AI & SOCIETY
https://doi.org/10.1007/s00146-024-02172-y

OPEN FORUM

Intelligence is not deception: from the Turing test to community-based ascriptions

Markus Pantsar[1]

Received: 29 April 2024 / Accepted: 16 December 2024

Abstract

The Turing test has a peculiar status in the artificial intelligence (AI) research community. On the one hand, it is presented as an important topic in virtually every AI textbook, and the research direction focused on developing AI systems that behave in human-like fashion is standardly called the "Turing test approach". On the other hand, reports of computer programs passing the Turing test have had relatively little effect. Does this mean that the Turing test is no longer relevant as a test, doomed to be a theoretical notion with little connection to AI practice? In this paper, I argue that there is one problem in particular with common traditional versions of the Turing test, namely their focus on deception. The criterion for passing the Turing test is standardly connected to an AI system's ability to deceive the interrogator about its identity. But why should we connect intelligence to the ability deceive? Here I present a revised version of an intelligence test that is not based on deception. In what I call the Community-based intelligence test (CBIT), an AI is introduced to a community of human subjects. If after a sufficient number of interactions within that community the humans are not able to identify the AI system as a computer, it is considered to have passed CBIT. I discuss whether that should be enough to ascribe intelligence to the AI, and if not, what more would be needed?

Keywords Artificial intelligence · Turing test · Intelligent machines · Philosophy of AI · Mathematical intelligence

✉ Markus Pantsar
markus.pantsar@gmail.com;
markus.pantsar@humtec.rwth-aachen.de

[1] RWTH Aachen/University of Helsinki, Human Technology Institute, Chair for Theory of Science and Technology, Theaterplatz 14, 52062 Aachen, Germany

Published online: 13 January 2025

1 Introduction

The Turing test has become a cornerstone of the literature on artificial intelligence (AI), up to the point that it is discussed in virtually every textbook and popular presentation.[1] However, it enjoys a rather curious status within the AI community. On the one hand, it is undoubtedly among the most widely known and influential ideas in the entire discipline. Indeed, it has been so influential that the "Turing test approach" has come to refer to an entire research direction concerning AI. As used in standard textbooks, such as (Norvig and Russell 2021, Sect. 1.1), approaches to AI research are divided into two dimensions. The first dimension concerns the question of whether intelligence should be connected to humans or to a wider notion of rationality. The second dimension concerns the question of whether intelligence should be understood as a property of behaviour or of internal processes. From the resulting four combinations, one approach thus takes the aim of AI research as creating machines that display human-like behaviour. This is called the Turing test approach, which is a good indicator of the theoretical importance of the test.

On the other hand, in actual AI research practice, the Turing test has been approached in very different ways. There have been reports of AI applications passing the Turing test, most famously in 2014 by the chatbot Eugene Goostman (Warwick and Shah 2016). Given the ubiquity of the Turing test in the AI literature, one would have expected this to be a remarkable event. Instead, the community mostly seemed to be in silent agreement that it was not a particularly important development. This ambiguous attitude toward the Turing test *as a test* prompts the question of how its status should be understood. Perhaps we should simply stop giving the Turing test so much attention, given that its relevance in AI research seems to be increasingly limited to historical treatments.

To some degree, I agree with this critical attitude toward the Turing test. Observing its role in modern discussions on AI, it seems unlikely that the Turing test will become a generally accepted method of ascribing intelligence to machines. Nevertheless, I believe that the Turing Test *approach* can be valuable both for AI research and its philosophy. The problem is not in detecting intelligence based on behaviour. Rather, it is in what kind of behaviour is considered intelligent. In Turing's original experimental setting of the "Imitation Game" (Turing 1950), a human interrogator presents questions to two players, one a machine and the other a human. Based only on the responses (transmitted in a way that does not reveal the players) the interrogator is then expected to identify the machine. If they cannot do that reliably, the argument goes, the machine should be considered intelligent in a similar way to how humans are.[2]

In this paper, I argue that the problematic part about the Turing test is its focus on *deception* as the relevant characteristic of intelligence. For the machine to pass the test, it needs to impersonate a human successfully enough to fool the interrogator. But this is puzzling in the wider context of intelligence ascriptions. Why would intelligence be connected to a form of deception?[3]

Here I present another approach to making intelligence ascriptions, one that is not fundamentally based on deception in the same way. The approach is community-based, so the AI system is placed within a community of human actors.[4] Instead of a single person (or a jury), the "interrogator" in this new test is formed by the community. If this community cannot identify the AI system as a machine, then it has passed what I call the *Community-based intelligence test* (CBIT). In addition to moving from individual interrogators to communities, CBIT is different from the Turing test also in another important way. Namely, it replaces the test setting in which the interrogator *knows* that they are being tested with a more natural setting in which the AI is introduced into existing human communities.

I argue that these two crucial differences make CBIT more relevant for making intelligence ascriptions than the Turing test, especially under its standard interpretations.[5] I also argue that for maximal relevance, in the setting of CBIT, the AI should not be developed with passing the test in mind. As we will see, this is in line with Turing's idea that machines should not be trained for the express purpose of passing the Turing test. In addition to that, however, this

[1] Throughout this paper, I discuss the Turing test in terms of machines, computers and artificial intelligence. I take all three to be in the present context equivalent. In modern usage, computer software are often called AI applications even when we agree that they are not intelligent. This results in questions of the type "is an artificial intelligence intelligent?", which admittedly are problematic. However, I will follow the modern usage of the term here. I justify the equivocation between machines and computers by assuming that any potentially intelligent machine is likely to be (essentially) a computer.

[2] Turing originally writes about thinking, but in modern literature the test is usually discussed in terms of intelligence. Here I follow the likes of (Proudfoot 2013) and (Wheeler 2020) in interpreting that the test can be discussed in terms of both.

[3] It should be noted that Turing most likely did not think of deception as a primary characteristic of intelligent behaviour. Rather, the ability to impersonate – and thus deceive – is central to the Turing test because it fits the empirical set up of his Imitation Game. Indeed, as we will see, he argued against training machines with the express purpose of passing the Turing test. Nevertheless, both in Turing's original set-up and in how the Turing test has consequently been applied, deception has remained as a central characteristic.

[4] By "community", I simply mean a bounded group of humans that interact regularly on specific platforms.

[5] There are also interpretations of the Turing Test that go into the kind of direction that I'm suggesting, such as the"Total Turing Test" (Harnad 1991), the "Lovelace test" (Bringsjord et al. 2001) and the "Questioning Turing Test" (Damassino 2020). These will be discussed in Sect. 4.

means taking focus away from deception as the key to passing the test. Hence, CBIT works best when computers are developed with the aim of *possessing* human-like intelligence, not *impersonating* human intelligence.

