Transcendence: Measuring Intelligence

Marten Kaas

University of York

Abstract

Among the many common criticisms of the Turing test, a valid criticism concerns its scope. Intelligence is a complex and multi-dimensional phenomenon that will require testing using as many different formats as possible. The Turing test continues to be valuable as a source of evidence to support the inductive inference that a machine possesses a certain kind of intelligence and when interpreted as providing a behavioural test for a certain kind of intelligence. This paper raises the novel criticism that the Turing test represents an example of Goodhart's Law operating in the field of artificial intelligence. As one measure towards the goal of creating genuinely intelligent machines, the Turing test must not be confused with the goal itself. Moreover, the Turing test ought to be augmented such that through its use additional evidence could be secured to support the strong inference that a machine, were it to pass the Turing Test, could think like a human.

In the film *Transcendence*, Will Caster (Johnny Depp) supposedly uploads his mind into a quantum computer after being fatally poisoned. I say "supposedly" because many of the characters in the film seriously doubt that the "real" Will has survived the uploading process. Even his wife Evelyn Caster (Rebecca Hall) eventually begins to question whether the image of her husband on the monitors and the voice coming through the speakers is in fact *her* husband, and not some other entity. In addition to philosophical questions of identity (e.g., is Will the same person post-upload or a new "person"?), the film also raises questions about the nature of consciousness and the possibility of conscious artificial intelligence. It is the latter set of questions with which this essay will engage, and in particular the epistemic problem of other minds and the relevance of the Turing test. In short, I argue that, properly understood, the Turing test continues to be valuable as a source of evidence to support the inductive inference that a machine possesses a certain kind of intelligence when interpreted as providing a behavioural test for a certain kind of intelligence.

Journal of Science Fiction and Philosophy vol.6 (2023)
 ISSN 2573-881X
 Published 2023 June 1.

Are You Self-Aware?

At the beginning of *Transcendence* we meet Will's friend Joseph Tagger (Morgan Freeman) who introduces the audience to Will's supercomputer named PINN (for Physically Independent Neural Network). To showcase some of PINN's abilities to FBI Agent Donald Buchanan (Cillian Murphy), Joseph asks PINN, "Can you prove that you are self-aware?" PINN responds: "That's a difficult question Dr. Tagger. Can you prove that you are?" The characters share knowing smiles with each other and, presumably, the audience, because we are all aware that PINN is not in fact self-aware (or conscious).[1] This exchange however is repeated. Later in the film, when Joseph meets post-upload Will, he is quite stunned. Post-upload Will asks, "Are you surprised to see me Joseph?" To which Joseph responds, "That depends. Can you prove that you are self-aware?" Like PINN, post-upload Will responds: "That's a difficult question Dr. Tagger. Can you prove that you are?" This time the characters share looks of shock and trepidation while Evelyn tries to defuse the situation by remarking that Will certainly has not lost his sense of humour.

Setting aside the plausibility and science of uploading a human mind, there are at least two interesting philosophical questions embedded in Joseph's question about self-awareness. The first is, how can one prove that they are indeed conscious and self-aware? In short, how does one prove that one has a mind? The second is, how do we know that other humans are indeed conscious and self-aware? Put another way, what sort of evidence justifies claims that one knows that other humans, and perhaps even certain other non-human entities like animals or machines, have a mind? I maintain that the answers to both of these questions is through their behaviour. In particular, I will attempt to address the second question by arguing that the Turing test is a valuable source of evidence to justify claims that one knows that another entity has a mind.

Background

Although the Imitation Game has been described many times since Turing, a brief recapitulation will be useful. In short, the Imitation Game involves three participants: a woman, a man, and an interrogator. In the first instance, Turing asks us to imagine whether it would be possible for the interrogator to successfully identify the woman if the interrogator is only allowed to interact with the woman and man via teletype (i.e., the interrogator only sees typewritten answers to their questions and so cannot rely on differences in handwriting, voice patterns, appearances, etc., to identify the woman) (Turing 1950). The woman's objective is to aid the interrogator as much as possible whereas it is the man's objective to imitate, as it were, a woman and thereby fool the interrogator. In the second instance, Turing asks us to imagine that a machine replaces the man in the game. If a machine was able to fool the interrogator, what might that tell us about such a machine? Turing maintains that questions arising from the second instance of the Imitation Game ought to replace the more contentious and ambiguous question, "Can machines think?" Indeed, since Turing suggested the Imitation Game (hereafter the Turing test) as a kind of test for thought or intelligence[2] in a machine,

[1] These two terms, "self-aware" and "conscious," are conflated in the film. I will also use these terms interchangeably unless otherwise specified.

[2] I will use the terms "thinking" and "intelligence" interchangeably unless otherwise specified.

most philosophical discussions have revolved around its value or usefulness. Now, more than 70 years after Turing first proposed the test, I maintain that it still has significant value.

