The Turing Test tested

Marjolein de Vries, Louis van der Burg, Sophie Horsman, Mayaan Shvo
student numbers: 111111111, 5981271, 111111111, 111111111

Abstract

to be done **Keywords:** Turing; Turing test; Spell errors; Validity; Relevance.

Introduction

We live in a world in which every year technology gets smarter and robots capable of more human tasks. Therefore, as Artificial Intelligence researchers, we often get asked (by both academic and non-academics) what this will entail for the future. Will the differences between humans and machines diminish? Will a machine (or a robot if you like) ever be able to truly think like a human being? The first real attempt of a practical way to test whether machines can think, came from computer scientist Alan Turing in 1936. He proposed a test in which one person (the so-called interrogator) talks simultaneously with two others, out of which one is a computer and the other one is a human being. When the interrogator makes the wrong identification at least half of the times, the computer passes the test and the machine is said to think.

80 years after Turings original ideas we would like to evaluate if and how a Turing Test could be used in future Artificial Intelligence research. Obviously, there have been many objections against the test. A lot of these objections are deeply philosophical that question whether the test measures machine intelligence at all. One example of a philosophical objection is the simulation objection. It says that success in the test only shows that a computer can give a good simulation of thinking, while it is not actually thinking. This raises philosophical issues such as what thinking actually means and whether computer simulations of thinking will always stay mere simulations or whether they can ever become the real thing (Copeland, 2015).

Up until now, no agreement has been reached and there is no way that any test for machine intelligence can solve this simulation objection. We will therefore firstly focus on two more practical objections that question whether the Turing Test could be practically used in science. In the light of these objections, we will discuss several alterations on the tests proposed by A.I. researchers. Finally, we will test some alterations to the Turing Test in practice by executing the Turing Test with a small manipulation ourselves. This experiment will provide some insights into the performance of the Turing Test. Moreover, the validity of the Turing Test will be discussed based on the outcomes of the experiment.

The research question of this paper is at the core of A.I. and therefore cannot be addressed by disciplinary viewpoints from either psychology, philosophy or computer science alone. Instead, it should be addressed with an interdisciplinary approach. Critical analyzing that is being done in philosophy helps understanding the central concepts and problems. Setting up an experiment and analysing the data using statistical methods needs to be learned from psychology. Finally, manipulating an experiment uses skills and tools from computer science. Therefore, we conclude that in answering our research question we need an integrative perspective that combines those three mentioned disciplines. To evaluate the validity of the Turing Test is both of theoretical and practical importance. It is practically important because it helps us in setting appropriate goals for A.I. research. It is theoretically important since a negative outcome (that passing a Turing Test should not be an appropriate goal in A.I.) has the theoretical implication that it might not be relevant to keep discussing it.

Turing Test

In 1950, Alan Turing elaborated his ideas about the Turing Test proposed an early version of the Turing Test in a paper titled Computing Machinery and Intelligence (Turing, 1950). The initial question proposed by Turing was whether machines can think. Turing argued that rather than discussing the definition of the words machine and think, the question can be answered by executing an experiment. The original experiment, called the Imitation Game, consists of a man, a machine, and an interrogator. The interrogator is located in a different room than the man and the machine, and his goal is to identify which of the two others is a man and which is a woman. The interrogator can do so by asking questions via some chat application to both the man and machine simultaneously. The goal of the man is trying to cause the interrogator to make a wrong identification, thus that the interrogator would mistake him for being a woman. The machine, which is clearly neither male nor female, has as goal to resemble a woman in order to cause the interrogator to classify it as a woman. If the interrogator decides wrongly as often as when the game is played with a woman in place of the machine, Turing would say that the machine is intelligent.

Over the years, this original experiment has been slightly changed into what we nowadays call the Turing Test. In the contemporary version of the Turing Test, the interrogator does not have to decide which of the two is the man and which is the woman, but rather which of the two is the human and which is the machine. If we accept the validity of the test, we would call a machine intelligent when the interrogator decides wrongly as often as correct when the experiment is executed multiple times.

Objections to the original Turing Test

(Oppy & Dowe, 2016; ?, ?)

The first objection is that the Turing Test is too hard for a computer to pass. Right now, many people think that we are so far away from passing the test that the goal of passing it might not be a realistic goal in A.I. research. Some people even think that no machine that man creates will ever pass the test.

