3.570.

Bender, Emily M., and Alexander Koller. 2020. "Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 5185–98. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463.

Berghel, Hal. 2025. "A Generative AI Perspective on the Turing Test: Passed but not Vindicated." *Computer* 58 (11):133–42. https://doi.org/10.1109/MC.2025.3600889.

Block, Ned. 1981. "Psychologism and Behaviorism." *The Philosophical Review* 90 (1):5–43. https://doi.org/10.2307/2184371.

Bohn, Einar Duenger. 2024. "The Moral Turing Test: A Defense." *Philosophy & Technology* 37:111. https://doi.org/10.1007/s13347-024-00793-1.

Chaudhri, Vinay K. 2025. "We Need a New Turing Test to Assess Al's Real-World Knowledge." *Nature* 646:1027. https://doi.org/10.1038/d41586-025-03471-0.

Cohen, Henry. 1952. "The Status of Brain in the Concept of Mind." *Philosophy* 27 (102):195–210. https://doi.org/10.1017/S0031819100033945.

[38] This follows many earlier alleged actual counterexamples (e.g., Deep Blue, Watson, AlphaGo).

Cooke, Di, Abigail Edwards, Sophie Barkoff, and Kathryn Kelly. 2025. "As Good as a Coin Toss: Human Detection of AI-Generated Content." *Communications of the ACM* 68 (10):100–109. https://doi.org/10.1145/3729417.

Copeland, B. Jack. 2000. "The Turing Test." *Minds and Machines* 10:519–39. https://doi.org/10.1023/A:1011285919106.

Copeland, B. Jack. 2004a. "Introduction to 'Computing Machinery and Intelligence'." In Copeland ed. 2004, 433–40.

Copeland, B. Jack. 2004b. "Introduction to 'Lecture on the Automatic Computing Engine'." In Copeland ed. 2004, 362–77.

Copeland, B. Jack. ed. 2004. *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, Plus the Secrets of Enigma*. Oxford: Clarendon Press. https://doi.org/10.1093/oso/9780198250791.001.0001.

Copeland, B. Jack. 2023. "Early AI in Britain: Turing et al." *IEEE Annals of the History of Computing* 45 (3):19–31. https://doi.org/10.1109/MAHC.2023.3300660.

Curzan, Anne. 2003. *Gender Shifts in the History of English*. Cambridge: Cambridge University Press. https://doi.org/10.1017/S0047404505290115.

Dennett, Daniel C. 1985. "Can Machines Think?" In *How We Know: Nobel Conference XX*, edited by Michael G. Shafto, 121–45. San Francisco: Harper & Row.

Dillion, Danica, Debanjan Mondal, Niket Tandon, and Kurt Gray. 2025. "AI Language Model Rivals Expert Ethicist in Perceived Moral Expertise." *Scientific Reports* 15:4084. https://doi.org/10.1038/s41598-025-86510-0.

Earp, Brian D. 2012. "The Extinction of Masculine Generics." *Journal for Communication and Culture* 2 (1):4–19.

Falk, Dan. 2014. "The Turing Test Measures Something, But It's Not 'Intelligence'." *Smithsonian Magazine*, June 10, 2014.

Fiedler, Alexandra, and Jörg Döpke. 2025. "Do Humans Identify AI-Generated Text Better Than Machines? Evidence Based on Excerpts from German Theses." *International Review of Economics Education* 49:100321. https://doi.org/10.1016/j.iree.2025.100321.

Gahrn-Andersen, Rasmus. 2025. "Beyond Symbol Processing: The Embodied Limits of LLMs and the Gap Between AI and Human Cognition." *AI & Society* 40:3105–7. https://doi.org/10.1007/s00146-025-02382-y.

Genova, Judith. 1994. "Turing's Sexual Guessing Game." *Social Epistemology* 8 (4):313–26. https://doi.org/10.1080/02691729408578758.