As Moor (1976) argued, the Turing test has value because it can be considered as a source of evidence to support the claim that a machine thinks. Like many other claims, the claim that a machine thinks is not one to endorse without good reasons or, preferably, evidential support. Turing, for his part, appeared to think that a purely philosophical debate would lead to interminable discussions surrounding the meaning of terms like "machine" and "think," so he offered the Turing test as a challenge instead (Turing 1950). The Turing test is, in effect, a challenge to put up or shut up (i.e., a challenge to set arguments aside and attempt to build a thinking machine), and there have been many attempts over the years to create a machine that could pass the Turing test, although all have failed thus far. Nevertheless, the test remains a goal for some in the field of artificial intelligence and a constant source of philosophical debate. Many scholars have objected, for example, that the Turing test fails as a test of genuine intelligence because it is a measure of behavioural fidelity. A common objection to the Turing test focuses on the fact that the test revolves around the generation of appropriate outputs or ends and that any appropriately designed system could produce the "correct" output for a given input (Gunderson 1964). Moreover, it has been objected that even if some machines could replicate the outputs typical of a human taking the Turing test, differences in the machine's information processing could be such that we are not warranted in ascribing it genuine intelligence (Block 1981).

Searle's famous "Chinese Room" thought experiment exemplifies these types of objections to the Turing test (Searle 1984). Imagine the following scenario. Suppose that you, having no understanding of Chinese, are placed in a room full of baskets of Chinese symbols and a rulebook with instructions that specify how you ought to manipulate these symbols. While in this room you are able to receive inputs of Chinese symbols from the outside and generate Chinese symbol outputs according to your rulebook, which you then send back out such that your "answers" are indistinguishable from those of a native Chinese speaker. Intuitively, as Searle argues, you do not understand Chinese even though, from an outside observer's perspective, you, or presumably the room, behave exactly as if you do understand Chinese (Searle 1984). Searle goes on to claim that, assuming that you, the human in the Chinese Room, are sufficiently analogous to a digital computer program that merely manipulates symbols according to syntactic rules, it follows that no digital computer program, as a formal symbol-manipulating system, could understand Chinese. Though the room appears to understand Chinese because it produces intelligible outputs for a given input, to attribute understanding or genuine intelligence to the room is, according to Searle, a mistake.

Another common objection to the Turing test concerns the link between mechanism and thinking. As many scholars have noted,[3] thinking about thinking machines raises a host of other, often emotional, considerations. Chief among these is the relationship between the mechanistic terms used to describe systems like digital

[3] Douglas R. Hofstadter, for example, in *The Turing Test: A Coffeehouse Conversation* uses dialogue between characters to explore the emotional baggage that accompanies many discussions of thinking machines.

computers, for example, and the naturalistic terms used to describe systems like the human brain.[4] Critics of the Turing test, and of artificial intelligence in general, have argued that intelligence does not have a mechanistic basis or that it is not something that can be described mechanistically. Similarly, Turing has been accused of begging the question because he assumes that the brain is a machine. In either case, critics maintain that a machine could never pass the Turing test.[5] A final common objection, and one that I argue is worth considering carefully (and which I will elaborate on later), is that the Turing test fails as a test for genuine intelligence because it is too narrow. Even if a machine is able to pass the Turing test by behaving appropriately in the context of a question-response game, that does not necessarily imply that it can think on the level of a normal human.

Objections to the Turing Test

In keeping with Moor (1976), I maintain that none of these common objections to the Turing test apply, with the exception of the final objection. Let us consider each objection in turn. The first objection concerns the validity of the Turing test vis-à-vis its ability to distinguish genuinely intelligent machines from those machines that merely appear to be intelligent. Again, this is because the Turing test revolves around the generation of appropriate responses to given questions, and it is conceivable that these responses could be brought about either accidentally or through trickery. Historically, it was objected that the Turing test presupposed a behaviouristic reduction of intelligence and that any such behaviouristic analysis was misguided. As Moor rightly points out, however, "our knowledge of thinking by others has an inductive basis" and it would be odd to hold machines to some higher standard (Moor 1976, 252-253). Moreover, rejecting the validity of the Turing test in this manner can lead to an extreme solipsistic viewpoint since it no longer concerns the question of whether machines can think, but whether any person besides oneself can think. After all, is it not on the basis of behaviour that we conclude other people can think? This epistemological problem of other minds, as it is known, is often briefly considered before being rejected for reasons that run the gamut from arguments by analogy (e.g., other humans behave in similar ways to me under similar circumstances, therefore they must think like me) to arguments from a shared history (e.g., other humans share a biological and evolutionary history with me, therefore they must think like me) to arguments from best explanation (e.g., the best way to explain your behaviour is on the basis that you have thoughts, desires, beliefs, etc., in short, a mind). Though potentially compelling, these arguments are ultimately chauvinistic[6] and fail to challenge the validity of the Turing test. Importantly, this is not to say that any machine that passes the Turing test must

[4] At one point in *Transcendence* Max Waters (Paul Bettany) tells Evelyn that humans can reconcile illogical conflicts arising from our *emotions* (e.g., loving someone and yet hating what they have done) whereas machines can do no such thing. This is presumably because, as Max claims in this particular scene, the human mind cannot be reduced to a series of *electrical impulses* such as those one might find in an AI.

[5] See, for example, Geoffrey Jefferson's Lister Oration, "The Mind of Mechanical Man" (1949), and Michael Apter's *The Computer Simulation of Behaviour* (1971).

