The Turing Test on Trial

Abstract

This essay considers the validity of the standard interpretation of the Turing test as a test of human-like intelligence, defined as the ability to reason, solve problems, think abstractly and learn. Several attempts to write programs that beat the test are described and their performance is discussed in relation to results from competitions like the Loebner prize. The conclusion in this essay is that even though computer programs have passed the Turing test, mostly through cheap tricks, they do not have a human's capacity for intelligent thought, although passing the test may still indicate some degree of intelligence. Finally, an extension to the Turing test is considered, namely the Winograd Schema.

1 Introduction

As the number of people communicating over the Internet is increasing, and at the same time the sophistication of artificial intelligence is rising — the question of who we are communicating with becomes more and more important. One of the most well-known methods of measuring artificial intelligence is the Turing test. The test was first proposed over 65 years ago, but perhaps it is more important today than ever before. In this essay we will explore what we can actually learn from the Turing test. We will also examine whether or not it is sufficient to prove the existence of artificial intelligence, or if it is just a necessary first step in such a proof.

1.1 Defining Intelligence

How to define intelligence is a heavily debated matter, but in order to be able discuss artificial intelligence, one must first decide on a clear definition of what intelligence is. The definition we choose to use in this essay is one from [1]: the capability to reason, solve problems, think abstractly and learn. It is also seen as the ability to comprehend the surroundings and from that draw conclusions and plan ahead.

We can measure the intelligence of humans by various tests such as intelligence quotient tests, Raven's matrices or visualising three-dimensional objects from two-dimensional diagrams [1]. But how do we measure the intelligence of a machine? The computer could probably outperform humans in intelligence tests designed for humans, but these tests could be argued not to measure the full spectrum of human intelligence. Another type of intelligence, emotional intelligence, can be measured by emotional quotient, which is analogous to the intelligence quotient [2]. A computer may not do as well when comparing the EQ to a human as it would if comparing the IQ. So how should we decide whether or not a non-human is intelligent? One suggestion is to examine if the non-human's behavior is remarkably human-like.

1.2 The Turing Test

One of the most influential pioneers in modern computing was Alan Turing. In 1950 he published the paper "Computing Machinery and Intelligence", where he proposes a test of a digital computer's intelligence [3]. There has been much debate over how the test is supposed to be conducted [4]. The two main interpretations of the test are called "The Standard Interpretation" and "The Imitation Game". The Standard Interpretation is a much more liberal reading of Turing in two aspects — it deviates more from what Turing actually wrote in his paper, and it is arguably easier to pass [4]. The Imitation Game is the more literal interpretation of Turings proposed test [4]. We will now explain the two.

In his original text, Turing asks the question "Do machines think?" but quickly dismisses the question to reformulate it as "Are there imaginable digital computers which would do well in the imitation game?" [3]. The Imitation Game is originally a game involving three people — one man, one woman and an interrogator of either sex. The interrogator can only communicate with the other two through typewritten notes, and by asking the subjects questions, he is supposed to determine who is the man and who is the woman. The man's goal is to trick the interrogator into thinking he is the woman, while the woman should try to convince the interrogator that she is in fact the woman. Turing then proposes that, if we were to replace the man with a machine, and the machine would do as well as the man — there is a reason to believe that the computer is intelligent.

In the Standard Interpretation, however, the goal of the interrogator is not to identify the woman, but rather to identify which of two entities spoken to is the computer. The goal of the computer is to trick the interrogator into thinking it is a human, and the human is supposed to convince the interrogator that it is in fact a human. If the interrogator fails to identify the computer, the computer passes the test [4]. The Standard Interpretation could be seen as a more general test, while the Imitation Game is more specific, since it involves the computer imitating a woman, which of course is also a human. In the original paper, Turing stated a prediction:

> I believe that in about fifty years' time it will be possible, to programme computers, [...] to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. [3]

In order to incite development, and make this prediction true, an annual competition "the Loebner Prize" was first organised in 1990 [5]. The prediction has been interpreted as a rule of success in these competitions.

In the following sections we will discuss to which extent the Turing test implies intelligence by first examining a few AIs (chat bots) that have tried to beat the test, and then the conclusions we can draw from the results of testing them.

2 Chat Bots and Testing

Four significant chat bots are covered in this section, named Eliza, Alice, Cleverbot and Eugene Goostman. After the basic operation of these has been described, some results of administering the Turing test on them are discussed — What percentage of judges did they manage to convince of their human-like behavior? What caused people to be fooled?