Thousands of pages have been written on the Turing test, but I limit the analysis to the literature needed to evaluate CBIT. Consequently, many aspects of Turing literature will not be dealt with. Most importantly, I will not propose an interpretation of what Turing himself ultimately meant by his test.[6] There is anecdotal evidence, for example, that he may have not been entirely serious in his 1950 paper, at least not in all parts.[7] This may be the case, but here I will take the Turing test seriously, and I take it seriously as an actual test to be conducted. This goes against the interpretation of Gonçalves (2023), for example, who has argued that the Turing test should be understood as a thought experiment rather than an actual test.[8]

While for Turing scholars these matters are important, they may not be the most relevant ones for research related to modern AI applications. Especially given the huge development of computers and AI in the more than 70 years since the publication of Turing's paper, there is no guarantee that Turing's own views represent the most fruitful understanding of his test in the modern context. The Turing test has evolved, for better or worse, as a concept since Turing proposed it. For this reason, I will focus on how the Turing test has been understood in modern AI research and the philosophy of AI, and how it should be understood and developed to be maximally relevant for future AI research.

Ultimately, I see Turing's greatest contribution with his test to be the way it replaced vague questions like "can machines think?" with an actual test that could be conducted in practice. While the details have prompted extensive discussion, the test indubitably gave a platform for fighting against an important bias in the question of machine intelligence. This bias is called by many names, but my preferred one is "meat chauvinism", which refers to the dogmatic stance that intelligence is only possible in biological systems (see, e.g., Clark 2008).[9] This bias still plays an important role in philosophy (as seen, for example, in the enduring influence of Searle's (1980) Chinese room argument) but the Turing test at the very least gives us tools to challenge meat chauvinists.

However, as stated above, the actual Turing tests performed so far have not had much impact in that sense. I believe that this is largely due to the way the Turing test is set up with a single interrogator and focus on deception. Since the Community-based intelligence test I present in this paper gets rid of both characteristics, hopefully it can become more relevant for the question of intelligence ascriptions regarding AI systems.

A key part of this approach is to not impose strict conceptual limits on what intelligence is. Almost a century ago, the psychologist Spearman lamented that "… 'intelligence' has become a mere vocal sound, a word with so many meanings that finally it has none" (Spearman 1927, p. 14). While much has happened since then, the notion of intelligence in psychology has remained elusive, with several mutually exclusive definitions proposed (see, e.g., Cianciolo and Sternberg 2004). Consistently with this understanding, the computer scientist Minsky has characterised intelligence as a "suitcase word" with several meanings to unpack (Minsky 2006). This is also my approach in this paper. I do not see intelligence as a meaningless concept, but I do advocate a pluralist reading of intelligence, in which experts in a field of activity are trusted to be the best evaluators of intelligence in that particular field. In the spirit of Turing, I also accept here that such intelligence can be evaluated on a behavioural basis.

In Sect. 2, I present different ways in which Turing's proposed test has been interpreted in the literature, focusing on Sterrett's (2000) distinction between "Original Imitation Game" and "Standard Turing Test". In Sect. 3, I discuss the Turing test as a means of making intelligence ascriptions. I distinguish between two ways in which such ascriptions can go wrong, false positives and false negatives. I show that Turing's main motivation in the historical context was to deal with false negatives, i.e., cases in which a machine is intelligent but we fail to recognise that. However, I argue that an adequate interpretation of the Turing test should also be able to avoid false positives, i.e., ascriptions of intelligence when none is present. Then in Sect. 4, I analyse the status of the Turing test and its proposed revisions in modern AI discussions, concluding that it has been increasingly losing its relevance as an actual test. To change this, I argue in Sect. 5, we need a test to fit the modern context of AI research. For this purpose, I propose the Community-based intelligence test. In Sect. 6, I then discuss how the proposed test should be understood in making intelligence ascriptions.

[6] For a recent book-length analysis of that topic, see (Gonçalves 2024).

[7] This is based on Turing's doctoral student Robin Gandy remarking that Turing "smiled and giggled" while reading parts of the paper (Gandy 1996, p. 125). This anecdote has been mentioned many times, including by Copeland (Turing & Copeland 2004, p. 433), Boden (2006, p. 1351) and Gonçalves (2022, p. 1).

[8] See Copeland (2000) for a discussion on this topic.

[9] Another variation of this type of counter argument has been presented by Chomsky (2023) who has argued that AI applications based on large language models (ChatGPT in particular) cannot be intelligent because they function on a statistical basis, and not symbolically like humans (according to Chomsky) do in their linguistic activity. Following this kind of argument, any artificial intelligence would need to fundamentally mirror human cognition.

2 The Imitation game and the Turing test

The Turing test is often described in terms of a machine impersonating a person. In this setting, a human interrogator presents questions to two players, one a machine and the other human. As mentioned in the Introduction, the test is expected to be organised so that the mode of communication does not reveal anything about the players, which in the standard setting means communicating via a computer interface of textual inputs and outputs. If the interrogators fail to identify the computer player *as a computer* often enough (in a modest version this is understood as 30 percent of the test runs (Copeland 2000)), the computer (or rather the software the computer is running) is considered to have passed the Turing test. This, in Turing's (1950) original formulation, implies that the computer can think like humans.[10]

That is how the Turing test is mostly understood nowadays, but Sterrett (2000, 2020) argues that two different tests can be identified in Turing's 1950 article. The test described above she calls the *Standard Turing Test* (STT). But Turing initially describes another test, what Sterrett calls the *Original Imitation Game* (OIG). The difference is small but important. In STT, one output comes from a machine and the other one from a man, and the interrogator then tries to decide which output is the man and which is the machine. In OIG, however, one output comes from a woman and the other either from a man or a machine. The interrogator then tries to ascertain which one is the woman. Thus, in STT the interrogator always knows that one output comes from a machine, while in OIG they do not. This implies that in OIG, unlike in STT, also the human player needs to impersonate something that they are not.

Aside from this difference between OIG and STT, as specified by Sterrett (2000, p. 544), there are two other key differences. First, STT seems to be very sensitive to the interrogator's level of skill, whereas in OIG it is not as important. Second, only in OIG can the machine do *better* than the man. This second point may seem irrelevant because the test is not aimed at detecting whether a machine can be more intelligent than a human being. While this is true, the fact that in STT the computer can at best match the human could be indicative of a larger problem. STT being sensitive to the interrogator's skill level, however, goes directly against Turing's views. In a 1952 radio interview, he stated that the interrogators (in this version forming a jury) should not be "experts about machines" (Copeland 2000, p. 524).[11]

While these two differences should not be dismissed, it is the difference concerning impersonating that I believe to be the most important between OIG and STT. In STT, the computer and the human are clearly facing different tasks. The human is just being themself, while the computer is impersonating human intelligence. In OIG, they are both impersonating, making their tasks much more similar. It is because of this (and to a lesser degree the two other differences Sterrett points out) that I see OIG as the superior test: it gives the machine a fairer chance, given that it does not face a more difficult task than the human player.