One of the reasons that the test is too hard was given by Robert M. French who thinks that nothing without a human subcognitive substrate could ever pass the test. According to French, there are obvious questions that people can use to discriminate between humans and computers, which reveal in his words low-level cognitive structure. By low-level cognitive structure he means the subconscious associative network in human minds. Humans develop many associations during their lives: they learn through experience that certain words or concepts are more commonly used together with second words or concepts than others, for example the words bread and butter in comparison with bread and dog. The interrogator in the Turing Test can make use of this associative network that he shares with other humans while he doesnt share this (at least not extensively) with the computer.

French expand these ideas and introduced rating games which are intentionally designed to be able to distinguish between humans and computers. One of his examples is that an interrogator would ask the participants to rate the name flugbots as an appropriate name for a breakfast cereal. The human participant in the game would

This relates to another reason that the test is too hard, which is that it does not only measure intelligence, but also how humanlike the machine is. It might be so that it is particularly hard to simulate certain human features, which have nothing or little to do with intelligence.

The second objection is not aimed at the difficulty of the test, but at the scope of the test, as it states that the Turing Test is too narrow for testing intelligence. The most known argument for this objection is given by John Searle (Searle, 1980), which we will discuss first. After Searles famous objection, we will discuss arguments supporting this objection made by other authors.

Searles main argument is described in his paper called Mind, brains, and programs (Searle, 1980) and describes the Chinese Room argument. The argument goes as follows. Suppose we have a computer program called Sam which is able to answer Chinese questions back in Chinese. We now replace this program Sam by a human Joe, and we envision this situation as if Joe is sitting in a room detached from the rest of the world. Joe is equipped with instruction books on the program instructions, thus the books describe which character to add in Joes answer when a specific character is present in the question. Joe is American and he does not understand anything from Chinese. Therefore, Joe has no idea what the questions or answers mean. However, to an external observer, it looks like that the program is intelligent as it gives reasonable answers to the questions.

The claim Searle wants to make with this thought experiment is that executing a program does not imply any understanding of what the program is doing or attaching meaning to the Chinese symbols which are used. A computer program uses symbols (zeros and ones) in a similar way, thus is also not capable of understanding these symbols or of attaching meaning to the symbols. Because the Turing Test only uses words, which can be interpreted as symbols, the test does not test real understanding.

Searles objection is related to the Sense organs objection given by Copeland (Copeland, 2015). This objections concerns the fact that the test is only aimed at question answering, and it does not test whether the computer can relate the used words to things in the world. According to this objection, a computer could thus pass the test without having any understanding of the words he is using. An alteration for the Turing Test is given, and argues that the test should be strengthened by providing the computer artificial sense organs such as vision and speech, which can be used to test the computers understanding of the meaning of the words.

The question then remains whether giving a computer artificial sense organs could solve Searles objection. We can argue that the computer now not only works with letters and words, which are coded as zeros and ones in the computer, so more thorough understanding is now possible. On the other hand, we can argue that images obtained by vision or sounds that the computer hears or produces can also be encoded into zeros and ones. Thus, images and sounds are in the end also symbols and adding those sense organs to the Turing Test would in no way resolve Searles objection.

From a less analytical and more practical point of view, Gerald J. Erion (Shieber, 2007) also argues that while computers might be able to pass the test, they can still not do much else than the limited tasks involved in passing the test. During the test, computers are only answering questions via some chat interface, which is a very limited task. We question whether his statement is correct. When passing the test, the computer must be able to solve a wide variety of every circumstances, related to common knowledge, memory, personal identity, and many more. Although the Turing Test is text-only, it thus requires the computer to do many subtasks which contribute to the computers credibility of being a human.

Alterations to the test

In order to sustain the validity/credibility of the Turing Test, it is advisable to alter the test in some ways. In this section, we will discuss two of those alterations which we will both use in our experiment as well.

Probabilistic Support

In the original Turing Test an interrogator has a conversation with both the human participant and the computer and has to give a yes or no answer to the question that (lets say) participant 1 was a computer. But given the first objection, it is very hard for a machine to fool one human into believing it

is itself a human, let alone fool the majority of people. How could we measure whether the performance of a computer is getting better or not?