2.1 Eliza

Eliza is an early and primitive implementation of a natural language processing program, popularly described as a chat bot. The program was developed by Joseph Weizenbaum in the 1960s and its ideas and functionality can be found in several different chat bots developed since then. The program detects if any of a given list of keywords are present in a given input string [6]. If there are, the string is transformed according to a rule associated with the highest-priority keyword. The result is then printed back to the sender. The associated transformation rules are given by a script supplied as a parameter to the program. In this way the script is the "intelligence" which provides the user with an interesting conversation.

One script which was used quite early was called DOCTOR. This script caused conversations that resembled those with an empathic psychiatrist. It had the benefit of being able to respond to statements in an "easy" way that would seem very strange in a normal conversation. In Joseph Weizenbaum's paper he gives the example where the user types "I went for a long boat ride" and the program would respond "Tell me about boats" which would be a perfectly fine question to ask in a conversation with a psychotherapist, but would probably sound very strange in most other conversational settings.

In relation to the Turing test, the author and the creator Joseph Weizenbaum did not set out to beat the Turing test with Eliza. However, in his paper he writes that "Some subjects have been very hard to convince that ELIZA (with its present script) is not human" which he thinks is a "striking" result in relation to Turing's test.

2.2 Alice

A further development of the ideas introduced by Eliza is Alice, ("Artificial linguistic internet computer entity") another program based on the recognition of keywords [7]. It has been enhanced in certain ways, however. First, the number of keywords that it reacts to has been increased greatly. Also importantly, the category for standard responses to be used when the input does not match any pattern, have been improved and added to. Second, the responses are formed by small computer programs which may even be recursive. Third, Alice is able to perform syntax reduction on the input, to streamline pattern matching for patterns with similar meaning. Finally, as her name implies, Alice was trained by internet conversations. A human, the botmaster, would read Alice's responses and improve the response schemas accordingly.

2.3 Cleverbot

Cleverbot is an online chat bot, uploaded to its current website in 2006 [8] but created by Rollo Carpenter as early as 1988. The chat bot primarily exists as a source of entertainment, meaning that people from all over the world can visit the website and talk to the chat bot.

According to the website, the responses that Cleverbot gives are not previously programmed. Not much has been published about the operation of Cleverbot, other than what is said on the webpage: the chat bot learns from its users and the conversations they have previously had in order to give a impression that it is smart. In operation, the chat bot also uses previous statements from its current user, meaning that the classical Hollywood movie line "anything said might be used against you" would also fit in this context.

2.4 Eugene Goostman

Recently, another computer program, Eugene Goostman, has received significant media attention, as it was reported that it passed the Turing test at an event at University of Reading [9][10]. Eugene is a chat bot that portrays a 13-year old Ukrainian boy. The program uses a knowledge base, or a database of general information, looks at the context of an utterance, and tries to estimate the mood of the discussion as rude or polite and tries to match its output to this.

2.5 Testing the Chat Bots

There have been several attempts to show that computer programs are able to pass the Turing test. Often this has been done in the form of competitions or events where many chat bots are tested against each other in a public setting. One example of such a series of competitions is the Loebner Prize.

The Loebner Prize was instituted in order to promote the advancement of AI research. The prize, handed out in competitions based on an interpretation of the Turing test, is 25 000 dollars to the first person to create a computer program that judges cannot distinguish from a human, and smaller prizes for programs that fail the test, but perform better than the others tested. Competitions have been held annually since 1991 [11].

Alice in the Loebner Prize

As an example of how the competitions have been organized, one can consider the 2000 Loebner competition. For this instance, there were six computers and four persons, or in total ten "entities", acting as conversation partners for a set of judges. The entities were only observed by the judges through the text that they produced as replies to the judges' questions. After five and after fifteen minutes, the judges were asked to decide if they had interacted with a human or a computer. Each entity was also rated by each judge on a ten-point scale on their estimated humanness, with 10 being the most human [12].

In 2000, Alice was deemed the best of the computer programs, however still not ranked as high as any of the humans on the scale of humanness [12]. None of the chat bots were taken for a human, but in some cases a human was thought to be a computer by some of the judges. The mean scores for the computers ranged from 0.56 to 3.75, whereas the humans were rated from 7.8 to 9.35. Alice's score on humanness was 2.35.