Not everyone agrees with Sterrett's analysis. Proudfoot, for example, argues that the man-imitates-woman game is only used by Turing to score the computer-imitates-human game, i.e., the success rate of man imitating woman can be used as a standard for assessing success in computer imitating humans (Proudfoot 2013, p. 395; see also Wheeler 2020, p. 525). This interpretation is definitely already present in Turing's work (Turing 1950, p. 434). Here I cannot enter the debate whether Sterrett's or Proudfoot's interpretation is more accurate, but ultimately the difference seems to be theoretically minimal, as long as we focus the analysis on OIG. Under Proudfoot's interpretation, the required success rate for the computer is determined by the man impersonating a woman in the other game, so both for the man and the computer the task is about impersonation. The tasks just take place in different games. In Sterrett's OIG, the difference is that only one game is needed.

The practical difference between having one game or two games can be debated, but ultimately both Sterrett and Proudfoot come to agree on the most important matter involved: that impersonation is essential to the Turing test. The computer's performance is compared to the performance of a man impersonating a woman, which as a means of detecting intelligence clearly implies that such impersonation demands intelligence. However, how can we be sure that impersonation and intelligence are in this way connected? Indeed, why did Turing focus on a test involving impersonation, rather than simply observing the behaviour of a machine? As Proudfoot has asked: "…why base a criterion of intelligence on deception, rather than simply giving the machine a series of tasks to perform" (Proudfoot 2013, p. 395). This is an important question to ask and to understand the answer, it is helpful to return to the historical context of Turing's original work.[12]

[10] This is often described in terms of the computer emulating the human brain. See, e.g., (Copeland 2000).

[11] Copeland helpfully adds that neither should they be experts about the human mind (Copeland 2000, p. 525).

[12] The point Proudfoot wants to make is somewhat different from what will follow. Recall from the Introduction that in modern AI literature the approach focusing on creating machines that display human-like behaviour is often called "the Turing test approach". Proudfoot (2013) argues that this idea of the Turing test as a behavioural test is wrong. For a behaviourist approach, she argues, we could give the machine tasks to perform, and not focus on deception. Thus, she argues for a *response-dependence* interpretation of intelligence in the context of the Turing test, rather than a behaviourist one, given that the interrogator's responses are the key to intelligent ascriptions. See Wheeler (2020) for a further analysis of the response-dependence

3 False positives and false negatives

As mentioned in the Introduction, the standard modern understanding of the Turing test takes it to be a tool for making intelligence ascriptions concerning machines. Such intelligence ascriptions can go wrong in two ways. First, we can ascribe intelligence to unintelligent systems, resulting in *false positives*. Second, we may fail to ascribe intelligence to intelligent systems, resulting in *false negatives*. To understand where Turing stood on these issues, we need to understand the background and intellectual climate of the time. The main purpose of Turing's (1950) paper seemed to be replacing vague questions like "Can machines think?" with something empirically tractable. In modern parlance, as stated in the Introduction, the "Turing test approach" has come to limit artificial intelligence research to constructing human-like intelligence (see, e.g., Norvig and Russell 2021), but that limitation did not seem to be crucial for Turing. Instead, as I understand it, he proposed the test because displaying human-like behaviour can be empirically testable, while other types of intelligence might not be.

This empirical testability was arguably Turing's key contribution to the question of intelligence ascriptions. To see why, we need to consider the Turing test in its historical context. In his original 1950 paper, Turing engages with only one contemporary author writing about intelligence, the neurosurgeon Geoffrey Jefferson. This is telling, since the paper can be seen at least partly as a response to Jefferson. Jefferson (1949) warned about the dangers of "anthropomorphizing the machine" (p. 1110), meaning that as electronical machines become more advanced, there will be greater temptation to read "qualities of the mind" (ibid.) into them. Jefferson lamented that this had a been a problem in research on animal behaviour, and it could be an even greater problem with machines. In both regards, it is important to recognize the contemporary intellectual climate in which Jefferson was writing. After all, the previous year had seen the publication of Norbert Wiener's (1948) book *Cybernetics*, with the telling subtitle *Or Control and Communication in the Animal and the Machine*. Wiener's groundbreaking idea was that different types of circular causal feedback systems can be explained with the same principles, regardless of their embodiment. The recent progress with electronic computers, most notably the ENIAC and the first Manchester computer, contributed to optimism that machines could capture human cognitive processes, perhaps even think and be intelligent.

Jefferson, as an opponent of such possibilities, can be seen as an early meat chauvinist, according to whom thinking and intelligence are biological phenomena that can only be ascribed to biological systems. But as his remark on interpreting animal behaviour reveals, he was also convinced that there is something special about the *human* mind, made possible by the human nervous system, especially the speech areas of the brain (Jefferson 1949). Jefferson did not dismiss the possibility of animals possessing minds, but he nevertheless followed the contemporary view that there is a dramatic leap in intelligence from non-human animals to humans. Consequently, he was worried about mis-attributing human-like intellectual qualities to animals.

From a modern perspective, however, research in Jefferson's time seemed to suffer more from the reverse phenomenon. If anything, the problem was that animal intelligence was not recognised even in cases where it should have (see, e.g., de Waal 2017). For Jefferson, intelligence was thus not only about biological exceptionalism but also about human exceptionalism. The latter position has become unpopular in recent times. The question is, what is there to support the former position?

The great contribution of Turing was to reject this entire idea of human (or other biological) exceptionalism by moving the discussion to an empirical test immune to biases and misconceptions that make us misattribute intelligence. Jefferson focused on the problem of false positives. The Turing test approach, however, appears to be more applicable against cases of false *negatives*. Indeed, recently both Shieber (2016) and Gonçalves (2023) have argued that this should be seen as Turing's original purpose: he wanted to provide a sufficient condition for attributing intelligence, i.e., capture the true positives. From this perspective, it is not damaging if the application of the Turing test results in ascribing intelligence to non-intelligent machines, as long as it captures all the intelligent machines that are tested.

In the big picture, I am sceptical of this reading of Turing. Seeing his 1950 paper (partly) as a response to Jefferson, it is indeed a crucial part of the Turing test that it can help get rid of false negatives. Jefferson stated his view quite clearly: electronical machines cannot have intelligence. Turing then wanted to respond by creating a test which could potentially make us reject Jefferson's view. At the same time, it seems that Turing ultimately saw his test as a means to make accurate intelligence ascriptions, including rejecting false positives. In research on animal intelligence, the matter has changed dramatically from Jefferson's times. As the intelligence of non-human animals came to be widely accepted, animal research reached a new stage in which both false positives and false negatives are possible (for more, see Pantsar forthcoming).

