Structural issues of the Turing test

A survey on the structural issues concerning two different interpretations of the Turing test

FABIO BALLABIO
Philosophical issues of computer science, Politecnico di Milano
July 23, 2018

Abstract

This essay explores the structural issues of the Turing test under two different interpretations, the Original one, also called Imitation game, and the Standard one, which is the one usually intended speaking about Turing test. Looking at some structural problems in both the two frameworks of the Turing test we want to show that none of the two, for different reasons, is a good replacement for the question "Can machines think?" lacking to capture the very essence of what human beings would identify as intelligence, introducing expedients and game mechanisms only useful within the test itself.

Introduction

In attempting to answer the question "Can machines think?", Alan Turing had to face the problem of defining "think". Trying to avoid it the question was turned into a game, presented in his paper "Computing machinery and intelligence" [8]. Many philosophers and scientist noticed ambiguities in the description of the game meant to substitute the question "Can machines think?" leading to two different interpretations of what Turing proposed in 1950, as presented in [7]. Taking behaviorism and so "whatever acts sufficiently intelligent is intelligent" [3], which is the backbone of the Turing's idea, as a proper foundation, I will examine how some structural problems of the game, under its two interpretations, affect the notion of intelligence. First, I will start introducing the test and what was the aim of Turing in proposing it, than I will identify clearly the two versions defined in [7] by Susan Sterrett.

Just to avoid confusion, along this essay I will use as synonyms the words test and game, version and interpretation.

Turing's aim and Turing test

Before discussing any criticism against the Turing test let me introduce the test itself, its aim and its two different interpretations.

How could researchers tell if a machine was capable of thought? The question "Can machines think" was directly unaddressable due to the impossibility of providing a comprehensive definition of "think". The genius of Alan Turing was in not being stuck in the question making a step further replacing it with an equivalent practical test, a game involving humans and machines, that separates the physical and intellectual capacities of humans, allowing to determine if a machine is capable of thought. The original game was the so called Imitation Game and works as described by Turing himself in [8]:

> *" The new form of the problem can be described in terms of a game which we call the imitation game. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A." [...] "We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think?" "*

Was a valid approach substituting the original question with a game? I think it was not just valid but probably the only one possible. We believe behavioral evidence is the best evidence we have for others intelligence and that's in fact the way we award intelligence to other human beings. We can't have any clue about what is going on in others mind, we simply assume that also other humans are intelligent just because they looks like us, even without any kind of interaction, just by looking. Among humans this approach works pretty well since all the human beings are quite similar systems. We haven't any difficulty in state that other humans are capable of thought just by inference, but the more a system differ from a human the more is difficult to make this inferential step. That's why a complex structure, as the one proposed by Turing, is needed to accomplish the task of granting intelligence in a human-like way when machines are involved. As a result of these considerations, Turing proposed the only reasonable solution. Having it stated clearly, my objections through this essay concerns only the specific game proposed which implies many features that have nothing to do with intelligence, hiding it.

The two interpretations of the test proposed by Turing

After a first reading of Turing's paper the game presented seems clear, but on a more literal reading, as pointed out by Susan Sterrett in [7], it results ambiguous in many aspects, leading not to one, but two such tests, which are not equivalent as they can appear. Trying to keep consistency in naming them with the work of Susan Sterrett, I will call "Original Turing test" or "Imitation game" the first version of the game proposed by Alan Turing immediately at the beginning of his paper "Computing machinery and intelligence", and "Standard Turing test" the second version which is the one that emerges going through it.

The original Turing test is the one described in the previous section, there are three players, A the man, B the woman and C the interrogator. Each one of the players is in a different room and C can communicate with A and B only by means of a terminal. C's objective is to determine who is the man and who is the woman, A's objective is to fool C in believing he's the woman while B's objective is to help C to achieve the right gender identification. Given that the question posed by Turing is: "What will happen when a machine takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" [8].

The standard Turing test is, instead, a game in which the man and the machine compete directly. The question becomes: "Let us fix our attention on one particular digital computer. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate program, it can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man?" [7]. Here A is the machine, B is the man and still C is the interrogator. C's objective is to identify who is the man and who is the machine, A's objective is to fool C in believing it's the man, and B's objective is to help C to achieve the right identification (notice that the concept of gender is lost here).