George, Frank H. 1956. "Could Machines Be Made to Think?" *Philosophy* 31 (118):244–52. https://doi.org/10.1017/S0031819100057284.

George, Frank H. 1958. "Finite Automata." *Philosophy* 33 (124):57–59. https://doi.org/10.1017/S0031819100038006.

Gibney, Elizabeth. 2025. "AI Language Models Killed the Turing Test: Do We Even Need a Replacement?" *Nature*, October 20, 2025. https://doi.org/10.1038/d41586-025-03386-W.

Gonçalves, Bernardo. 2024. "Turing's Test, a Beautiful Thought Experiment." *IEEE Annals of the History of Computing* 46:36–49. https://doi.org/10.1109/MAHC.2024.3432278.

Gunderson, Keith. 1964. "The Imitation Game." *Mind* 73 (290):234–45. https://doi.org/10.1093/mind/LXXIII.290.234.

Guo, Meng-Hao, Jiajun Xu, Yi Zhang, Jiaxi Song, Haoyang Peng, Yi-Xuan Deng, Xinzhi Dong, Kiyohiro Nakayama, Zhengyang Geng, Chen Wang, Bolin Ni, Guo-Wei Yang, Yongming Rao, Houwen Peng, Han Hu, Gordon Wetzstein, and Shi-min Hu. 2025. "R-Bench: Graduate-Level Multi-Disciplinary Benchmarks for LLM & MLLM Complex Reasoning Evaluation." Preprint, arXiv, May 4, 2025. https://doi.org/10.48550/ARXIV.2505.02018.

Hadid, Vanessa, Karim Jerbi, and John W. Krakauer. 2026. "The Illusion of AI Consciousness: Lessons from Human Unconscious Processing." *The Transmitter*, June 8, 2026. https://doi.org/10.53053/VAGJ4451.

Hayes, Patrick, and Kenneth Ford. 1995. "Turing Test Considered Harmful." In *IJCAI'95 Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, vol. 1, edited by Chris S. Mellish, 972–77. San Francisco: Morgan Kaufmann.

Hodges, Andrew. 2014. *Alan Turing: The Enigma, The Book That Inspired the Film The Imitation Game—Updated Edition*. Princeton, NJ: Princeton University Press.

Hosseini, Parsa, Sumit Nawathe, Mazda Moayeri, Sriram Balasubramanian, and Soheil Feizi. 2025. "SpurLens: Automatic Detection of Spurious Cues in Multimodal LLMs." Preprint, arXiv, March 11, 2025. https://doi.org/10.48550/arXiv.2503.08884.

Jefferson, Geoffrey. 1949. "The Mind of Mechanical Man." *The British Medical Journal* 1 (4616):1105–10. https://doi.org/10.1136/bmj.1.4616.1105.

Johnson-Laird, Philip N., and Marco Ragni. 2023. "What Should Replace the Turing Test?" *Intelligent Computing* 2:0064. https://doi.org/10.34133/icomputing.0064.

Jones, Cameron R., and Benjamin K. Bergen. 2024. "People Cannot Distinguish GPT-4 from a Human in a Turing Test." Preprint, arXiv, May 9, 2024. https://doi.org/10.48550/arXiv.2405.08007.

Jones, Cameron R., and Benjamin K. Bergen. 2025. "Large Language Models Pass the Turing Test." Preprint, arXiv, March 31, 2025. https://doi.org/10.48550/arXiv.2503.23674.

Jones, Cameron R., and Benjamin K. Bergen. 2026. "Large Language Models Pass a Standard Three-Party Turing Test." *Proceedings of the National Academy of Sciences* 123 (21):e2524472123. https://doi.org/10.1073/pnas.2524472123.

Kazemian, Atlas, Eric Elmoznino, and Michael F. Bonner. 2025. "Convolutional Architectures Are Cortex-Aligned De Novo." *Nature Machine Intelligence* 7:1834–44. https://doi.org/10.1038/s42256-025-01142-3.