In the field of artificial intelligence, the current state of the art is quite different from that in animal intelligence research. It is also different from early AI research, where false positives abounded. The 1956 theorem-proving program *Logic Theorist* by Newell, Simon and Shaw, for

Footnote 12 (continued)
notion of intelligence and the Turing test.

example, was understood by at least one of its creators to show genuine properties of the mind:

> [W]e invented a computer program capable of thinking non-numerically, and thereby solved the venerable mind/body problem, explaining how a system composed of matter can have the properties of the mind. (Simon 1991, pp. 206–207)

Nowadays few would say that the standard theorem-proving computer programs – obviously much more developed than *Logic Theorist* – have intelligence or other properties of the mind. They are understood as mechanical, rule-based tools, comparable to pocket calculators in their functioning (Pantsar 2024). But as false positives have become rare in AI research, it is reasonable to ask whether the AI community has come to embrace Jefferson's credo again. In the modern discussion, is it even feasible that an AI application could be accepted as intelligent?[13]

This is a timely and pertinent question, indeed probably more so than it was in 1950 when Turing presented his test. Much has happened in terms of development of AI applications, but the general form of one main objection remains the same. Whether it is stated explicitly or accepted implicitly, an important AI-sceptical argument remains that the physical functioning of machines is not conducive to the emergence of intelligence. This has been the main argument in the best-known philosophical anti-AI work, such as Searle (1980) and Dreyfus (1992), But aside from clever thought experiments (which Searle's Chinese room admittedly is), I fail to see a significant difference between their views and that of Jefferson. Consequently, if Turing's Imitation game is relevant for modern AI research, it is mostly relevant in the exact same way as it was when proposed in 1950: Turing's approach can provide us with a way to deal with false negatives in AI research.

4 Revising the Turing test

The focus on false negatives above may sound odd given the several much-publicised reports of computers passing the Turing test. Until 2020, researchers competed for the Loebner Prize for passing the Turing test, which the chatbot Eugene Goostman won in 2014. It was widely reported as the first program to pass the test (Warwick and Shah 2016). So much has been written on the Turing test in the AI literature that one would have expected this to be a monumental event in the history of AI, followed by extensive literature on the chatbot and the experiment. However, while there certainly was some reaction, the community's response was lukewarm at best. The report of the program passing the Turing test was met with scepticism and it certainly did not become a watershed moment in the history of AI.[14] Hence, Eugene Goostman and other reported cases of passing the Turing test should not count as false positives in intelligence ascriptions. Their success in the experimental setting notwithstanding, they were not generally seen as intelligent.

Even though my focus will be mainly on false negatives, I am not suggesting that false positives do not exist in AI intelligence ascriptions, nor that they are unimportant. There certainly are cases in which researchers describe artificial systems as being genuinely intelligent, and such cases are likely to be false positives with the current state of AI research. However, I contend that so far, the Turing test has not resulted in genuine false positives. No AI system has been generally accepted as being intelligent based on passing the Turing test. Indeed, in that sense, not much progress has been made. While adequate criteria for passing the Turing test have been discussed, there is little agreement on how the Turing test should be applied in intelligence ascriptions of artificial systems.

There have been, however, efforts to revise the Turing test to make it more suitable for intelligence ascriptions. Harnad (1991) proposed the "Total Turing test", in which the machine is situated in a real-world environment and considered intelligent if it could generally do everything that humans can. Harnad's suggestion took the Turing test outside the confines of verbal interactions, but in doing so, it seems to become needlessly limiting, since only self-propelling robots could possibly pass the test. A less demanding revision was proposed by Bringsjord and colleagues (2001) who, based on an observation by Ada Lovelace, suggest *creativity* as the key to intelligence ascriptions. Their proposed "Lovelace test" understood creativity as requiring a high degree of autonomy from the machine, which the authors took to be at the very least unlikely for computers (p. 25). The Lovelace test has since been revised by others (see, e.g., Riedl 2014), but it is safe to say that it has not been able to challenge the Turing test in the kind of attention it has drawn.

[13] One problem is the moving of the goalposts: tasks previously thought to require intelligence are no longer seen as such once AI applications for completing them are developed. This is the case with many of the AI success stories, such as playing games like chess and Go, as well as translation and image recognition.

[14] In 2000, the pre-eminent Turing expert Jack Copeland made two predictions. First, following Wilkes (1953), that passing the Turing test will be "hailed as one of the crowning achievements of technical progress" and, second, that passing the test by Turing's modest standards of three out of ten judges misidentifying the computer may happen in the near future (Copeland 2000, p. 537). The second prediction proved to be accurate as Eugene Goostman achieved exactly that level in 2014. However, the first prediction, at least in relation to that success, was way off the mark. While the event certainly made newspapers, ultimately it was not hailed as any kind of crowning achievement of AI research.

This history of the Turing test and its revisions may prompt the question of whether passing the Turing test as a target of AI research is obsolete, as argued by (Vardi 2014), or perhaps even harmful, as argued by (Hayes & Ford 1995). Or that it is perhaps confined to being a thought experiment, as suggested by Gonçalves (2023). Here I submit that if any of those interpretations are true, it is because of one central aspect of the Turing test both as it was first presented and in its later interpretations: namely, that it is based on *deception*. In this paper, by deception I simply mean that the computer programme in the Turing test presents itself as something that it is not, namely a human being. The chatbot Eugene Goostman, for example, presented itself as a thirteen-year-old boy from Ukraine who has, among other things, a pet guinea pig. Thus the whole enterprise of designing the chatbot was aimed to deceive the human interrogator in the Turing test (Warwick and Shah 2016).

Turing wisely warned about training computers in deception with the particular task of passing the test in mind (1951), but deception still remained at the core of his test. But whether understood as mimicking or more sophisticated impersonation, the test is only testing for one type of activity. This limitation can cause both false positives and false negatives. False positives can occur because impersonation may not require intelligence. A statistical model like GPT4.0 may make the chatbot ChatGPT appear intelligent by impersonating human answers, but few experts would ascribe any intelligence to it.[15] False negatives, on the other hand, occur because the computer could be intelligent in other ways but not be able to impersonate humans adequately.

That is why I make the following proposal: to make Turing-type tests relevant again, we need to get rid of the focus on deception.[16] Turing made an astute observation when he remarked that the computers should not be trained for the test. I believe that a Turing-type test works best when computers are developed without the test in mind. The question is, what should this kind of revised Turing test be like? Following the distinction made my Sterrett (2000), both the OIG and STT are based on deception since the computer under both interpretations is trying to impersonate something that it is not. As argued by Sterrett, the OIG is the superior test because in it also the human (the man) engages in deception, thus having a similar task to the computer. I agree that for the Turing test to be relevant, it is important that the task of the computer is similar to the task of the human subject.[17] However, I want to challenge the idea that this task should be based fundamentally on deception.