[6] Consider an alien, for example, passing the Turing test. Ought we to deny that this alien can think? Probably not, especially if the only evidence of its thinking was collected from the Turing test.

necessarily be able to think. Rather, passing the Turing test is one piece of valuable evidence to justify the induction that a certain machine can think (Moor 1976).

The second objection resembles the first in that it concerns the validity of the Turing test but with regard to the internal operations of the machine being tested. Suppose that a machine does pass the Turing test and in our curiosity we take it apart in an attempt to understand how the machine passed the test. Imagine that upon closer inspection we realize that the machine was able to pass the Turing test because its memory simply contained a vast but finite collection of responses to innumerable conversational prompts and questions (think of the rulebook that one might find in Searle's Chinese Room). In light of this new evidence, and in spite of the fact that the machine passed the Turing test, it seems reasonable to deny the machine intelligence. Despite the machine's initial appearance of intelligence, its internal operations are such that it does not possess genuine (human) intelligence, one aspect of which is conversational flexibility and which is something this hypothetical machine lacks. The right question or conversational topic would reveal that this machine can only respond in a preprogramed way, and so lacks genuine intelligence. But this objection does not apply to the Turing test *per se*, only to the claim that a certain machine thinks. As Moor correctly points out, we must be careful to distinguish between two claims: (1) that evidence of the internal operations of a system might falsify a justified inductive inference that a system can think like a human and (2) that evidence of the internal operations of a system is necessary to make a justified indicative inference that a system can think like a human (Moor 1976). Claim (1) is entirely reasonable whereas claim (2) is far too strong to warrant serious endorsement. As mentioned above, the common sense inductive inference that other humans think is based on nothing more than their outward behaviour. Critics of the Turing test and of artificial intelligence in general have been unable to articulate why machines ought to be held to some different standard.

The third common objection to the Turing test is the least forceful, in part because it appears to stem from emotional outrage that human intelligence could ever be understood in mechanical terms and a widely held belief that human beings are somehow unique and set categorically apart from the rest of the natural world. This attitude permeates pop culture and is evident in the film *Transcendence*. In short, it has been objected that the Turing test is either question-begging, i.e., assumes that human intelligence is mechanistic in nature (or can be described mechanically), or is a valid test for intelligence but is one that no machine will ever pass because human intelligence is unique and does not yield to a mechanistic analysis. But developments in the field of artificial intelligence research, e.g., the development of deep neural networks and sophisticated machine learning techniques, have allowed systems to attain levels of "intelligence" unimagined even a few decades ago. DeepMind's machine AlphaZero, for example, cannot pass the Turing test but there is no denying that it plays the games of chess, Go, and shogi intelligently (Silver et al. 2018). So this third objection is revealed for what it truly is, not an objection to the Turing test, but an empirical claim concerning the impossibility of constructing an intelligent machine.

The final common objection to the Turing test concerns its scope. That is, it can be argued that the Turing test is inadequate as a test for intelligence because it is only one test of intelligence, i.e., conversational intelligence. However similar to the second objection concerning the internal operations of the system being tested, Moor highlights

that we must be careful to distinguish between two different claims: (3) that additional evidence that cannot be directly obtained from the Turing test might falsify a justified inductive inference that a system can think like a human and (4) that additional evidence that cannot be directly obtained from the Turing test is necessary to make a justified inductive inference that a system can think like a human (Moor 1976). I agree with Moor that claim (3) is true; however, I disagree with his assessment that claim (4) is false. Moor maintains that "it is simply a misleading numbers game to suggest that the Turing test is only one test" and that the "test provides a format for directly or indirectly examining any of a wide variety of activities which would count as evidence for thinking" (Moor 1976, 256). Progress in research on artificial intelligence, machine learning and robotics has revealed that, contrary to what Moor thought (and indeed what most proponents of artificial intelligence thought, not that Moor was one of those proponents), the Turing test is only one measure of a certain kind of intelligence. Despite the fact that we might be justified in making the inference that a machine is intelligent if it passes the Turing test, it would be more accurate to say that we are justified in making the inference that the machine possesses a certain kind of intelligence. Claim (4) that Moor identifies can be challenged on the grounds that human thought is complex and multi-dimensional, so much so that additional evidence that cannot be directly obtained from the Turing test *should be desired* to make a justified inductive inference that a system can think like a human. In short, I grant that additional evidence beyond the Turing test is not necessary to make a justified inductive inference that a system can *think*, but I argue that one ought to desire additional evidence before making the stronger inference that a system can think *like a human*.