Main differences between the two interpretations

While the Original and the Standard tests may seem quite similar they set rather different questions as a replacement for "Can machines think?". The two main differences I want to highlight here are about the fairness of the games and the interpretation of the results.

The Original game is fairer assigning the same task to the man and the machine, trying to screen out the man's advantage of being human as much as possible. The role of impersonation is crucial since it requires a critical approach, it allows to "de-emphasize training and emphasize thinking" [7], both the man and the machine have to critically evaluate and reflect on their answers, none of the two has lived a life a woman has, so none of the two can rely upon his own cognitive habits. In the standard game, instead, the only entity who has to impersonate something is the machine, which is required to behave like a man, while the man doesn't have to do nothing but being himself. In this second version the game is quite unfair since the man can rely on his whole background of having lived a human life.

For what concern the outcomes, we have to ask ourselves what does it mean to pass the Original Turing test, and what does it mean to pass the Standard one. In the original formulation we compare statistically how well the task of cross-gender impersonation is performed by the machine with respect to the man. Both the man and the machine play the same game, with the same objective of imitating a woman to mislead the interrogator, so they can be compared each other. We have a performance measure (from the machine) and a benchmark (the man's score), the test is passed if the machine scores higher than the benchmark. In the standard formulation the machine and the man are directly competing in an unfair game. The kind of measurements we can get from it are based on how often the interrogator takes the machine as human, but we miss a target. Which kind of result, numerically, represents a sufficient score to pass the test? May be that if a machine is taken by the man at least 50% of the times then the test is passed. Is it a good guess? Probably not, Susan Sterrett, arguing about the Standard interpretation, says: "given a skilled and determined interrogator only a human could pass the test" [7], that's why we are more likely to measure the interrogator skills rather than the machine's.

Turing test's problems overview

In the following paragraphs I'am going to present some structural criticism about the Turing test. I will also argue on how each of them applies to both the two interpretations, to support the thesis that the Turing test, whichever interpretation you would like to choose, is structurally ill-posed, hiding intelligence with requirements and expedients which have nothing to do with it as experienced by human beings. The issues I am going to present are the anthropocentrism of the test and the bias towards human intelligence posed by French in [3], the representation of intelligence as a decision problem and the consequent absence of gradient and some flaws of the game involving the useless pretense to be human even in their defects, the focus on imitating rather than communicating presented by Cullen in [1] and the subjectivity of the outcome relying it on the interrogator.

Culturally-oriented human-like intelligence

The first argument against the Turing test, stated by French in [3], is that not intelligence will pass the Turing test, but only an intelligent entity that has experienced the world as humans do. To clarify the idea behind this thesis I am going to report in brief the parable used by French himself: there are philosophers on a island focused on derive the essential concepts of "flying". On that island the only flying entities are seagulls. Unable to get a suitable and complete definition for the phenomenon and knowing the writings of Alan Turing, the philosophers decide to set-up a Turing test-like test for flight, named Seagull test. The Seagull test works much like the Turing test, the philosophers observe the behavior of the seagulls and of the entity under investigation through a radar screen. The entity will be said to have passed the Seagull test if the philosophers are unable to distinguish the two subjects. It's quite obvious that planes and many different species of birds will never be identified as flying entities by this test. Metaphors aside, the same applies within the Turing test, no entity will get intelligence awarded unless it displays a human-like intelligence.

The argument from French needs to be argued differently under the two versions of the test. Starting from the easiest case, the Standard interpretation, the claim from French appears strong. To pass the Standard Turing test a machine has to impersonate a man better than the man itself. We are explicitly looking for human-like intelligence, the goal is to outperform a man in being a man. The Original interpretation needs, instead, a deeper inspection. Here both the man and the machine have to display a behavior which falls outside their habits. In trying to imitate a woman they have to critically think about when and how their answers should be modified. This approach, not only make the game fairer, but eliminates, as much as possible, the common background of human experience between the man and the woman, trying to isolate the very essence of thinking, such that the man can no more rely on his own experience of the world as human. Through this expedient the problem is smoothed as much as possible, even if among human beings, there will be an unavoidable common notion of the world and the way we experience it which machines lack.