Indeed, I believe that the focus on deception is one important reason why the AI community does not seem to take Turing tests seriously as actual tests for intelligence. Eugene Goostman (or rather, its developers) managed to deceive 33 percent of the interrogators, which is an impressive feat, given that the only task of the interrogators was to recognise the computer as a computer. But the way it did was in many ways suspicious. Indeed, the very fact that the chatbot was programmed to present itself as a Ukrainian boy in his early teens should raise concerns. By impersonating a non-native speaking boy, the chatbot's functioning made the interrogators less likely to recognise unusual answers as computer-generated. Hence Eugene Goostman was based on deception in a particularly problematic way, due to its failings being masked by the character it was impersonating.[18]

However, the problem with the focus on deception goes deeper than that. Why would we connect intelligence with deception in the first place? This connection can cause both false negatives and false positives in intelligence ascriptions. Many birds can deceive predators by acting injured, but such distraction displays are standardly considered to be instinctive rather than intelligent behaviour.[19] False negatives, on the other hand, can occur because animals may be intelligent in other ways while not possessing the ability to deceive. Indeed, this danger of false negatives can be even more serious in the case of artificial intelligence. What would be the motivation to develop an AI that can conceal its identity as a computer? For the purposes of the Turing test, the motivation is obvious. But that is a very artificial and special circumstance. Indeed, it seems likely that in most cases we *want* to be able to identify computers as computers, making the ability to deceive something that the developers are likely to avoid in AI systems. From this background, I want to explore the possibility of retaining a Turing-like setting for making intelligence ascriptions, while getting rid of the focus on deception. In the next section, founded on that principle, I will propose a rough framework for replacing the Turing test with a Community-based intelligence test (CBIT).

[15] Some researchers claim that ChatGPT based on the GPT-4 model has passed the Turing test (James 2023; Mei et al. 2024) but recent evidence based on an online public test does not bear this out (Jones and Bergen 2023).

[16] I am not alone in proposing taking the focus away from deception. The above-mentioned "Total Turing test" and "Lovelace test" can both be seen as similar ventures. Another revision, recently proposed by Damassino (2020), explicitly distances itself from deception. In his "Questioning Turing test", the machine is tested on its ability to do *enquiry* (like a doctor or a detective, for example), rather than its ability to imitate humans in conversation. In making intelligence ascriptions, the success and strategy of the enquiry are then assessed by the evaluator.

[17] As mentioned earlier, this can be done in two ways. Either we can have the OIG test where both the man and the computer impersonate a woman (Sterrett 2000) or the man-imitates-woman test can be used for scoring the computer-imitates-man test (Proudfoot 2013).

[18] This kind of approach is sometimes called "artificial stupidity" in the literature (see, e.g., Damassino 2020).

[19] However, it has been argued that some such behaviour is learned (Walters 1990).

Before that, however, we need to distinguish between two types of deception. In my proposed test, the identity of the computer as a computer is concealed, just like in the Turing test. In practice, this may mean creating human-like identifiers for the computer. In this way, also in my proposed test some deception may be necessary. But beyond this first type of deception, there is a second, in the present context more important, level of deception. On this second level, the *output* of the computer is designed so that it can deceive the interrogator, as in the Turing test. This may or may not be the primary purpose of the computer programme, but on the second level, it is a key characteristic of it. It is this second type of deception that I will focus on.

5 The Community-based intelligence test

Above I have identified the focus on deception (of the second type) as the key problem with the Turing test, whether understood as the Original Imitation Game or the Standard Turing Test. In this section, I will present a test for making intelligence ascriptions for AI systems that is not fundamentally based on deception. The Community-based intelligence test (CBIT) I propose is not based on the AI system impersonating humans. It is, however, based on the AI *behaving* in a human-like fashion. Hence, while my approach will be a departure from the traditional understandings of the Turing test, it is still part of the Turing test approach in the AI literature.

This distinction between impersonating humans and behaving in human-like fashion is important to make, even though it may not always be clear in practice. For an AI to learn to behave in a human-like manner, it may use imitation as a key learning method. Here I understand imitation in the standard psychological way as copying the behaviour of another agent (see, e.g., Zentall 2006). In modern robotics research, this kind of imitation learning is an important implicit training method (Billard and Grollman 2012). However, given that imitation is important also for human children in developing their intelligence, the presence of impersonation strategies is not by itself detrimental for developing genuine intelligence, and hence for the possibility of intelligence ascriptions.[20] The distinction I propose is thus meant to emphasise that the impersonation of humans is not the *target* of the development of AI systems. The AI system may impersonate humans in the learning process by, for example, detecting and adopting common patterns in human behaviour. But we need to distinguish between this type of imitation learning and impersonation with the specific purpose of passing as humans. Both can be part of the Turing test approach, as it is generally understood in the AI literature. But only the latter is necessarily connected to deception (of the second type).

An example makes this clearer. In the case of developing artificial mathematical intelligence, we may aim to train the AI to be as human-like as possible in its outputs. In this way, it could develop human-like criteria for recognising interesting theorems and proofs, therefore providing mathematical outputs (ultimately even research papers) modelled after those provided by human mathematicians (Pantsar 2024). However, the purpose of this training and development process is not to create an AI that can deceive the mathematical community about its identity; it is to create an AI application that can help mathematics progress. The possible intelligence of such an application is a secondary question. It is, however, a particularly interesting question. If a mathematical AI does contribute in this way to the mathematical community, in what kind of scenario would we be prepared to ascribe intelligence to it? Here I propose that its putative intelligence could be assessed by a form of Community-based intelligence test.

What exactly is that form? That is a question that should be considered separately in each field where we want to assess machine intelligence. The aim of AI research is often disclosed in terms of developing artificial *general* intelligence (AGI), but this approach seems unnecessarily limiting. If an AI can possess human-like mathematical intelligence, for example, I believe it is a *bona fide* question how we can recognise that intelligence. Therefore, consistent with the understanding of intelligence presented in the Introduction, the CBIT approach is not limited to general intelligence. Instead, it is meant to be applicable also for evaluating domain-specific artificial systems in terms of their intelligence.[21] The distinction between domains, which is potentially a difficult prospect, is itself not crucial for the present approach. The important matter is that in CBIT, we can limit our considerations to a specific subfield of intelligent activity and the way an artificial system behaves within a relevant community.

In practice, the CBIT approach means inserting an artificial system into a community of human agents, such as those formed by mathematicians. It is this community that is then

[20] While psychologists agree that imitation is important for human and primate learning, there is a lot of debate how widely imitation is used in the animal kingdom. For a review, see (Zentall 2006).

[21] One important question is whether such intelligence needs to be human-like, as understood in the Turing test approach. In the case of mathematical intelligence, for example, could the aim not be maximal progress without any prior limit to human-like mathematics? Indeed, I believe that this may prove to be a more fruitful approach. However, for the present context of *detecting* intelligence, it may be more feasible that human interrogators ascribe intelligence to human-like AI systems. In any case, this worry of an anthropocentric bias was already noted by Turing (1950, p. 435), so it is hardly a problem exclusive to the present approach.

responsible for evaluating the intelligence of the system. Importantly, in CBIT, the members of the community are not informed that an AI system has entered the community. Instead, the AI system enters the community similarly to how a new human member would. The evaluation phase of the test then determines, after a sufficient number of interactions involving the AI within the community, whether the human members of the community are able to detect the AI as a machine. The way the intelligence evaluation takes place in the CBIT is thus more "organic" than traditionally associated with Turing tests, in the sense that the test settings are not (necessarily) created with the express purpose of the test.

