Castro, J Theor Comput Sci 2017, 4:2
DOI: 10.4172/2376-130X.1000160

Journal of
Theoretical & Computational Science

Review Article Open Access

Computing Machinery, Intelligence and Undecidability

Paulo Castro*

Centre for Philosophy of Sciences of the University of Lisbon, Faculty of Sciences, University of Lisbon, CFCUL UID/FIL/00678/2013, Lisbon, Portugal

Abstract

In 1950, Alan Turing proposed a decision criterion for intelligence validation in a computer. Most simply, if a human judge was incapable of deciding from two witnesses which was the computer and which was the human, the machine would have acquired artificial intelligence. Here I will argue that the Turing test has a fundamental problem, making it impossible to provide human intelligence validation. In fact, the test is undecidable and thus cannot be considered a valid methodology to test for artificial intelligence. This does not mean that human intelligence simulation in a machine is unattainable. It means that we need a general theory offering common characteristics of intelligent agents and specific metrics to test for it. A theory able to predict intelligence emergence independently of our own subjective appreciation about how a system socially interacts with us. If such a theory is attainable or within our reach, in the coming years, remains an open problem.

Keywords: Turing test; Artificial intelligence; Intelligent agent; Undecidability

The Turing Test

In 1950, Alan Turing proposed a decision criterion for intelligence validation in a computer. Most simply, if a human judge was incapable of deciding which of two hidden witnesses was the computer and which was the human being, the machine would have become intelligent. Turing starts his seminal paper «Computing Machinery and Intelligence» by asking «can machines think» [1]. As he himself recognized, this is a problem that demands being equated in a more practical way. To do that Turing devised a testable situation, using the so called "game of imitation", where before a blind jury; two different gender witnesses are instructed to answer by writing to a list of questions. One of the witnesses will try to imitate the gender of the other, trying to lead the judge into believing that he is a woman when he is a man or otherwise. The judge's mission is naturally to provide the right gender identification, thus ending the game. Turing then asked what would happen if we were to substitute the role of one of the human witnesses by a computer program to imitate human behaviour. The question «can machines think» therefore becoming an empirical one. Namely, «can a human judge distinguish between another human and a machine». He was, of course, taking very seriously the hypothesis that human intelligence as a set of producible behaviours is a computable procedure, within our powers to simulate given an enough amount of memory, processing power and programming complexity.

The Turing debate, concerning the meaning of the Turing test to Artificial Intelligence feasibility and to what human intelligence is or is not, has been one of the major philosophical research domains over the last decades. Quite significantly, Horn's efforts to produce an argumentation map on the debate revealed, at the time he mapped it, 800 major "moves" «carried on by 400 scholars, researchers, and scientists worldwide, from at least ten academic disciplines» [2]. The debate includes claims, rebuttals, and counterrebutals upon the Turing test and his further claim that «at the end of the century… one will be able to speak of machines thinking without expecting to be contradicted» [1]. Fifty or so years debating over the issue, and the imitation game involving machines has been performed again and again. Using Eliza like strategies [3], chatterbots have been programmed and tested before human judges, only to find that we are still far from seriously consider that machines do think. Floridi et al., evaluating the 2008 Loebner Contest, reported:

«despite the brevity of our chats, a couple of questions and answers, were usually sufficient to confirm that the best machines are still not even close to resembling anything that might be open-mindedly called vaguely intelligent» [4].

More recently, and perhaps more productively, Warwick and Shah also reported about the Turing test sessions held in 2014 at the Royal Society. Their stand, however, was not philosophical, concerning whether machines have or have not acquired intelligence. Rather, they have focused on «the practical nature of the test as an operational test of intelligence in which a machine's conversational abilities are directly compared with those of a human». They added, and I find the remark quite substantial for what I will be arguing, that «we do…agree with Turing that engineering a machine to think can help us to understand how it is that we humans think» [5].

It should furthermore be noted that the game of imitation has an empirical status. It slightly recalls of what Einstein did when he conformed a constant accelerating frame to be empirically equivalent to a gravitationally acted frame. The observer cannot distinguish between the two situations and so it must be that they are at least equivalent. I think Turing, using the game of imitation, adopted a similar strategy to identify a humanely intelligent system. The Turing test is therefore what can be called an empirical validation test as Harnad seems to agree:

«It is important to understand that the Turing Test (TT) … sets AI's empirical goal to be to generate human-scale performance capacity. This goal will be met when the candidate's performance is totally indistinguishable from a human's. Until then, the TT simply represents what it is that AI must endeavour eventually to accomplish scientifically» [6].