Kind, Amy. 2022. "Computing Machinery and Sexual Difference: The Sexed Presuppositions Underlying the Turing Test." In *Feminist Philosophy of Mind*, edited by Keya Maitra and Jennifer McWeeny, 54–69. New York: Oxford University Press. https://doi.org/10.1093/oso/9780190867614.003.0003.

Köbis, Nils C., Barbora Doležalová, and Ivan Soraperra. 2021. "Fooled Twice: People Cannot Detect Deepfakes but Think They Can." *iScience* 24 (11):103364. https://doi.org/10.1016/j.isci.2021.103364.

Langdon-Davies, John. 1946. "Electronic Brains Can't Reason." *Daily Mail*, November 19, 1946, 2.

MacKay, Donald M. 1951. "Mindlike Behaviour in Artefacts." *The British Journal for the Philosophy of Science* 2 (6):105–21. https://doi.org/10.1093/bjps/II.6.105.

Marcus, Gary, and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon Books.

Mauldin, Michael L. 2009. "Going Under Cover: Passing as Human." In *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, edited by Robert Epstein, Gary Roberts, and Grace Beber, 413–29. Dordrecht: Springer. https://doi.org/10.1007/978-1-4020-6710-5_24.

Mays, Wolfe. 1952. "Can Machines Think?" *Philosophy* 27 (101):148–62. https://doi.org/10.1017/S003181910002266X.

Mei, Qiaozhu, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2024. "A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans." *Proceedings of the National Academy of Sciences* 121 (9): e2313925121. https://doi.org/10.1073/pnas.2313925121.

Minsky, Marvin. 2013. "The Turing Test Is a Joke!" Interviewed by Nikola Danaylov on *Singularity Weblog*, July 12, 2013. https://www.singularityweblog.com/marvin-minsky/.

Mitchell, Melanie. 2024. "The Turing Test and our Shifting Conceptions of Intelligence." *Science* 385 (6710):eadq9356. https://doi.org/10.1126/science.adq9356.

Mondorf, Philipp, and Barbara Plank. 2024. "Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models—A Survey." *Conference on Language Modeling* [COLM 2024], University of Pennsylvania, 7–9 October, 2024. https://colmweb.org/2024/AcceptedPapers.html.

Moor, James H. 1976. "An Analysis of the Turing Test." *Philosophical Studies* 30:249–57. https://doi.org/10.1007/BF00372497.

Moor, James H. 2001. "The Status and Future of the Turing Test." *Minds and Machines* 11:77–93. https://doi.org/10.1023/A:1011218925467.

Morrin, Hamilton, Luke Nicholls, Michael Levin, Jenny Yiend, Udita Iyengar, Francesca DelGuidice, Sagnik Bhattacharyya, James MacCabe, Stefania Tognin, Ricardo Twumasi, Ben Alderson-Day, and Thomas Pollak. 2025. "Delusions by Design? How Everyday AIs Might Be Fuelling Psychosis (and What Can Be Done About It)." Preprint, PsyArXiv, July 12, 2025. https://doi.org/10.31234/osf.io/cmy7n_v5.

Mountbatten, Louis. 1946. "Presidential Address of Admiral the Viscount Mountbatten of Burma, K.G., G.C.V.O., K.C.B., D.S.O." *Journal of the British Institution of Radio Engineers* 6 (6):221–25.

Østergaard, Søren Dinesen. 2025. "Generative Artificial Intelligence Chatbots and Delusions: From Guesswork to Emerging Cases." *Acta Psychiatrica Scandinavica* 152 (4):257–59. https://doi.org/10.1111/acps.70022.

Pinsky, Leonard. 1951. "Do Machines Think About Machines Thinking?" *Mind* 60 (239):397–98. https://doi.org/10.1093/mind/lx.239.397.

Proudfoot, Diane. 2011. "Anthropomorphism and AI: Turing's Much Misunderstood Imitation Game." *Artificial Intelligence* 175 (5–6):950–57. https://doi.org/10.1016/j.artint.2011.01.006.