As described above, the CBIT is divided into three stages:

(1) In the *introduction stage*, an AI system is introduced into a human community.
(2) In the *interaction stage*, the AI system communicates with the members of the community.
(3) In the *evaluation stage*, the human members are tested for whether they recognised the AI system as a computer or not.

The setting of CBIT is clearly a departure from the Turing test. In the Turing test, the interrogator(s) know all along that they are involved in a test involving both humans and computers. In CBIT, they are unaware of their participation in a test before the evaluation stage. Thus, in the introduction stage of CBIT, the AI is introduced as part of a community without revealing its identity as a machine, but also without revealing to the community that one of its members is a machine. Consequently, for the first two stages of CBIT, the human interrogators (i.e., the human members of the community) are unaware of being interrogators in a machine intelligence test. Instead, they assess the AI through its communicational behaviour just like they would do with new human agents introduced to the community. It is only at the evaluation stage of the CTT when the human members of the community are queried about their assessment of the AI member, that they learn that they are being tested.[22]

In CBIT, it is obviously necessary to hide the AI's identity as a computer from the community. In this, CBIT follows the Turing test. Amidst all the discussion about the Turing test, one may forget the brilliance of the fundamental part of Turing's proposal: namely, that to minimise biases, the identity of the players in the test should be concealed. In practice, this crucial aspect of the Turing test limits CBIT to online communities where all communication is text-based, where it is at least in principle possible to conceal the identity of the AI as a machine.

The practicalities of CBIT pose important questions and potential problems in all three stages. In the introduction stage, should the AI be the only new member introduced? Or should it be introduced alongside new human members? These are questions that have important consequences for the evaluation stage. If the AI system is introduced together with new human members, then there is a natural way to conduct the evaluation phase: namely, the interrogators are revealed that one of the new members was in fact a computer, and it is their task to identify which one. If they cannot identify the computer above a certain threshold, the AI system is thought to pass CBIT. If, on the other hand, the AI system is introduced as the only new member of the community, the evaluation phase needs to be different. The interrogators could be asked, for example, about their evaluation of the intelligence of different members of the community, one of them being the AI. If the AI is not estimated to be less intelligent than the other members, it should be considered to have passed the test.[23]

There are other practical questions that need to be considered before launching such a test. For example, the duration of the interaction stage needs to be decided before starting. Above I have proposed that the human members provide their evaluation of the AI system after a "sufficient number of interactions" within the community. But what is a sufficient number and what should those interactions be like? While it is possible to speculate on rough guidelines for such matters, ultimately they depend on the particular community. In the case of a mathematical community, for example, I have proposed that a suitable interaction could be based on AI producing articles presenting mathematical proofs (Pantsar, forthcoming). Rather than a single event (as in the Sokal hoax,[24] for example), the idea is that during CBIT the AI system would generate articles and communicate them, and about them, autonomously. If the mathematical community could detect it as a computer, then the system would fail CBIT. But if, after an extensive period (say, a

[22] It is crucial that the interrogators do not learn that they are being tested before the evaluation phase. Before that, they could always reveal the identity of the AI system by asking "are you a computer?" Since the AI system is not trained to deceive, it would reveal itself. But if the interrogators do not know that they are tested, this would be an unusual question to ask. Indeed, if they did ask the question, it already shows that they had suspicions about the identity of the new member, thus making it fail the test.

[23] The most straight-forward query, simply asking the interrogators whether the new member is human or a computer, would not work: in such settings, the interrogators have a bias for estimating even the human members as computers (Copeland 2000, p. 525). Presumably, such a bias would also be present if the question would not be direct. For example, the question "did you notice anything unusual about the member X?" could already prompt such a bias.

[24] In 1996, the physicist Alan Sokal published a nonsensical paper in the journal *Social Text* with the purpose of exposing flaws in certain academic fields and their publishing policies (Sokal 1996). For more on the "Sokal affair", see (Bricmont and Sokal 2003).

couple of years) the AI has been contributing actively to the community without having been detected as an AI, it would be considered to pass the CBIT. In this version of the test, the evaluation phase would be passive, but it could also be replaced by an active query of the human members of the community.

To carry out such a test, some deception of the first type may need to be used. For example, in the mathematical community there are typically identifiers – such as email addresses, academic titles and affiliations – that reveal the true identity of the members. Hence, for CBIT to work as an actual test, we would need to find a way of concealing the identity of the AI from the human members of the community. While this may be problematic, I assume that there can be online communities into which an AI system can be introduced without revealing it as a machine. If this assumption is feasible, I contend that CBIT is a step forward from the Turing test. Most importantly, that is because CBIT is not based on deception. While some deception (of the first type) may be needed to conceal the identity of the AI, the behaviour of the AI in the test is not based on deception (of the second type). In the mathematical scenario described above, for example, the behaviour of the AI is based essentially on the same aims as those of the human members of the community: generating and communicating mathematical articles (and other mathematical content).

Hence CBIT carries all the advantages that Sterrett (2000) identified in OIG over STT. First, there is nothing to prevent AI from outperforming humans as part of an online community. In CBIT, the human members set the standard for intelligence, but the (possible) intelligence of the computer is not limited to that of the human members. Second, the skill of the interrogators (i.e., the human members of the community) is not crucial for detecting the AI as a machine. Indeed, one key advantage of CBIT over both OIG and STT is that, before the evaluation part, the interrogators do not *know* that they are being tested. Any skill involved in detecting machines is therefore likely to play a smaller role than in traditional Turing tests. And finally, third, in CBIT the AI and the human perform similar tasks. This is because the AI is not trained to deceive by imitating or impersonating humans. Rather, it is trained to have human-like ability and use it for the kind of tasks in which humans in that community use their intelligence.[25]

[25] Of course, the CBIT framework could be abused so that the computer programme is designed with the express purpose of passing the test in mind, thus moving to the second type of deception. While there is no way to prevent that, I trust that the introduction of CBIT would discourage such approaches. After all, the designers would ultimately need to own up to this focus on deception when their results would be published.

6 CBIT and intelligence ascriptions

Earlier, I argued that a successful version of the Turing test should be able to prevent both false positives and false negatives. Now we can assess how CBIT fares in that regard. Clearly CBIT, just like the Turing test, is not immune to both types of misattributions of intelligence. However, I contend that it can fare better than the Turing test in both regards. In CBIT, false positives can occur when the community is not careful enough and accepts unintelligent systems, without detecting them as such. But this problem also extends to incompetent human members. Hence, we can assume that (at least some) communities organically try to weed out incompetent members, including non-intelligent artificial systems.[26] Therefore, while we cannot assume communities to be immune to false positives, it is in their own interest to find ways of minimising them.