Obviously, Turings original test can measure the difference in performance between different digital? participants. When the first computer participant can fool 6 out of 20 people it does better than the second one which can only fool 3 out of 20. But A.I. researchers have suggested that it is easy to alter the Turing Test in such a way that it yields more fine-grained statistical data (Shieber, 2007). To do this, the altered test should ask participants to provide probabilistic answers, such as: I am 75% sure that participant 1 is a computer (and therefore I would give a 25% chance that participant 2 turns out to be a computer). While it might be difficult (or even impossible) to pass the test, we can now at least compare the differences in performances between different digital participants.

This alteration does not only provide scientists with more elaborated statistical data, it also opens up the possibility for an experiment to use a within subject design. In the original Turing Test it would be only possible to do an experiment to use a between subjects design, since the participants would only give a yes or no answer.

Introducing Spelling Errors

The Turing Test requires computers to be as human as possible. In order to pass the test, the computer should exhibit human-like behavior and thus should make human-like errors such as not having a perfect memory (not knowing which day of the week it was 8 years ago), not doing complex calculations very fast, and making errors while typing (Epstein, Roberts, & Beber, 2009). Not only is it the question whether this actually has anything to do with being intelligent, it also makes it hard for the computer which has to take all these subtle human trans logical reasoning errors into account when constructing an answer in order to pass as a human.

Because these trans logical errors have nothing to do with intelligence in general, it may be a good idea to help the computer with that in order to let it have a fair chance of passing the test. In an altered version of the Turing Test, the computer could thus get equipped with some software which automatically prevents the computer of exhibiting non-human behavior such as giving an answer which requires extensive memory (which day of the week was it 8 years ago) or giving an answer to a very difficult mathematical question (which no human can possibly give). Moreover, the software could help the computer by automatically making spelling errors which occur frequently with humans, and by automatically delaying the computers answer (according to some distribution). There are of course also more elaborate functions which can be equipped in the software.

From Literature to Experiment

One of such functions of the software, namely automatically adding spelling errors to the computers answer, will be evaluated in our experiment. By adding this manipulation, we are going to investigate whether we can modify the Turing Test in favour of the computer. In this way, we will also try to mitigate the first objection which states that the test is too hard for a computer to pass. Moreover, in line with the first alteration, we will also measure the participants belief which of the two conversation partners is a computer on a probability scale. Therefore, we measure the outcome on a ratio scale instead of on a simple nominal scale.

When designing or executing a Turing Test, the output criterion (Copeland, 2015) is important to take into account. The output criterion states that the interrogator should talk to the human and machine simultaneously, in order to be able to compare the conversational output of them both when assessing which is the human and which is the machine. Therefore, we make sure to satisfy this criterion in our execution of the test.

The experiment

The main question we would like to answer during the experiment is: *Does the introducion of spelling errors make the chatbot come across as more human?*

The experiment will now be described in detail.

Participants

The participants were 20 students of Utrecht University who volunteered. Participants ranged in age from 18 to 26, with a mean age of 21.65. Of the participants 50% were male and 50% female. The majority (40%) had not previously heard of the Turing Test, the minority (25%) had prior knowledge of the Turing Test and the rest (35%) had heard of the test but not in detail.

Procedure

The participants were placed before a computer with two chat windows open, with Person 1 and with Person 2. The participants were then told that they had to converse by chat interface with two people and they had to choose which one of

the people was a real person and which one was a computer. There was a time limit set at three minutes, but they could stop at any moment before that when they had made a decision. They were prohibited from using emoticons and they could only ask one question per time. The participants each did two rounds of conversations in total.

Person 1 was always the chatbot and Person 2 was always the same real person. The human controlling the chatbot followed a schedule, see figure 1, so that the first ten participants talked to the chatbot Mike[1] and the last ten participants talked to the chatbot Rose[2]. The schedule also indicated that each person had one conversation with the chatbot without added spelling errors and the other conversation had introduced spelling errors by a script. The script used the most common misspellings done by humans(citation needed) introduced according to the length of the sentence and a chance

[1] http://www.eslfast.com/robot/english_tutor.htm

[2] http://brilligunderstanding.com/rosedemo.html

variable. The maximum amount of errors is calculated as 10% of the amount of words in the sentence rounded down. And then there is a chance of 1/3 to introduce a spell error. The algorithm checks the words in the sentence given by the chatbot against a library of the most commonly misspelled words. When a match is found it replaces the original word with the misspelled word in the sentence. After the two rounds of conversations the participants proceeded to fill in the questionnaire, as described in section ??.