The Turing Test and Goodhart's Law

Given the above considerations, I maintain that one valid criticism of the Turing test is that it represents a paradigmatic example of Goodhart's Law operating in the field of artificial intelligence. First articulated by Charles Goodhart (1984) to describe certain economic practices, Goodhart's Law roughly states that when a metric for a target becomes the new target, that metric ceases to be a good one. As a toy example consider that one measure of the tidiness of my bedroom is the amount of clothing lying on the floor. If I substitute the measure for my goal, then some odd consequences might arise. For example, if I take all of the clothes lying on the floor and glue them to my wall and ceiling, I will have "successfully" cleaned my bedroom. The point is that, by treating what was once a good measure of the goal as the goal itself, I have rendered it a poor measure of the goal altogether. The same can be said of the Turing test. It is certainly one good metric insofar as we are interested in gathering evidence to support the inference that a machine can think, but passing the Turing test is not the target itself. The target is the creation of a genuinely intelligent machine. This criticism is linked to my earlier disagreement with Moor about the falsity of his claim (4) that additional evidence that cannot be directly obtained from the Turing test is necessary to make a justified inductive inference that a system can think like a human. And this is because I maintain that while the Turing test is still useful, it does not in fact provide a format for examining a wide variety of activities that could count as evidence for thinking.

Various scholars have similarly pointed out this connection between the Turing test and Goodhart's Law. Crosby (2020) for example highlights that passing the Turing test is a very different goal from trying to build a thinking machine. Researchers

interested in the former goal might consider attempting to win the Loebner Prize, a Turing-test-inspired competition,[7] whereas researchers interested in the latter might consider getting machines to complete animal cognition tasks, for example. This is precisely the view taken by Crosby (2020) who asserts that there are non-verbal tests (e.g., the Aesop's Fable task)[8] that better test for evidence of thinking in machines. Crosby (2020) claims that the Turing test is a mostly successful operationalisation of the question "Can machines think?", depending on the quality of the judges, with the exception that it sets the bar too high as it is not passable by most exemplars, i.e., humans. Moreover, as Bieger and Thórisson (2018) note, it is task-oriented evaluations in general that tend to fall victim to Goodhart's Law. In the field of artificial intelligence research, because problems and tasks are often defined precisely (e.g., create a machine that can play chess), what is usually measured when testing machines is not intelligence or capacity for thought *per se*, but performance (Hernández-Orallo 2017). This results in a specialization drift, i.e., "the conscious or unconscious tendency of AI researchers to specialize to a particular task, or even worse, to overfit to a benchmark (known in other areas as Goodhart's Law)" (Hernández-Orallo 2020, 2). Importantly, while other scholars argue that the Turing test ought to be replaced by some other type of test or abandoned altogether, I maintain that the Turing test is still a useful and powerful tool. My aim in what follows is therefore to defend the novel position that, properly understood, the Turing test is a valuable source of defeasible evidence to support the claim that a machine possesses a certain kind of intelligence. For the moment, however, let us consider *Transcendence* once more.

Intentionally Ignoring the Evidence

It cannot be overstated that the Turing test is a source of *defeasible* evidence, and one of the reasons for this is that the test is concerned only with outputs. The entity under scrutiny is treated simply as a black box whose internal operations are ignored. In fact, the problem is considerably worse, especially if we consider modern AI, humans and imagined artificial hyperintelligences such as post-upload Will. Modern deep neural networks, human cognition and artificial hyperintelligences running on networks of quantum processors are not merely treated as black boxes, they often are black boxes whose internal operations are either too difficult to grasp, not well understood, or some combination thereof.

Intelligence, like intentions or a mind, though they presumably have internal origins, manifests through outputs, i.e., outward behaviour. *Transcendence* is therefore a cautionary tale concerning the risks of disregarding behavioural evidence. Consider that, throughout the film, although the characters doubt post-upload Will's intelligence, consciousness and identity, none are doubted to the degree that post-upload Will's (supposedly) good intentions are doubted. And yet, we might ask, on what basis are the characters in the film arriving at this conclusion? Among other benevolent acts, post-upload Will returns sight to a blind person, heals a man after he was beaten, heals

[7] Launched in 1990 by Hugh Loebner, cash prizes were awarded to those who could build a computer program that judges considered to be the most human-like via textual conversation.

[8] The Aesop's Fable task or experiment is based on a story of a thirsty crow and tests a subject's understanding of the world. In short, a subject must obtain an out-of-reach reward floating in a tube of water by dropping stones into the tube and thereby raising the water level to reach the reward, or the water itself in the case of the thirsty crow.

multiple people after they are shot and restores mobility to a paralyzed person. Indeed post-upload Will appears to anticipate that people will doubt his good intentions and so invites Joseph Tagger and Agent Buchanan to his facility, the Brightwood Data Center (BDC), to demonstrate his capabilities, e.g. returning sight to a blind person using nanotechnology. Post-upload Will is adamant that they (the team at the BDC, including Evelyn) are not hiding anything and not coercing anyone, only helping those that seek their help.

And yet, such displays fail to convince Tagger and Agent Buchanan, among many other main characters, of post-upload Will's good intentions. They instead latch onto the fact that post-upload Will, when healing people using nanotechnology, also connects those people to himself such that he can remotely control them whenever he desires. After visiting the BDC, Agent Buchanan remarks that regardless of whoever or whatever post-upload Will is, he appears to be building an army and therefore ought to be considered a threat, one that must be dealt with swiftly via a pre-emptive strike. There is, in short, essentially no amount of behavioural evidence that post-upload Will could produce to convince many of the humans in *Transcendence* that he intends only to help and not harm them. This is because intentions, like a mind and intelligence and consciousness, as a result of their internal origins, can only be inferred on the basis of observable evidence. Coupled with the opaque, black-box quality of human cognition and current deep neural networks (never mind futuristic hyperintelligent AIs), it follows that there are epistemic limits that preclude the possibility of ever having *certain* knowledge of another entity's intentions, for example.