In CBIT, false negatives occur if an intelligent AI is distinguished as a machine through some external characteristics and deemed *ipso facto* unintelligent. This is the meat chauvinism problem mentioned in the Introduction, and it cannot be avoided in CBIT. Indeed, in this respect CBIT may initially seem to be weaker than the Turing test. While the latter takes place in a controlled test setting which makes it easy to hide the physical composition of the AI, in CBIT the identity of the AI may be detectable through external factors. There are two ways to deal with this problem. First, we can run CBIT in experimental settings just like traditional Turing tests, making sure that the true identities of the members are not revealed. However, that would entail setting up an online community for the purpose of the text, which may not be feasible in practice. Hence a second way of dealing with the problem seems more practical: we create a human-like online identity for the AI and enter it into an existing online community. This identity may include a name, email address, avatar, and perhaps even a university affiliation. For carrying out CBIT, an optimal online community would thus be one with a minimum of possible identifiers of the members as humans.

As detailed above, I do not want to ignore potential practical issues in setting up and running the Community-based intelligence test. However, assuming that such issues can be resolved, I contend that compared to the Turing test, CBIT has a better chance of getting rid of also false negatives. This is mainly because in the kind of setting that I have described, the interrogators do not know that they are assessing the intelligence of one of the members of the community. Hence, among other advantages, it would remove one important bias in traditional tests: namely, the fact that the interrogators tend to mis-identify the human players as

[26] This certainly is the case with primitive bots that are a constant scourge of online communication platforms.

computers more frequently than vice versa (see, e.g., Copeland 2000, p. 525). This is likely due to the test setting, the interrogators knowing that one of the players is a computer. With the CBIT, this problem disappears.

CBIT also includes other advantages against biases. Due to the indispensable role of the community in the test, it is open to different notions of intelligence. Thus, it can avoid bias in AI research that limits the scope of research to particular manifestations of intelligence. Intelligence comes in many types that are likely to include significant cultural variation. The CBIT approach is open to such variation in different communities and hence it can tackle bias based on geography, language, education, gender, socio-economic status, ethnicity, and other divergent aspect of cultural backgrounds. Instead of focusing on finding general, uniform intelligence ascriptions, CBIT is sensitive to the diversity of intelligent communities already in its set-up.

One potential problem with CBIT is that it is characteristically a longer-span endeavour than the kind of test settings that the Loebner prize events, for example, used. However, I do not consider this objection to be a damaging one, because those types of tests do not seem to be taken seriously by the AI community, anyway. That CBIT involves a longer time span is in no way conducive to it being less of a scientific test, given that scientific experiments are not constrained to a short timeframe. That the test is conducted in the context of a community is not a major problem either, because there are multiple ways to record community reactions (e.g., responses, reviewer and editor reports, commentaries, queries, interviews, etc.). There are, as mentioned above, important issues concerning when the test can be considered passed, ones that should not be downplayed. However, most of these issues are present in similar (although not equivalent forms) also in the Turing test and its variations. We need to agree, for example, on the rate of misidentification that is considered the threshold for passing the test, and the number of interactions that is considered sufficient for making the evaluation. While in the CBIT such questions may need more consideration, they are not fundamentally different from the Turing test.

7 Conclusion

The Turing test refuses to go away. There is something about its central idea that fascinates people and still makes it a central platform for the question of detecting artificial intelligence. New tests, often going into wildly different directions, are still often framed in terms of the Turing test.[27] In addition, other types of intelligence tests have proven to be unfeasible for modern AI applications. The chatbot ChatGPT, for example, got an IQ of 155 on the Verbal IQ test used for humans (Roivainen 2023). Yet, as Roivainen points out, the same version of ChatGPT often fails in simple reasoning tasks. Some claim that models like GPT4.0 that is behind (as of writing this) the latest version of ChatGPT should already be considered cases of artificial general intelligence (see, e.g., Arcas and Norvig 2023), but many more are sceptical about that. In these discussions, I still see the potential for Turing-type behaviour-based tests in making intelligence ascriptions.

Against this background, I have proposed a framework for a new Community-based intelligence test (CBIT). Instead of developing AI applications to pass the Turing test, the CBIT is based on a more organic approach in which AI applications are entered as part of (online) communities. If the human members of those communities cannot detect the AI applications as computers, the AI is considered to have passed the test and should be considered intelligent. This does not imply that they are necessarily intelligent in the sense of artificial general intelligence. Rather, the idea is that an AI can possess domain-specific intelligence, for example, in areas like mathematics.[28]

I am hopeful that passing CBIT would be considered a breakthrough on a different level than in the reported cases of passing the Turing test, like that of Eugene Goostman. However, if this did not happen, it would be very interesting to ask *why*. If CBIT can be passed by an artificial intelligence, what else would be required to make intelligence ascriptions to machines? Are there any objections left, except for the meat chauvinist view that intelligence is limited to biological systems? If not, there are two ways forward. Either we accept that there are genuinely intelligent machines, or otherwise we simply need to treat intelligence as a notion for psychology and biology, not for computer science.

As of writing this, CBIT is still a theoretical prospect with no plans to implement it in practice. Before that can be done, many practical and ethical considerations need to be carried out. For example, we need to assess the ethical issues involved in creating a human-like online identity for an artificial system. We should also be careful about the effects that AI systems may have on communities, especially when

[27] For example, Suleyman (2023) proposes a test making an AI turn a $100.000 investment into a million dollars in a matter of months. However, remembering Sterrett's (2000) criticism that in STT computers are given different tasks than humans, certainly the Suleyman's test is in that sense a disaster. After all, how many humans would be able to pass it? Aside from this problem of false negatives (and the whole avaricious nature of the test), there is also a potential problem of false positives: is it clear that successful investing requires intelligence?

[28] In this scenario, the AI can of course also have intelligence in other domains.

their identities are concealed. If introducing an AI system can potentially damage the community, these risks need to be mitigated. This worry is particularly pertinent in cases in which the system is accepted as part of the community. In such scenarios, we need to trust the AI contributions, just like we have learned to trust human agents. However, trust in AI is an important research topic with its own characteristics (for a review of empirical research on the topic, see Glikson and Woolley 2020).[29]

Yet, it is important to recognise that these are worries that we should take seriously regardless of the status of CBIT. Given the immense progress in machine learning technology in recent years, having AI systems capable of passing CBIT may be possible sooner than we expect. Such AI systems may begin to play important roles in communities even if no tests like CBIT are ever run. I believe it is time that we as philosophers start preparing for that possibility. In that preparation, I hope that the Community-based intelligence test I have proposed in this paper can provide a fruitful platform for discussions.[30]

Acknowledgements This paper was developed during my time as a Senior Research Fellow at the Käte Hamburger Kolleg "Cultures of Research", RWTH Aachen University, Germany. I would like to thank Regina Fabry for very helpful comments on an early version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. No further funding for this research was provided.

Data availability No data was collected or processed.

Declarations

Competing interests The author declares that there are no competing interests.

[29] I am grateful to an anonymous reviewer for emphasising this issue.

[30] I would like to thank Regina Fabry for very helpful comments on an earlier version of this manuscript.