There was also an additional expert judge present at the 2000 competition [12]. He asked questions to test the capacity of the computer programs for understanding, reasoning and learning, such as "What is the color of a blue truck?", "What letter does the letter 'M' look like when turned upside down?" and "PLEASE IMITATE MY TYPING STYLE". None of the chat bots were able to correctly answer any of these prompts, whereas no human failed on any of them.

What caused some people to be more or less fooled by the programs? There are several tricks used by the chat bots that seem to be part of the explanation [12]. For instance, chat bots tried to cause people to talk about themselves. This ELIZA-inspired, therapist-like behavior had the advantages that the bot said little that it could later contradict, and also that people like to talk about themselves, which may have improved their opinions of the bot.

Testing Cleverbot

In 2011 at the Techniche festival in Guwahati, India, Cleverbot took part in a formal Turing test [13]. The test involved 30 volunteers chatting on three screens with either a chat bot or a human (which of these, they did not know) for 10 rounds of 4 minutes each, all with an audience watching. The audience then got to vote on a scale of 0 to 10 how human the conversations were, not knowing that half of these conversations were human-to-human. The cleverbot achieved a score of 59.3% out of the maximum while the humans got a score of 63.3%, all from a total of 1334 votes cast.

Testing Goostman

In 2008, Eugene Goostman was one of the tested programs in the Loebner competition [14]. It did not win, but it was noted that it was able to remember and use information given to it. For instance, one judge told the program that his car was red and then asked the program what his car's color was, which the program was able to correctly answer. Also, Eugene received high ranks for its conversational ability in part because it was considered good at discussing music preferences.

In 2012, Eugene Goostman was again tested, this time in Bletchley Park, UK, an event described in an article by Warwick and Shah [15]. A procedure like the one used in the Loebner competition was used, with interrogators speaking to computers and humans via computer terminals. Here, Goostman managed to convince 28 percent of the judges that it was a human. Another example is a 2014 event, held at the University of Reading [16]. There, judges held simultaneous conversations with humans and computers for five minutes. There were 30 judges and in total 300 conversations. Goostman was judged as human in one third of the conversations it participated in, meaning that it managed to pass the Turing test in the sense that it fooled more than 30 percent of the participants into believing it was a human.

In considering some of the transcripts, there was one recurring factor that appears to cause several of the judges to misclassify Goostman. The machine succeeded in steering the conversation to topics that it is familiar with. Several times it was possible for the program to completely ignore the content of a question asked to it, and just ask another question about another topic back, and still be judged as a human. Warwick and Shah made the comment that "humans are easily distracted and allow the subject to be changed directly" [15].

3 Discussion

In retrospect, the Turing test has led to the development of quite a few seemingly intelligent programs ranging from Eliza to Goostman. These programs have improved with time, supposedly to the point of passing the Turing test. But do they really pass? What can we conclude about the intelligence of the bots? What can we infer about the test itself? Is intelligence testing based on use of language even a valid paradigm?

3.1 Passing the Turing Test or Not

Recall the Standard Interpretation of the Turing test where, for the computer to pass the test, the judges should not be able to correctly classify a computer as a computer more than 70% of the time. Since the various tests and competitions that have been held have been performed in different ways, it is somewhat difficult to unequivocally say that the tests show that a program passed the Turing test.

For instance, the Cleverbot and the later Eugene Goostman underwent tests and could be considered to have passed. The Cleverbot achieved a score of 59.3% "maximum humanness" at a test, which could be seen as pass, if it had not been the case that the way of determining "degree of humanity" was quite dissimilar to the requirements laid out in Standard Interpretation. This is also quite different compared to the event in 2014 at University of Reading where Eugene was tested and thought to be human by a third of the judges. In this test the judges only had two options — to decide whether or not the subject was a computer. The Standard Interpretation states nothing of this, which therefore brings an uncertainty to the question, "Did the chat bots pass the Turing test?". An argument could certainly be made for either answer. Regardless, all of the chat bots in this essay have tricked some people into thinking that the chat bot is in fact a human. The difference is the fraction or number of people tricked by the bots.

3.2 Passing the Turing Test may not Imply Intelligence

But does tricking a number of people really show that the computer possesses the same capacity for learning, reasoning and abstract thinking as a human does? In the Loebner price competition of year 2000 where an expert judge asked questions to try to see if the the chat bot showed signs of human abilities, all of the chat bots failed, but still tricked some of the other judges in the Turing test.