***Corresponding author:** Paulo Castro, Centre for Philosophy of Sciences of the University of Lisbon, Faculty of Sciences, University of Lisbon, Lisbon, Portugal, Tel: +351217500000; E-mail: jpcastro@fc.ul.pt

Received August 05, 2017; **Accepted** September 18, 2017; **Published** September 21, 2017

Citation: Castro P (2017) Computing Machinery, Intelligence and Undecidability. J Theor Comput Sci 4: 160. doi:10.4172/2376-130X.1000160

Before starting my own analysis, I would like to point out a rather surprising result about the Turing test. Sato and Ikegami studied the computability aspect of the imitation game. What they did was to replace the judge by a Turing Machine and look for a universal effective procedure the machine could follow, in the imitation game, in order to distinguish between a human and another Turing Machine. It so happens that such a procedure does not exist. The Turing test is computationally undecidable, as the authors concluded:

«We have shown that "No machine can be an interrogator who can distinguish a man from a machine." However whether a man can play the role of the interrogator or not is still an interesting open question» [7].

It is this last open question that will be dealt here.

Undecidability within Turing's Version of the Game of Imitation

I will now proceed by asking, not if computers can think or even succeed in the imitation game, but instead if the imitation game can be used to assert that a machine has performed humanely. That will be my main question. It so happens that there is a flaw in Turing's version of the imitation game. Simply put it, nowhere in the test has been asserted that the judge has been proven to be an intelligent agent and, therefore, that he is able to foresee human intelligence in other systems. We just have always taken it for granted. We are humans and humans are thought to be intelligent by definition, and so able to identify likely intelligent systems. However, if one argues that in a certain sense human beings are also machines, perhaps biological and physical complex mechanical devices, the Turing test stands incomplete, because it needs prior approval on the judge supposedly intelligent skills.

Another way to realize the need for such prior validation would be noting that we could easily replace the judge by a computer, programmed to imitate the judge's behaviour. And, therefore, before asking the judge to choose between human and artificial witnesses, we should extend Turing's intelligence criteria to assert that the agent performing as a judge is himself able to assert correctly. That is, before beginning the test, we must validate the judge human abilities to perform as expected. We know from the Sato-Ikegami theorem that there will be a problem replacing the judge by a Turing Machine, since this mechanical judge will not be able to comply [7]. It is precisely to avoid such a replacement that we now must ensure that the judge is a human being and not a Turing Machine.

So let us investigate further about such a generic intelligence validation procedure. We will restrict ourselves to Turing's original configuration, assuming however that we do not know which of the three agents is the judge. Our problem will now be if there is a feasible procedure to identify intelligent behaviour. After the judge authority validation the Turing test can perform as usual.

Let us call three agents A, B and C, assuming that there exists a procedure to validate human intelligence in any agent, having the following two properties:

(i) No agent can apply the procedure to itself.

(ii) The procedure is only valid if it is performed by a human intelligent agent.

The first property ensures that the procedure is truthful. Since each agent is to be treated like a black box, if we would allow ourselves to accept a self-intelligence assertion coming from the agent itself, any procedure doing just that would be valid. That would make the procedure trivial, ill applicable and thus not trustful. We would never know if an agent was applying a legitimate procedure to its own case.

The second property ensures that the procedure is sound. If a less than a humanly intelligent machine could truly identify human behaviour, it could also truly imitate it. That would make the procedure useless. Property **(ii)** means that no machine can compute correctly the procedure unless it is human intelligent. Note, however, that a machine can still execute the procedure although giving an answer which can be either wrong, inconsistent or random, inconsistent or random. In a certain sense what has been said stands equivalent to the Sato and Ikegami conclusion.

Assuming that the procedure above exists and can be performed by either a human or computer agent (although this last will perform incorrectly), let us now accept that all agents can communicate with each another. For generality sake, let us also assume that the agents can be either humanly intelligent or computer intelligent and that each agent can perform like a judge, deciding which of the other two agents is human. Each agent will, of course, stand as a witness before each of the other two. All possible configurations for the true nature of each agent are listed in Table 1.