The point from French presented here lead us to look for the answer to the wrong question, while Turing's aim was to answer the question "Can machines think?" the test seems to answer to "Can machine think exactly like human beings?" which French himself tags as "significantly less interesting than the former" [3]. Such strong position from French needs to be discussed, the question "Can machine think exactly like human beings?" seems way far from being uninteresting since "It presumes that we know everything about the way humans think, everything about human intelligence" [2]. Another objection about the whole point from French is based on the fact that "general intelligence divorced from our own intelligence is a chimera"

[4]. Turing test is undoubtedly testing for human intelligence, and couldn't be otherwise, at least until we have a general theory of intelligence, which to be developed has for sure to start from the only evidence of intelligence we can study and understand experiencing it everyday, human's.

No gradient of intelligence

Taking for granted that Turing test looks for human intelligence and being fine with it, as argued in the previous section, we can start looking inside the spectrum of human intelligences. The absence of gradient regards how intelligence is perceived against how it is represented, both in the Original and the Standard tests. Starting from an explicative example I will go on arguing about the problem caused by turning intelligence into a decision problem, i.e. a problem that can be posed as a yes-no question of the inputs. Consider, just for the sake of the example, the Original interpretation of the Turing test, even if there is no particular reason to choose this over the Standard one. We have unchanged the first round of the game, in which a man (A) tries to fool the interrogator (C), in believing he is the woman (B), who in hers turn tries to help C. Now, in the second stage, instead of "What will happen when a machine takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" we ask ourselves "What will happen when a child takes the part of A in this game? Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?". There are plenty of reasons in believing the answer should be "no". Even for children of about 10 years or more, which have already fully developed all their human conversational skills, passing the test remains a mirage. Children for sure don't lack intelligence nor conversational abilities, so what brings them to failure? They just have not experienced all the aspects of a human life a man had. Imagine any kind of question regarding the sexual sphere, the professional life, or even the task of imitation itself for which a child is probably too naive for, they will put the child in serious troubles. One may argue that in principle, each one of us have a wide range of problematic questions based on the background we have. I agree, but, the arguments that may put in trouble a man are not known a priori, while in the case of the child the interrogator could easily guess them. From this example two conclusions can be drawn, either the child in not intelligent, or the Turing test has a problem in its formulation.The first hypothesis can be discarded, there are no reasons in thinking a child cannot be capable of thoughts while the second holds and the problem is that there isn't the notion of "gradient of intelligence" within the test, both the Standard and the Original one. Turing test reduces intelligence to be a binary problem, you can win or lose the game, you can be either intelligent or non-intelligent.

According to the possible outcomes of the Turing test systems are organized in two crisp sets (sets whose membership functions are boolean) the first one representing intelligent systems, the second one representing non-intelligent systems. This is not the way we experience intelligence. We, as humans, are used to rank people about any kind of characteristic and skill, and intelligence is no exception. The way humans experience intelligence is not a matter of crisp sets but fuzzy's (sets in which the membership function define, not the mere membership or not, but a measure of how strong is this membership). Turing turns a graded problem into a binary problem thresholding it, setting as a threshold the level of intelligence of grown-up humans, not just humans but specifically adults, that's why children will fail the Turing test. To avoid this problem my view is that has to be devised a multi-task test, like the one described by Cullen in [1], treating intelligence as an n-dimensional space, whatever it is and whichever are the n features defining that space (is not matter of this discussion trying to derive salient aspects of intelligence).

Encourages mistakes

Imagine a test instance in which the interrogator asks only for mathematical calculations of increasing difficulty. For sure at some point humans will start making errors becoming the task very challenging, while for the machine it would remain rather easy. To mimic humans in doing calculations machines have to introduce deliberately errors, otherwise an interrogator will be easily able to discriminate them. The problem is directly addressed by Turing in [8], he asks "Are they [machines] any worse for that [making no errors in calculations]?" The implicit answer "no", for the sake of the test, turns out to be "yes", within the test it is a weakness that should be avoided. There is a contradiction, there isn't any reason to penalize machines for handling perfectly mathematical problems, but the test does it. Definitely we are taking away some "intelligence" to achieve intelligence.