Even Evelyn comes to doubt post-upload Will's good intentions. Late in the film she expresses that what they are doing at the BDC is no longer fulfilling for her because "Will" is no longer with her. In his effort to understand her change of heart, post-upload Will reveals that he has been monitoring Evelyn's vitals and biochemistry in an effort to empathize with her. Though post-upload Will complies with Evelyn's demand to show her all of the information he has been collecting about her, she responds to his honesty with shock and revulsion, claiming that he has intruded upon her private thoughts and feelings.

All of this is to say that when it comes to intentions, and perhaps other aspects of mind, epistemic certainty is an unachievable ideal. Epistemic uncertainty is the rule, not the exception. Defeasible inferences made on the basis of behavioural evidence, for example, are the pragmatic goal to strive towards. It is therefore of utmost importance that we do not ignore, disregard or misconstrue behavioural evidence. For the characters in *Transcendence*, their misconstrual of the evidence results in Evelyn's and post-upload Will's deaths, depriving the world of a benevolent hyperintelligent AI. For us, disregarding or misconstruing the evidence might mean denying a machine intelligence when we really ought to consider it intelligent. In the near future, disregarding or misconstruing the evidence might also mean denying a machine rights or moral status when we really ought to consider it as having interests and needs. Such a discussion however is far outside the scope of this paper, and so I return now to the Turing test and the evidence one might collect to support the inference that a machine possesses intelligence, or at least a certain kind of intelligence.

Passing the Test

Before considering different kinds of intelligence and the different metrics one might use to gauge whether a system possesses a certain kind of intelligence, some clarifications are necessary. First, Moor does recognize that "further testing beyond the Turing test would be valuable and that the results of such further testing might make one revise inferences based on the results of the Turing test alone" (Moor 1976, 255-256). It is worth repeating that the Turing test is valuable because it provides a good format for collecting inductive evidence to support the claim that a machine thinks. In contrast to Moor however, I have argued that further testing should be conducted and additional evidence collected in order to justify the inductive inference that a machine thinks *like a human*, an inference, I maintain, that requires more evidence to support than the inference that a machine *can think*. Second, unlike Moor, I believe that the Turing test is valuable if one interprets it as a behavioural test for a certain kind of intelligence. To reject the usefulness and value of the Turing test as a behavioural test for intelligence is to hold machines (or non-humans) to some different standard than what is commonly held for humans. Behavioural (and functional) tests for intelligence are often challenged on the grounds that they fail more generally as adequate theories of mind, e.g., they fail to explain the existence of qualia, to use the philosophical jargon.[9] And yet other humans are assumed to have minds not because we have privileged access to their minds but because of observations we make of them at a behavioural (or functional) level. Given that it is on the basis of indirect behavioural evidence that we conclude other humans can think, machines ought to be held to this same standard. This is especially true considering that no compelling reasons or arguments have been given to demonstrate why it is that machines ought to be held to some different standard. In short, treating the Turing test as the basis for a behavioural test of a certain kind of intelligence is valuable because behavioural evidence alone ought to be sufficient to justify the inductive inference that a machine can think and has a certain kind of intelligence. Danaher (2020) makes a similar claim concerning the moral status of machines. He argues that "performative artifice, by itself, can be sufficient to ground a claim of moral status" as long as it is roughly equivalent to "another entity to whom we afford moral status" (Danaher 2020, 2025). Importantly however, in the case of thinking machines, and I cannot stress this enough, this does not mean that one should not desire additional evidence beyond the Turing test (i.e., beyond performative artifice) nor does it mean that evidence obtained from the Turing test is indefeasible. I am simply asserting that the Turing test has value both as the basis for a behavioural test of a certain kind of intelligence and, like Moor, as a source of good inductive evidence to support the claim that a machine has a certain kind of intelligence.

Yet despite its value, my worry is that the Turing test will be confused as *the* test for intelligence in machines and mistaken for the goal, as opposed to what it truly is, a metric for the progress being made on a certain kind of artificial intelligence. On the first confusion, while it might be tempting to assume that the Turing test provides a format for examining a wide variety of activities that would count as evidence for thinking, this is not entirely true. If the first 50 years (1950-2000) of research into

[9] Qualia are those ineffable "raw" experiences that a person has when they, for example, drop a stone on their toe. The pain that a person experiences "feels like" something to them that is unique to their experience.