Proudfoot, Diane. 2013. "Rethinking Turing's Test." *The Journal of Philosophy* 110 (7):391–411. https://doi.org/10.5840/jphil2013110722.

Proudfoot, Diane. 2024. "Turing's Test vs the Moral Turing Test." *Philosophy & Technology* 37:134. https://doi.org/10.1007/s13347-024-00825-w.

Rahimov, Avraham, Orel Zamler, and Amos Azaria. 2025. "The Turing Test Is More Relevant Than Ever." Preprint, arXiv, May 5, 2025. https://doi.org/10.48550/arXiv.2505.02558.

Restrepo Echavarría, Ricardo. 2025. "ChatGPT-4 in the Turing Test." *Minds and Machines* 35:8. https://doi.org/10.1007/s11023-025-09711-6.

Russell, Jenna, Marzena Karpinska, and Mohit Iyyer. 2025. "People who Frequently Use ChatGPT for Writing Tasks Are Accurate and Robust Detectors of AI-Generated Text." In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, vol. 1, edited by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, 5342–73. https://doi.org/10.18653/v1/2025.acl-long.267.

Scriven, Michael. 1953. "The Mechanical Concept of Mind." *Mind* 62 (246):230–40. https://doi.org/10.1093/mind/LXII.246.230.

Shah, Huma, and Kevin Warwick. 2010. "Testing Turing's Five Minutes, Parallel-Paired Imitation Game." *Kybernetes* 39 (3):449–65. https://doi.org/10.1108/03684921011036178.

Sloman, Aaron. 2013. "Aaron Sloman Absolves Turing of—the Mythical Turing Test." In *Alan Turing: His Work and Impact*, edited by S. Barry Cooper and Jan van Leeuwen, 606–11. Waltham, MA: Elsevier.

Sterrett, Susan G. 2000. "Turing's Two Tests for Intelligence." *Minds and Machines* 10:541–59. https://doi.org/10.1023/A:1011242120015.

Sterrett, Susan G. 2020. "The Genius of the 'Original Imitation Game' Test." *Minds and Machines* 30:469–86. https://doi.org/10.1007/s11023-020-09543-6.

Temtsin, Sharon, Diane Proudfoot, David Kaber, and Christoph Bartneck. 2025. "The Imitation Game According to Turing." Preprint, arXiv, January 29, 2025. https://doi.org/10.48550/arXiv.2501.17629.

Traiger, Saul. 2000. "Making the Right Identification in the Turing Test." *Minds and Machines* 10:561–72. https://doi.org/10.1023/a:1011254505902.

Turing, Alan M. 1947. "Lecture on the Automatic Computing Engine." In Copeland ed. 2004, 362–94.

Turing, Alan M. 1948. "Intelligent Machinery." In Copeland ed. 2004, 410–32.

Turing, Alan M. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236):433–60. https://doi.org/10.1093/mind/LIX.236.433.

Turing, Alan M. 1951. "Can Digital Computers Think?" In Copeland ed. 2004, 482–86.

Turing, Alan M., Richard Braithwaite, Geoffrey Jefferson, and Max Newman. 1952. "Can Automatic Calculating Machines Be Said to Think?" In Copeland ed. 2004, 494–506.

Warwick, Kevin, and Huma Shah. 2016a. "Can Machines Think? A Report on Turing Test Experiments at the Royal Society." *Journal of Experimental & Theoretical Artificial Intelligence* 28 (6):989–1007. https://doi.org/10.1080/0952813X.2015.1055826.

Warwick, Kevin, and Huma Shah. 2016b. *Turing's Imitation Game: Conversations with the Unknown*. Cambridge: Cambridge University Press.

Cite this article: Proudfoot, Diane, Sharon Temtsin, and Christoph Bartneck. 2026. "Why No One Has Implemented the Turing Test: Identifying Turing's Design and Putting it into Practice." *Philosophy of Science*. https://doi.org/10.1017/psa.2026.10282