An objection may be that the purpose of the mistakes is to make the machine more human. This objection can be turned out thinking about how humans and machines make such errors. A human doesn't make errors based on complex statistical models a machine uses. Introducing errors basically embeds in our machine much more non human processes, going down under the surface, than the amount we are taking away. My view is that a good test should preserve qualities rather than introducing complex processes to hide them.

Now let's look at a practical scenario in which we have an intelligent (according to the Turing test) machine ready to be deployed on the market, probably the first thing the company will do before launching the new product will be to eliminate the voluntary mistakes introduced to mimic humans in doing mathematics. Deliberate errors are just a trick to pass the test, they have nothing to do with intelligence, nor with real world applications.

What argued so far represents a strong objection against the Standard interpretation of the Turing test, since the machine and the man are directly compared, and the "imitation" task has to be accomplished only by the machine. A test instance as the one presented at the beginning of this section represents a strategy likely to be implemented form the interrogator since his task is to directly discriminate between the two. Regarding the Original interpretation, the issue still apply, and still lead to easily discriminate the machine from others participants, but further considerations about the goal of each player bring us to think that, even if this strategy (still the one described in the example at the beginning of the section of asking only for mathematical calculations) should work, is not so likely to be implemented from an interrogator. May be A a man or a machine, the task of A is to imitate B (woman). Is there something useful, from the interrogator's perspective, in asking for calculations to distinguish between a woman and a man? In principle not, so this analysis seems strong enough to state that an average interrogator is not so likely to implement this strategy in trying to accomplish his own goal, even if, probably, exploiting it leads to identify the machine most of the times.

Lying as a key feature of intelligence

Whichever interpretation of the test we take into account lying is fundamental in passing it. Makes this lying a key feature of intelligence? Definitely "intelligent" but too sincere answers lead machines to failure.

Jamie Cullen in [1], describes a scenario in which a game called "Guessing game" is in place and works as follows: each "topic" is a word that the "interrogator" is looking for the machine to reproduce, but without explicitly naming it. In the instance presented, the topic is "stomach" and the conversation going on is:

I : *I'm thinking of one of your body parts as the topic.*

M : *Ok. What is its function?*

I : *it is the place where the food gets digested*

M : *I do not eat food. I use batteries for energy. Is this similar?*

I : *[...] Now that I'm aware that you're not human, perhaps I should correct my earlier statement: The topic is actually a human body part.*

M : *Is the topic "stomach"?*

I : *Yes.*

In this dialogue the machine shows strong evidence of intelligence, even beyond expectations demonstrating some sort fo self-awareness, moreover the dialogue is carried out successfully since the machine guesses the topic "stomach". Setting up and this conversation in the context of the Turing test gives us a machine that exhibits something that seems intelligence, but that fails in having it granted. Machine revealing itself as machine invalidate all the other accomplishments. That's why shifting the attention from imitation to communication as proposed by Cullen itself may be a better solution. "The objective of a Communication Test is not to convince an interrogator that an examinee (machine) is a human being, but simply to attempt to achieve a goal via sharing meaning with a conversation partner." [1]. The idea is to set up a test made of multiple games highlighting different aspects of human interactions, each one scored individually based on the quality of the communication and the accomplishments in these games, to build up a general statistics to be compared to previous runs of the game involving humans. This approach change the paradigm, is not necessary anymore to lie about its own identity, machines are judged on their "abilities" of being intelligent without any pretense to be human. Being aware of itself, of its own status of machine, may seem a more important feature of intelligence rather than the ability of lying, so communication games needs to be taken into account.

Relies on the subjectivity of the interrogator

Within the Turing test, although his role is often forgotten, the influence of the interrogator on the outcome may be way greater than expected. The interrogator, being human, has opinions, desires and biases, which contributes to bring some preconceptions about the outcome. Moreover, not just the mindset but also the skills of the interrogator have an influence. Turing in [8], defining the interrogator's characteristics, speaks about "average interrogator" without further specifications. What "average" means? We are looking for questioners without a background in AI? or not even any high level of education? or what else? is quite clear that the same game, changing the specification about the interrogator, may become way more, or way less difficult.

Another issue, is that the game, as it is formulated, creates a competition between participants, and the interrogator may be pushed, by human nature, in trying to win it at all costs. Will be a better solution to design a game not centered on a competition but on a mutual-satisfaction as the one proposed by Cullen in [1] assuring a fairer behavior from all the participants.