artificial intelligence have demonstrated anything, it is that the kinds of intelligence that researchers initially suspected would be hard to implement in a machine have actually been relatively easy to implement, and vice versa (i.e., the kinds of intelligence that researchers initially suspected would be easy to implement in a machine have actually been difficult, if not impossible thus far, to implement).[10] In the former category are things like mathematical prowess, logical reasoning skills and board game playing abilities, all of which are kinds of intelligence that have been implemented in machines with varying levels of success to date. One impressive recent example is an artificial system called AlphaZero, developed by DeepMind, which exhibits intelligent board-game playing behaviour. In contrast to a system like Deep Blue, which utilized handcrafted features, expert human knowledge, and adhered to explicitly stated rules to play the game of chess, AlphaZero learned to play chess (in addition to Go and shogi) via millions of games of self-play, without any human knowledge or handcrafted features. Interestingly enough, if there were a "Chess Turing test" both Deep Blue and AlphaZero would likely pass this test. This evidence alone might lead one to make the inference that both Deep Blue and AlphaZero possess chess-playing intelligence. However, this toy example also highlights why I have insisted that one ought to seek additional evidence to support such an inference. Looking under the hood, as it were, at the information processing taking place in Deep Blue and AlphaZero, it would become obvious that Deep Blue is essentially using a sophisticated brute force method to compute the next best move, whereas AlphaZero can contextually evaluate a next best move based on prior experience (Silver et al. 2018). In light of this new information, it might be prudent to revise the inference that Deep Blue possesses chess-playing intelligence.[11]

On the other hand, the kinds of intelligence that researchers initially thought would be easy to implement in machines have actually been far more difficult to implement than suspected. These kinds of intelligence include things like visual image recognition, motor skills, and conversational abilities. Despite Turing's attempt to avoid explicitly defining terms like "machine," "think," "mind" and "intelligence," philosophers, psychologists, computer scientists and many others besides have nevertheless stipulated various definitions. My aim in this section is not to introduce a definitive taxonomy of the kinds of intelligence or distill their essences into a single definition, but rather to highlight that "intelligence" (or perhaps more accurately "human intelligence") is not just one thing. I have made numerous references to the idea that there are

[10] As the late Ronald de Sousa (1991) aptly remarks, "It is a pregnant irony that computers are now relatively good at some of the reasoning tasks that Descartes thought the secure privilege of humans, while they are especially inept at the 'merely animal' functions that he thought could be accounted for mechanically."

[11] Modern descendants of Deep Blue, such as Stockfish, still use a sophisticated brute force approach to play the game of chess. Although these systems play at a superhuman level, they arguably have no chess-playing intelligence because their "knowledge" of the game is explicitly encoded by humans. AlphaZero, in contrast, arguably "knows" how to play the game. AlphaZero is not only significantly better at playing chess than Stockfish (e.g., AlphaZero searches around 60,000 positions per second whereas Stockfish searches around 60 million per second, yet AlphaZero wins far more often) but it learned to play chess without any explicitly encoded human knowledge. See, Silver et al. 2018, for more information.

different kinds of intelligence, one of which is measured by the Turing test, and so in this section I aim to defend that idea.

Consider first that if intelligence was distilled into a single definition, then such a definition would likely be quite ambiguous. For example, in their survey of over 70 different definitions of intelligence, Legg and Hutter recognize that "it is difficult to argue that there is an objective sense in which one definition could be considered to be the correct one" (Legg and Hutter 2007, 9). They nevertheless offer the following informal definition of intelligence: "Intelligence measures an agent's ability to achieve goals in a wide range of environments" (Legg and Hutter 2007, 9). One can see how the Turing test could be subsumed under this general definition, given that it involves an agent attempting to achieve a particular goal in a relatively small range of environments.[12] That is, the format of the Turing test, although it is conversational or question-based in nature, allows for the discussion of virtually any topic the interrogator wishes to discuss. This point serves as a basis for the rejection of Moor's claim that the Turing test provides a format for examining a wide variety of activities, which would count as evidence for thinking or intelligence (Moor 1976). The Turing test really only tests a machine's linguistic and/or conversational intelligence. This is because a machine that is capable of passing the Turing test may not need to possess the different kinds of intelligence that humans also possess.

Machines could also pass the Turing test if shortcuts, as it were, are employed, but such approaches are indicative of Goodhart's Law. Improperly understood, the Turing test amounts to a behavioural challenge to trick and deceive a human interlocutor, and this is largely the form most, if not all, Turing test-like competitions have taken since Turing first proposed the Imitation Game. Machines like ELIZA (Weizenbaum 1966) and PARRY (Colby et al. 1971) that "pass" the Turing test do so via the shortcut of imitating a Rogerian psychotherapist and a paranoid individual respectively. Over short bursts of conversation both ELIZA and PARRY can convince a human interlocutor that they are human, but this does not amount to passing the Turing test. Even before conducting a white box[13] analysis of ELIZA or PARRY by looking at their internal functioning, a serious human interrogator would have no problems identifying that the entity they are conversing with is a machine based on ELIZA's or PARRY's behaviour. Arguably, machines like ELIZA and PARRY tell us far more about how willing humans are to infer, under black box conditions, that an interlocutor has some kind of mind, and much less about how we ought to approach building a thinking machine.

Improperly understood, any number of chatbots could be seen as "passing" the Turing test. However properly understood, the Turing test is not merely about passing or failing. Properly understood, the Turing test is a *game*, the object of which is to time and time again proffer evidence (on the part of the machine) of an ability to hold an intelligible conversation. Short bursts of conversation on a narrow range of topics

[12] Importantly, this definition also excludes certain other agents, like thermostats, from being considered intelligent. A thermostat can only ever achieve a single goal in one particular environment which, at best, might justify the inference that it is minimally intelligent.