Since intelligence is a multifaceted quality, there does not exist one universal test for measuring the intelligence of humans, but rather there is a multitude of different tests. The Turing test is only one test. It seems like wishful thinking or presumption to think that it alone could provide enough basis for measuring every aspect of artificial intelligence. Different chat bots do seem to excel at different aspects of intelligence. For instance, Cleverbot shows signs of the human aspect of learning, since it continuously learns from its previous conversation with other humans, while Goostman shows the ability to remember facts from previously in a discussion and use these in later responses, a basic form of reasoning and learning.

In addition, tests of human intelligence do not give a binary response to whether or not a human is intelligent or non-intelligent — but rather they indicate where on a scale of intelligence a person rates. Perhaps artificial intelligence can only be measured in this fashion as well. If a chat bot does better in the Turing test than another chat bot — maybe it cannot be concluded that one is intelligent and the other is not, but instead it can be inferred that one is more intelligent than the other.

3.3 The Role of Language in Intelligence

Since the Turing test is based on conversing with a machine in a human language, some questions arise. Is conversing the best way to test intelligence? Or are there any alternatives? This query may have solution as follows.

How do we interact with humans to know that they are intelligent? It is obviously through language, which is acknowledged by many renowned philosophers such as J.H Moor, P Schweizer and S Watt [17]. Although most of the world is convinced that language captures intelligence, Donald Michie, a famous AI researcher argues that language does not capture sub articulate thought: a type of cognitive operations we are not consciously aware of [18]. Michie expands on this thought by defending with the example that it is impossible for a programmed machine to find the plurals of words "platch", "snorp" and "brella" as there are no explicit rules written down to find the plural of these words, whereas human would guess them as "platchez", "snorps", and "brellz" from their perception. But the argument of Michie can be refuted if given a large language corpus a machine learning algorithm would discover countless examples of words and their plurals, and recognise the phonetic rules for pluralising. Thus proving the method of using language to test intelligence is a very widely accepted way to test intelligence.

3.4 Shortcomings of and Extensions to the Turing Test

Prior to the discussion on the shortcomings of the Turing test, it is good to further analyse the basis of the Turing test, that machines can display intelligence, which Turing defends in the later part of his paper. A renowned philosopher, Searle, holds contrary views to those of Turing [19]. Searle's argument is that only machines with internal causal powers equivalent to those of brains can understand and think and no program by itself is sufficient for thinking. Searle tries to prove this by imagining a situation of locking

up a person who knows nothing about Chinese in a room. Then he is given a story written in Chinese, questions he needs to answer about the story and a set of rules in English to decrypt the Chinese symbols, sufficient for him to answer the questions. In this situation, the person is acting like an intelligent agent (a computer) and the Chinese script and the questions are inputs to the agent and the English rules is the program. So in order for the person to answer the questions he needs not understand Chinese, he just needs to use the rules and correlate them to the Chinese symbols and give the answers out. Searly thus tries to prove that the agent has no understanding, it is just performing computations on the Chinese characters and giving out the answers. Searle, in his paper, then states a lot of other examples to prove his standing and concludes by answering the questions

"Can machines think?"

"Yes, a machine with a nervous system, neurons with axons and dendrites, and all the rest of it, sufficiently like ours can."

"But could something think, understand, and so on solely in virtue of being a computer with the right sort of program?"

"No, because the formal symbol manipulations by themselves don't have any intentionality". Thus Searle analyses the question "Can machines think?" more deeply and holds very strong views contrary to the Turing on one of the bases of the Turing test.

From the points discussed so far, the Turing test has its shortcomings. The imitation game on which the entire Turing test is based has limitations and needs consideration and evaluation. An appropriate program to take the Turing test would be a chat bot, which is a script made to mimic human conversation. Then arises the question whether these chat bots are the representation of intelligence. Also, these chat bots rely a lot on techniques such as tricking and fooling humans, although it's fun to think about, creating a chat bot that can fool humans but it is not really the same thing as creating artificial intelligence.

Considering the human participants in the imitation game, if participants falls to the tricky questions of the interrogator, then the interrogator may assume the human to be a machine and thereby increasing the chances of the machine to beat the Turing Test.

In the Loebner Prize contest held in 1991, a human participant showed extraordinary knowledge about Shakespeare, and the interrogator wrongly concluded that the human is a machine. This resulted in increasing the probability of a machine to be judged as a human[20]. With these shortcomings, Turings test may just be a necessary but may not be a sufficient test of intelligence.