Looking at the Standard Turing test, the machine and the man are directly competing answering questions from the interrogator, which knows that there's a machine and a man. Skills and bias are everything in this framework, as already argued, with a sufficient skilled interrogator no machine will ever pass this test. Looking at scores it's more likely that we're evaluating interrogator skills rather than machine's intelligence.

Different considerations can be made for what concern the Original Turing test, here two aspects are crucial, first the task of imitation makes the game fairer, the machine have no more to beat a man in being a man, but to beat a man in imitating a woman, competing in a task in which none of the two is experienced, so both,"whatever" the question is, have to go back to the very essence of "thinking" which seems independent from the interrogator. Than, second, a point which is not clear in the Turing's paper, who, while describing the role of the interrogator in the first round of the test says "he (interrogator) knows them (man, woman) by labels X and Y", but than nothing is said about the substitution. Is the interrogator aware of the presence of a machine in the game or still believe the game is played among human beings only? For sure would be better that the interrogator isn't aware to shield the outcome from the interrogator's bias and tricks.

Conclusion

The Turing test, whatever interpretation one takes into account, as argued through this essay, suffers some structural issues making us missing the very essence of intelligence, spending too much efforts in building a gaming superstructure which introduces elements completely disconnected from the concept of intelligence. While some of these problems may be neglected, as the French thesis that the test is focused towards human intelligence, which is actually true but, as argued, no alternatives are on the table until a comprehensive theory of intelligence will be available, which in its turn seems reasonable to be derived form a comprehensive theory of human intelligence, some others applies and are due, not to the idea of turning the question "Can machines think?" into a game, but to the framework of that particular game which poses intelligence as a binary problem, focused on deceiving rather than communicating and relying on a subjective judgment. For what concern the two versions of the test specifically, from what argued along this essay I can safely state, in accordance with Susan Sterrett, that for sure the Original version is way more robust and reliable, "despite the similarity of the two tests, the first test is vastly superior" [7] , even if unfortunately nowadays the most widespread notion of Turing test is the Standard one.

As a real world evidence of the misleading path forced by the Turing test structure, we can take look at some results coming from the Loebner Prize competition. The Loebner prize is "an annual contest between computer programs to identify the most human programs, and eventually to award $100,000 to the program that first passes an unrestricted Turing test" [5].

Historically, most of the famous bots which have gained attention within the competition, such as ELIZA and PARRY, resort to tricks suited to fool human minds rather than implementing true intelligence. ELIZA is a chatbot proposed by Joseph Weisenbaum in 1966, which fools some judges in believing it is human by answering to questions with questions to continuously change the topic avoiding contradictions, "this works because most people like to talk about themselves, and are happy to believe the program is listening." [5]. PARRY written in 1972 by Kenneth Colby, simulates paranoid people and alternates times in which continuously change the topic to times in which it goes deep in specific stories embedded in its memory, and moreover it does not try to give an answer to everything, sometimes it just admit ignorance answering "I don't know". None of the two chatbots is intelligent, both implement nothing more than if then else rules, they just know how to deceive humans exploiting their weaknesses.

As shown through these examples, using the Turing test as our tool towards machine intelligence, may drive us out of the way. If research totally commits its efforts in passing the Turing test it's likely that we will find new tricks rather than intelligence. "We won't learn much about AI from Loebner Prize but we will learn some non negligible things about social psychology" [6].

References

1. Jamie Cullen. Imitation versus communication: Testing for human-like intelligence. *Minds and Machines*, 19(2):237–254, 2009.

2. Adam Drozdek. Human intelligence and turing test. *AI and Society*, 12(4):315–321, 1998.

3. Robert M. French. Subcognition and the limits of the turing test. *Mind*, 99(393):53–66, 1990.

4. David Hillel. Gelernter. The muse in the machine: computers and creative thought / david gelernter. pages 149–162, 1994.

5. Michael L. Mauldin. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. pages 16–21, 1994.

6. Stuart M. Shieber. Lessons from a restricted turing test. *Commun. ACM*, 37(6):70–78, June 1994.

7. Susan G. Sterrett. Turing's two tests for intelligence. *Minds and Machines*, 10(4):541–559, 2000.

8. Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(October):433–60, 1950.