One can easily see that if we would have only a humanly intelligent agent, the procedure for human intelligence validation will be undecidable, since there is no other human intelligently agent to validate the former. On the other hand, with only computer intelligent agents acting, the procedure will also be undecidable fundamentally by the same reason, a lack of expertise power. Very much the same can be asserted using the Sato-Ikegami theorem.

Let us now analyse the two remaining situations. That is, Turing's original configuration for the game of imitation and the situation where only humanly intelligent agents are involved. Do they allow for human intelligence validation?

Consider first Turing's classical configuration: two humans and a machine. Suppose humanly intelligent agent A judges B to be a humanly intelligent agent and C to be a computer intelligent agent. Meaning that:

a) A finds B humanly intelligent and C computer intelligent.

Now suppose humanly intelligent agent B judges A to be humanly intelligent and C to be a computer intelligent agent. That is:

b) B finds A humanly intelligent and C computer intelligent.

Since by property **(ii)** A could only have performed correctly if it was humanly intelligent in the first place, one can thus write:

A	B	C	Classification
Humanly intelligent	Computer intelligent	Computer intelligent	Undecidable
Computer intelligent	Computer intelligent	Computer intelligent	Undecidable
Humanly intelligent	Humanly intelligent	Computer intelligent	?
Humanly intelligent	Humanly intelligent	Humanly intelligent	?

Table 1: All possible configurations for a human intelligent validation procedure involving three agents. Classification refers to the procedure possible results.

c) A finds B humanly intelligent only if A is humanly intelligent.

Similarly for B, one can assert that:

d) B finds A humanly intelligent only if B is humanly intelligent.

This, of course, leads us into an undecidability situation, since from **c)** and **d)** one must conclude that:

e) A finds B humanly intelligent only and only if B finds A humanly intelligent.

And since each agent can only find the other humanly intelligent, if it is itself humanly intelligent in the first place, we must finally conclude that:

f) A is humanly intelligent only and only if B is humanly intelligent.

This means that the human intelligence validation procedure can only work for a Turing classical configuration if there are two humanly intelligent agents in the first place. We immediately see that the same argument can be applied to the last situation in Table 1, where there are no computers. This shows that neither the presence nor the absence of a machine is relevant to the procedure soundness. Hence, any procedure in the Turing test style, conceived to identify the existence of at least one humanly intelligent agent, requires *ab initio* the presence of at least two humanly intelligent agents. Most concretely, we have found that the human intelligence validation procedure, in the Turing test style, whatever it may be, can only serve his purpose being redundant. Consequently, no such procedure exists.

Considering all possible situations in Table 1, we have arrived to the conclusion that any human intelligence validation procedure in the context of the Turing test is a non-computable task. And it so follows that the question it was supposed to answer, namely if there is a humanly intelligent agent among three possible candidates, becomes an undecidable one.

A third agent wouldn't contribute to alter this state of affairs because if it is computer intelligent, it will perform badly according to **(ii)**, and if it is humanly intelligent, it will be in the same situation as the one described for the other two agents, either in relation to A or B. Still we could try to sort the difficult out, allowing the procedure to be performed by more than three agents. Let us say we have n agents and that these have been grouped in subsets of three agents. This means that there will be one, two or none agents left. Each set of three agents as we now know would end in an undecidability situation. From **f)** it is clear that the same would happen if two agents remained. As to the situation where only one agent is left, it cannot be validated by any one of the other n-1 agents, since none of these can be asserted to be humanly intelligent.

Since the procedure to be used wasn't further specified save for properties **(i)** and **(ii)** and since it is to be applicable by observing any behaviour performed by a candidate agent, we can assume that it is a generic empirical procedure. Thus our final and complete statement must be that "the problem of human intelligence identification by empirical means, strictly using intelligent agents as the only authoritative resource, among a set of n candidates, is undecidable".

From this the answer to my initial question must be that the imitation game cannot be used to assert that a machine has performed humanely. The reason is that it is impossible to assert in the first place that any agent in the test, including the judge and the human witnesses, have performed human intelligently. In other words, the Turing test stands as an utterly undecidable scheme for identifying human intelligence.

Epistemological Consequences for Artificial Intelligence

The prior statements are not about mechanical computability, as the one encountered in Turing machines. The kind of undecidability herein stated is independent on the nature of the performing agents. It's more like an epistemological boundary, imposing on each agent the impossibility for proof reading the same degree of intelligence in similar agents. We have already seen that same phenomena, according to Sato and Ikegami, since Turing machines are computably unable to identify themselves as such.