[13] Here "white box" is contrasted with "black box." In the former one has access to the internal workings of a system in addition to the inputs and outputs, whereas in the latter one has access only to the inputs and outputs.

should not be sufficient to persuade a human interrogator that their interlocutor possesses conversational intelligence because such conversations generate small amounts of evidence. Multiple rounds of conversation, in contrast, over long periods of time, covering a range of topics discussed at various depths, can generate far more evidence from which it could be inferred by the human interrogator that their interlocutor possesses conversational intelligence. Consider by analogy the kind of intelligence AlphaZero possesses and the way in which it was tested. AlphaZero's superhuman chess-playing intelligence was rigorously tested in one thousand different games against Stockfish (a modern descendent of Deep Blue), in additional games against Stockfish when Stockfish was augmented by a strong opening book, and even in games that started from common human openings (Silver et al. 2018). AlphaZero has no awareness that it is playing a game, that it is a machine, that humans exist, etc. Even so, it is undeniable that AlphaZero possesses some kind of intelligence, i.e., it is able to achieve a goal (win a board game) in a small range of environments (chess, Go and shogi).

Similarly, it is undeniable that in a "Chess Turing test" AlphaZero would be indistinguishable from Stockfish; both play chess at a superhuman level. But when compared to each other and when evidence of their internal differences is available, the proper inference to make is that AlphaZero is closer to a genuinely intelligent machine than Stockfish. To clarify a point raised earlier, it may be prudent in light of this information, i.e., information of the internal operations, to revise the inference that Stockfish (or Deep Blue) is similar to AlphaZero. While Stockfish's significant use of explicitly encoded human knowledge does not disqualify it from being considered a thinking intelligent chess-playing machine (especially if the only evidence available is outward behavioural evidence), AlphaZero's minimal use of explicitly encoded human knowledge (e.g., only the rules of chess had to explicitly encoded) makes it more impressive by comparison. As the researchers at DeepMind note, "common human openings were independently discovered and played frequently by AlphaZero during self-play training," something that certainly cannot be said of Stockfish insofar as its "discoveries" of common human openings were explicitly coded into it (Silver et al. 2018, 1143).[14] Whether conversational intelligence is similar to chess-playing intelligence (or more generally the board-game playing intelligence AlphaZero exhibits), i.e., whether one can imagine a machine passing the Turing test without ever "understanding" that it is being tested, what a human is, the meanings of the words it uses, that it is a machine, etc., is an open question. What is clear, and what I have been insisting on, is that far more evidence would need to be collected from and *should* be collected from various behavioural "Turing test for X kind of intelligence" before one would be justified in making the inference that a machine can think on the level of a human being.

[14] Consider two people tasked with baking a cake. Person A is given a recipe and told to bake the cake according to the recipe, whereas person B is not given a recipe and is told to bake the same cake. Assuming they both succeed, I take it that what person B accomplished was more impressive and required a more sophisticated kind of intelligence in comparison to person A. Importantly, this is not to say that what person A accomplished was not impressive and did not require intelligence. Analogously, the same could be said of Stockfish (A) and AlphaZero (B).

Kinds of Intelligence

To reiterate, the Turing test properly understood is simply one measure of one kind of intelligence.[15] Nowhere is this fact more obvious than in the field of robotics. Humans do not only possess the kind of intelligence that allows them to carry on conversations and answer questions intelligibly; they also possess the kind of intelligence that allows them to move around in their environment intelligibly. This kind of morphological intelligence, which humans and other biological organisms possess, can be thought of as the kind of intelligence that the physical body confers to its owner (Winfield 2017). Note that it is entirely possible that a machine could pass the Turing test and yet utterly fail to possess any morphological intelligence.

Indeed instead of settling on one single general definition of intelligence, Winfield outlines a taxonomy of intelligences: morphological, swarm, individual and social intelligence (Winfield 2017). Morphological and swarm intelligence are the kinds of intelligence that a robot (or robots) might possess, and these could be measured using different behavioural tests (e.g., a "morphological Turing test" involving a robot navigating some kind of obstacle course) capable of generating evidence to justify the inductive inference that a machine can navigate its environment intelligently. AlphaZero is an artificial intelligence that lacks morphological intelligence but possesses individual intelligence given that it has the "ability to both respond (instinctively) to stimuli and, optionally, learn new, – or adapt existing – behaviours through, typically, a process of trial and error" (Winfield 2017, 2). Finally, a machine capable of passing the Turing test would likely possess some combination of individual and social intelligence, the latter understood as "the kind of intelligence that allows animals or robots to learn from each other" through either imitation or instruction (Winfield 2017, 3). The post-upload Will from *Transcendence* is one such machine.