References

Arcas B, Norvig P (2023) Artificial General Intelligence is already here. Noema. https://www.noemamag.com/artificial-general-intelligence-is-already-here

Billard A, Grollman D (2012) Imitation learning in robots. In: Seel NM (Ed.), Encyclopedia of the Sciences of Learning (pp. 1494–1496). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_758

Boden M (2006) Mind as machine: a history of cognitive science. Oxford University Press

Bricmont J, Sokal A (2003) Intellectual Impostures (Main-Re-issue edition). Profile Books

Bringsjord S, Bello P, Ferrucci D (2001) Creativity, the Turing test, and the (Better) Lovelace test. Mind Mach 11(1):3–27. https://doi.org/10.1023/A:1011206622741

Chomsky N, Roberts I, Watumull J (2023) Opinion | Noam Chomsky: The False Promise of ChatGPT (March 8 2023). The New York Times. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Cianciolo, A. T., & Sternberg, R. J. (2004). Intelligence: a brief history. Blackwell Pub.

Clark A (2008) Pressing the flesh: a tension in the study of the embodied, embedded mind? Philos Phenomenol Res 76(1):37–59. https://doi.org/10.1111/j.1933-1592.2007.00114.x

Copeland BJ (2000) The Turing test. Mind Mach 10(4):519–539. https://doi.org/10.1023/A:1011285919106

Damassino N (2020) The questioning Turing test. Mind Mach 30(4):563–587. https://doi.org/10.1007/s11023-020-09551-6

de Waal FBM (2017) Are we smart enough to know how smart animals are? Norton

Dreyfus HL (1992) What computers still can't do: a critique of artificial reason. MIT Press

Gandy R (1996) Human versus mechanical intelligence. In: Millican P, Clark A (Eds) Machines and Thought. Oxford University Press, pp 125–136

Glikson E, Woolley AW (2020) Human trust in artificial intelligence: review of empirical research. Acad Manag Ann 14(2):627–660. https://doi.org/10.5465/annals.2018.0057

Gonçalves B (2022) Can machines think? The controversy that led to the Turing test. AI Soc 38:2499–2509. https://doi.org/10.1007/s00146-021-01318-6

Gonçalves B (2023) The Turing test is a thought experiment. Mind Mach 33(1):1–31. https://doi.org/10.1007/s11023-022-09616-8

Gonçalves B (2024) The Turing test argument. Routledge

Harnad S (1991) Other bodies, other minds: a machine incarnation of an old philosophical problem. Mind Mach 1(1):43–54. https://doi.org/10.1007/BF00360578

Hayes P, Ford K (1995) Turing test considered harmful. Proceedings of the 14th International Joint Conference on Artificial Intelligence1:972–977

James, A. (2023). ChatGPT has passed the Turing test and if you're freaked out, you're not alone. *TechRadar*. https://www.techradar.com/opinion/chatgpt-has-passed-the-turing-test-and-if-youre-freaked-out-youre-not-alone

Jefferson G (1949) The mind of mechanical man. BMJ 1(4616):1105–1110

Jones C, Bergen B (2023) Does GPT-4 pass the Turing test? (arXiv:2310.20216). arXiv. http://arxiv.org/abs/2310.20216

Mei Q, Xie Y, Yuan W, Jackson MO (2024) A Turing test of whether AI chatbots are behaviorally similar to humans. Proc Natl Acad Sci 121(9):e2313925121. https://doi.org/10.1073/pnas.2313925121

Minsky M (2006) The emotion machine: commonsense thinking, artificial intelligence, and the future of the human mind. Simon & Schuster

Norvig P, Russell S (2021) Artificial intelligence: a modern approach, Global Edition (4th edition). Pearson
Pantsar M (2024) Theorem proving in artificial neural networks: new frontiers in mathematical AI. Eur J Philos Sci 14(1):4. https://doi.org/10.1007/s13194-024-00569-6
Pantsar M (forthcoming) How to recognize artificial mathematical intelligence in theorem proving. Topoi
Proudfoot D (2013) Rethinking Turing's test. J Philos 110(7):391–411
Riedl MO (2014) The Lovelace 2.0 test of artificial creativity and intelligence (arXiv:1410.6142). arXiv. https://doi.org/10.48550/arXiv.1410.6142
Roivainen E (2023) I gave ChatGPT an IQ test. Here’s What I discovered. Scientific American. https://www.scientificamerican.com/article/i-gave-chatgpt-an-iq-test-heres-what-i-discovered/
Searle JR (1980) Minds, brains, and programs. Behav Brain Sci 3(3):417–424. https://doi.org/10.1017/S0140525X00005756
Shieber SM (2016) Principles for designing an AI competition, or Why the Turing test fails as an inducement prize. AI Mag 37(1):91–96. https://doi.org/10.1609/aimag.v37i1.2646
Simon HA (1991) Models of my life. Basic Books
Sokal A (1996) Transgressing the boundaries: towards a transformative hermeneutics of quantum gravity. Social Text 46(47):217–252
Spearman C (1927) The abilities of man (pp. xxiii, 415). Macmillan
Sterrett SG (2000) Turing's Two tests for intelligence. Mind Mach 10(4):541–559. https://doi.org/10.1023/A:1011242120015
Sterrett SG (2020) The Genius of the "Original imitation Game" test. Mind Mach 30(4):469–486. https://doi.org/10.1007/s11023-020-09543-6
Suleyman M (2023) My new Turing test would see if AI can make $1 million. MIT technology review. https://www.technologyreview.com/2023/07/14/1076296/mustafa-suleyman-my-new-turing-test-would-see-if-ai-can-make-1-million/
Turing AM (1950) Computing machinery and intelligence. Mind, LIX(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433
Turing AM (1951) Intelligent machinery, a heretical theory. In: Copeland BJ (Ed.) (2004), The essential Turing: The ideas that gave birth to the computer age. Oxford University Press, 472–475
Turing AM (2004) The essential turing: seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life plus The secrets of Enigma (Copeland BJ, Ed.; 1st edition). Clarendon Press
Vardi MY (2014) Would Turing have passed the Turing test? Commun ACM 57(9):5. https://doi.org/10.1145/2643596
Walters JR (1990) Anti-predatory behavior of lapwings: field evidence of discriminative abilities. Wilson Bull 102(1):49–70
Warwick K, Shah H (2016) Can machines think? A report on Turing test experiments at the Royal Society. J Exp Theor Artif Intell 28(6):989–1007. https://doi.org/10.1080/0952813X.2015.1055826
Wheeler M (2020) Deceptive appearances: the turing test, response-dependence, and intelligence as an emotional concept. Mind Mach 30(4):513–532. https://doi.org/10.1007/s11023-020-09533-8
Wiener N (1948) Cybernetics: or control and communication in the animal and the machine. MIT Press
Wilkes M (1953) Can machines think? Proc IRE 41(10):1230–1234. https://doi.org/10.1109/JRPROC.1953.274272
Zentall TR (2006) Imitation: definitions, evidence, and mechanisms. Anim Cogn 9(4):335–353. https://doi.org/10.1007/s10071-006-0039-2