Due to the tricks used by the chat bots, and the humans falling to the tricks of the interrogator — the Winograd schema has been proposed as an extension to the Turing test. It involves the subjects responding to a broad range of English questions, which any native English-speaking adults can easily answer correctly, and it can be administered and graded without expert judges [21].

For example, questions of the following format are posed in the test: "The trophy doesn't fit in the brown suitcase because it's too big. What is too big?", where the answers can be either "the trophy" or "the suitcase".

In order to answer such questions, the program needs to understand the subtext that may not expressed in the words of the sentence and be able to sort out what is going on. However, the author points towards cautious construction of questions, since the answers to certain questions are very obvious and can be picked up even by weak algorithms. Thus, the Winograd schema, an extension of the Turing test, can be a good test to overcome the shortcomings of the Turing test.

4 Conclusion

In this essay we have presented the Turing test, cases when it has been used and discussions about its strengths and shortcomings. We feel fairly confident in concluding that the Turing test is not sufficient to prove that a computer is as intelligent as a human, even though the test does indicate some basic level of intelligence. Furthermore, with extensions like the Winograd schema it could be made into an even better indicator. In conclusion, we could express passing the Turing test as a necessary but insufficient condition for proving human-like artificial intelligence.

References

[1] I. J. Deary and P. G. Caryl, "Neuroscience and human intelligence differences," *Trends in Neurosciences*, vol. 20, no. 8, pp. 365–371, 1997.

[2] A. M. Colman and O. R. (.-b. collection), *A dictionary of psychology*, Fourth. Oxford: Oxford University Press, 2015.

[3] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[4] S. Traiger, "Making the right identification in the turing test[1]," *Minds and Machines*, vol. 10, no. 4, pp. 561–572, 2000.

[5] K. Frankish and W. M. Ramsey, *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, 2014.

[6] J. Weizenbaum, "Eliza - a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966, ISSN: 0001-0782. DOI: 10.1145/365153.365168. [Online]. Available: http://doi.acm.org/10.1145/365153.365168.

[7] R. S. Wallace, *The anatomy of ALICE*. Springer, 2009.

[8] R. Carpenter. (2006). Cleverbot, [Online]. Available: www.cleverbot.com (visited on 04/20/2016).

[9] I. Sample and A. Hern, *Scientists dispute whether computer eugene goostman' passed turing test*, theguardiantheguardian.com, Ed., [Online; posted 9-June-2014], Jun. 2014. [Online]. Available: http://www.surfline.com/surf-news/maldives-surf-access-controversy-update_75296/.

[10] L. Ulanoff, *The life and times of 'eugene goostman,' who passed the turing test*, mashable.com, Ed., [Online; posted 12-June-2014], Jun. 2014. [Online]. Available: http://mashable.com/2014/06/12/eugene-goostman-turing-test/.

[11] M. L. Mauldin, "Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition," in *AAAI*, vol. 94, 1994, pp. 16–21.

[12] J. H. Moor, "The status and future of the turing test," *Minds and Machines*, vol. 11, no. 1, pp. 77–93, 2001.

[13] R. Carpenter. (2011). Cleverbot comes very close to passing the turing test, [Online]. Available: www.cleverbot.com/human (visited on 04/20/2016).

[14] M. Bishop, H. Shah, and K. Warwick, "Testing turing's five minutes, parallel-paired imitation game," *Kybernetes*, vol. 39, no. 3, pp. 449–465, 2010.

[15] K. Warwick and H. Shah, "Good machine performance in turing's imitation game," *Computational Intelligence and AI in Games, IEEE Transactions on*, vol. 6, no. 3, pp. 289–299, 2014.

[16] reading.ac.uk, Ed., *Turing test success marks milestone in computing history*, [Online; posted 8-June-2014], Jun. 2014. [Online]. Available: http://www.reading.ac.uk/news-and-events/releases/PR583836.aspx.

[17] K. LaCurts, "Criticisms of the turing test and why you should ignore (most of) them," *Official Blog of MITs Course: Philosophy and Theoretical Computer Science*, 2011.

[18] D. Michie, "Turing's test and conscious thought," *Artificial Intelligence*, vol. 60, no. 1, pp. 1–22, 1993.

[19] J. R. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–424, 1980.

[20] S. M. Shieber, "Lessons from a restricted turing test," *ArXiv preprint cmp-lg/9404002*, 1994.

[21] H. J. Levesque, E. Davis, and L. Morgenstern, "The winograd schema challenge.," in *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.