Table 1: Distribution of conversations

Chatbot	Control	With spelling errors
Rose	10	10
Mike	10	10

Results

To test our main hypothesis, namely that the introduction of generated spelling errors would allow a chatbot to perform better in a Turing test, a Wilcoxon signed-rank test has been performed, yet the result was not significant ($Z = -1.14$, $p < 0.05$).

From an analysis of the post-experiment questionnaire we learn that 40% of participants have never heard of the Turing Test. 35% of participants had a vague notion of what the test is and 25% of participants knew exactly what the Turing Test is.

It is worth noting that we have found a significant positive correlation between prior knowledge of the Turing Test and the participants confidence in identifying the computer as the computer ($r(40) = .39$, $p < 0.05$).

Discussion

In this paper, we tested the hypothesis that the introduction of generated spelling errors would allow a chatbot to perform better in a Turing test. Contrary to our hypothesis, our results did not yield a significant effect. The aim of this discussion is to offer a number of methodological and theoretical arguments, as to why the results did not match our hypothesis. First, the spelling errors were automatically generated, and the amount of errors was set relative to the length of the chatbots response. We argue that there is a fine line between introducing too many spelling errors, and too few. In the former case, the chatbots response might be perceived as unreliable, while in the latter case, no effect will come of introducing the spelling errors, as they are too few to notice. As part of future work, pre-tests should be run so that an optimal amount of introduced spelling errors might be found. Furthermore, the sample size in our experiment was rather small, and future experiments would benefit from a larger set of participants.

As part of the data analysis, we have found a statistically significant positive correlation between prior knowledge of the Turing Test and the participants confidence in identifying the computer as the computer. It seems that participants who had prior knowledge of the test, were better equipped to test the chatbots, for example by asking questions that more effectively single out the computer; another example is that those participants were not surprised by the chatbot's ability to parse informal language. For future work, considering the participants knowledge of the test might prove important when designing the experiment and analyzing the results.

We learn from the participants responses in the questionnaire, that a number of participants did find the chatbots more human when spelling errors were introduced (Person 2 made a typo [thats why I think hes the human]). Conversely, some participants found the computers perfectly punctuated responses to give away the computers identity (the wording of person 2's answers and the perfect punctuation made me feel like they were a computer). A recurring theme found in the participants responses, is that they found their questions ill understood by the chatbots (person 2 did not answer the last question correctly) or the replies given to them, odd (Person 2 was the computer because he did not communicate in a logical manner. His answers were quite weird) or out of place (I don't think someone would respond "I'm happy to know you're doing fine" in a casual chat conversation..). Further, it is evident by these responses that the chatbots behaviour heavily influenced their attempt to pass as humans as part of the test.

To conclude, in this experiment we have set out to investigate whether we can modify the Turing Test in favour of the computer, by manipulating the computers responses via the introduction of spelling errors. In some cases it shows that adding spelling errors can have a benefit for the computer, but in general the hypothesis cannot be tested yet as the chatbots are still too primitive to give correct conversational responses.

In order to zoom out and return to the larger picture, we would like to understand the implications of our findings on the main goal of the current research, namely to evaluate if and how a Turing Test could be used in future Artificial Intelligence research.

Objections Variations ...

We can say that passing the turing test is not yet viable, but investigating the test and how to make computers pass it is a good way of learning more about human language interaction.

Appendix

References

Copeland, J. (2015). *Artificial intelligence: A philosophical introduction*. John Wiley & Sons.

Epstein, R., Roberts, G., & Beber, G. (2009). *Parsing the turing test*. Springer.

Oppy, G., & Dowe, D. (2016). The turing test. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2016 ed.). http://plato.stanford.edu/archives/spr2016/entries/turing-test/

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(03), 417–424.

Shieber, S. M. (2007). The turing test as interactive proof. *Noûs*, *41*(4), 686–713.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.