We must now ask if the previous conclusion brings any appreciable epistemological consequences to Artificial Intelligence (AI) main goal. That is, to fully simulate intelligent human behaviour, thoughts and perhaps conscience in a machine or to achieve what is called the "singularity".

We can theorize this, as with other technological endeavours, to correspond to a set of carefully planned procedures, producing outputs brought about from empirical available data, that was inputted and processed by human intelligent agents. In a symbolic manipulation sense, while doing Technology, we can see ourselves as Turing machines, computing some set of instructions while assembling whatever the program tell us to do. Whatever the task at hand, if we bring it to an end, we would then have computed some plan in our heads, making the task computable.

Now, can we assume that assembling a human intelligent robot is a computable technological endeavour? It seems clear that if a human set of actions is to produce a preplanned effect, that effect should be verifiable at the end of the task. If, for instant, we have in mind building a car, by the end of the process we must be sure that the complex object thus obtained is, indeed, a car. An object that behaves in all acceptable ways like an automobile. In the given sense, for a task to be computable it is then crucial that the expected result should be checkable by some verification procedure. One that once applied to the object allows us to classify it, accordingly to what was intended for the object to be.

The Turing test has so far been considered the Golden Standard for AI identification. However, as we have seen, it cannot be applicable as such, since it is undecidable. This simply means that at the end of any sound procedure, put forward to simulate human intelligence, one cannot use the Turing test. This however does not mean that human intelligence simulation is unattainable. However, since the Turing test is an empirical validation test, one should ask if its undecidability has some substantial consequences for empirical tests, intended to identify human intelligence in a machine. In fact, what has been shown seems to suggest that all empirical tests, involving intelligent agents as the only authoritative resource, are undecidable and so unfitted for the task. One can argue that the same kind of undecidability scheme holds whenever we are faced with what may be called a «subjective» appreciation about intelligence, coming from a supposedly intelligent agent. It thus stands that we are in need of more general and objective criteria in order to successfully perform such a validation. From an epistemological point of view, we are in need of a general theory about human intelligence and how this can be implemented in a machine. Since we cannot, strictly by ourselves, validate intelligence (whether natural or artificial) in any given system, we need to properly define common characteristics of intelligent agents and devise specific metrics to test for such characteristics. Apparently, we need a theory able to predict intelligence emergence, from observable data in any given system, independently of our own appreciation about how the system

socially interacts with us.

In other words, it seems that the AI quest for intelligence singularity demands a rational endeavour far beyond such empirical and social tests, as the one proposed by Alan Turing. If such a theory is attainable or within our reach, in the coming years, remains an open problem.

Conclusion

From the analysis of Alan Turing 1950 proposed test for evaluating intelligent performance in a machine, I have suggested that "the problem of human intelligence identification by empirical means, strictly using intelligent agents as the only authoritative resource, among a set of n candidates, is undecidable". In other words the Turing test or any other test using the same scheme is undecidable. This does not mean that human intelligence simulation in a machine is unattainable. It means that we need a general theory offering common characteristics of intelligent agents and specific metrics to test for it. A theory able to predict intelligence emergence, independently of our own subjective appreciation about how the system socially interacts with us. If such a theory is attainable or within our reach, in the coming years, remains an open problem.

References

1. Turing AM (1950) Computing Machinery and Intelligence. Mind LIX, pp: 433-460.
2. Horn RE (2009) Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking. Computer Dordrecht: Springer, Netherlands, pp: 73-88.
3. Weizenbaum J (1966) ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine. Commun ACM 9: 36-45.
4. Floridi L, Taddeo M, Turilli M (2008) Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges - An Evaluation of the 2008 Loebner Contest. Minds and Machines 19: 145-150.
5. Warwick K, Shah H (2016) Can machines think? A report on Turing test experiments at the Royal Society. Journal of Experimental & Theoretical Artificial Intelligence 28: 989-1007.
6. Harnad S (1992) The Turing Test is Not a Trick: Turing Indistinguishability is a Scientific Criterion. SIGART Bulletin 3: 9-10.
7. Sato Y, Ikegami T (2004) Undecidability in the Imitation Game. Minds and Machines 14: 133-143.