So we return to Turing's original terminology, "imitation" and "game." As Dennett notes, Turing's game is like a well-composed challenge to proponents and critics of artificial intelligence alike; "it seems fair, demanding but possible, and crisply objective in the judging" (Dennett 1981, 93). To imitate human conversational abilities is to possess a certain kind of conversational intelligence and, moreover, presupposes the ability of an agent to learn from interactions with other agents. Moor was not wrong to suggest that the Turing test provides a format for examining a wide variety of activities that would count as evidence for thinking, but there are far more activities that the Turing test cannot examine that would also count as evidence for thinking in machines. The Turing test therefore stands as a valuable source of evidence of thinking in machines, but also one in need of augmentation in recognition of the many different kinds of intelligence that humans possess and that machines may potentially possess.

Conclusion

Most, if not all, of the characters in *Transcendence* eventually accept that post-upload Will has a mind. Though they question whether "he" is the same Will, and doubt his capacity for emotional connection, empathy and altruism, his behaviour clearly indicates that post-upload Will is self-aware, conscious and hyperintelligent. In short,

[15] Or, more generously, the Turing test is one measure of multiple kinds of intelligence, i.e., not a test of all of the different kinds of intelligence that humans possess, but certainly a test for more than one kind of intelligence.

the behavioural evidence provided by post-upload Will through his Turing test-like interactions with other characters in the film is such that we are warranted in making the inductive inference that he has a mind. I have argued that the Turing test is valuable both because it provides a good format for collecting inductive evidence to support the claim that a machine is intelligent, and because it can be interpreted as a behavioural test for a certain kind of intelligence. I maintain that the three common objections to the Turing test do not apply, but that it is valid to criticize the test on the basis of its scope. Further, I have highlighted that emphasizing the Turing test as a target for research in artificial intelligence confuses the test for what it really is: one measure of one kind of intelligence. That being the case, I maintain that the Turing test ought to be augmented given the complex and multi-dimensional nature of intelligence. Ultimately, the Turing test stands as one of the many different tests a machine must pass before it can be inferred that the machine possesses genuine human-level intelligence and, perhaps, self-awareness.

References

Apter, Michael. 1971. *The Computer Simulation of Behaviour*. New York: Harper & Row, Publishers, Inc.

Bieger, Jordi & Thórisson, K. 2018. "Requirements for general intelligence: A case study in trustworthy cumulative learning for air traffic control." *Workshop on Architecture & Evaluation for Generality, Autonomy, & Progress in AI. International Joint Conference on Artificial Intelligence,* 1-11.

Block, Ned. 1981. "Psychologism and behaviorism." *The Philosophical Review, 90* (1), 5-43.

Colby, Kenneth, Weber, S. & Hilf, F. D. 1971. "Artificial paranoia." *Artificial Intelligence, 2*, 1-25.

Crosby, Matthew. 2020. "Building thinking machines by solving animal cognition tasks." *Minds and Machines,* 30, 589-615.

de Sousa, Ronald. 1991. "Does the eye know calculus? The threshold of representation in classical and connectionist models." *International Studies in the Philosophy of Science, 5*(2), 171-185.

Dennett, Daniel. 1981. "Reflections on 'The Turing Test: A coffeehouse conversation.'" In *The Mind's I: Fantasies and Reflections on Self and Soul*, edited by Douglas Hofstadter and Daniel Dennett, 92-95. New York: Basic Books, Inc.

Goodhart, Charles. 1984. *Monetary Theory and Practice: The UK Experience*. London, UK: The MacMillan Press.

Gunderson, Keith. 1964. "The imitation game." *Mind, 73*(290), 234-245.

Hernández-Orallo, José. 2017. “Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement.” *Artificial Intelligence Review, 48*, 397-447.

Hernández-Orallo, José. 2020. “AI evaluation: On broken yardsticks and measurement scales.” *Association for the Advancement of Artificial Intelligence,* 1-7.

Hofstadter, Douglas. 1981. “The Turing Test: A coffeehouse conversation.” In *The Mind’s I: Fantasies and Reflections on Self and Soul*, edited by Douglas Hofstadter and Daniel Dennett, 69-92. New York, New York: Basic Books, Inc.

Jefferson, Geoffrey. 1949. “The mind of mechanical man.” *Lister Oration for the British Medical Journal, 1*, 1105-1121.

Legg, Shane & Hutter, M. 2007. “A collection of definitions of intelligence.” *Frontiers in Artificial Intelligence and Applications, 157*, 17-24.

Moor, James. 1976. “An analysis of the Turing Test.” *Philosophical Studies, 30*(4), 249-257.

Searle, John R. 1984. *Minds, Brains and Science*. Cambridge, USA: Harvard University Press.

Silver, David, Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science, 362*(6419), 1140-1144.

Transcendence. 2014. Dir. Wally Pfister. [Perf. Johnny Depp]. Warner Bros. Pictures. Film.

Turing, Alan. 1950. “Computing machinery and intelligence.” *Mind, 59*(236), 433-460.

Weizenbaum, Joseph. 1966. “ELIZA – A computer program for the study of natural language communication between man and machine.” *Communications of the ACM, 9*(1), 36-45.

Whitby, Blay. 1996. “The Turing Test: AI’s biggest blind alley?” In *Machines and Thought: The Legacy of Alan Turing,* edited by Peter Millican & A. Clark, 53-62. New York, New York: Oxford University Press.

Winfield, Alan. 2017. “How intelligent is your intelligent robot?” *arXiv: 1712.08878v1,* 1-8.