**Computing Machinery, Intelligence and Undecidability**

P. Castro

Ph.D
University of Lisbon, Faculty of Sciences, Center for Philosophy of Sciences
Campo Grande Ed. C4
1749-016 Lisboa Portugal
paulo.castro.pi@gmail.com

Abstract:

In 1950 Alan Turing proposed a decision criteria 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. In this paper I will argue that the Turing test has a fundamental problem, making it impossible to provide human intelligence validation. Furthermore I will reason that all empirical tests destined to identify intelligence behavior cannot provide an answer and that consequently artificial intelligence identification is an undecidable problem. Since this must mean that there is no algorithmic procedure to list humanely intelligent systems, it must also mean that human intelligence is not computable and therefore that there is no pragmatic theory one can write to build an humanely intelligent robot.

Keywords: Turing test, Artificial Intelligence, intelligent agent, undecidability.

## 1. The problem with the game of imitation

In 1950 Alan Turing proposed a decision criteria 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.

We owe Turing the first formulation of an empirical criteria for identifying a machine's behavior no different from our own, that is for identifying an expression of artificial intelligence.

Turing starts his seminal paper «Computing Machinery and Intelligence» (Turing, 1950) by asking the question «can machines think». As he himself recognized this is a problem that demands being equated in a more empirical way. To do that Turing devised an experiment, using the game of imitation where before a blind jury, two different gender witnesses are instructed to reply in writing to a preset list of questions. One of the witnesses will try to imitate the gender of the other and so trick the judge in 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 ending the game.

So far, so good, and Turing asks what would happen if we were to substitute the role of one of the human witnesses by a computer programmed to imitate human behavior. The question «can machines think» therefore becomes an empirical one. Namely, «can a human judge distinguish between another human and a machine».

We now see what Turing was getting at. He was taking very seriously the hypothesis that human intelligence is a computable procedure. That is, with an enough amount of memory, processing power and programming complexity, one should be able to build a "human robot" so to speak. This assumption is not very far from the one shared today by most computer scientists, although I have to state my admiration for Turing's early conviction, when he says in the same article sixty years ago:

«It will simplify matters for the reader if I explain first my own beliefs in the matter. Consider first the more accurate form of the question [concerning a machine's performance in the imitation game]. I believe that in about fifty years' time it will be possible, to programme computers, with a storage capacity

of about 109, 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. The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. I believe further that no useful purpose is served by concealing these beliefs. The popular view that scientists proceed inexorably from well-established fact to well-established fact, never being influenced by any improved conjecture, is quite mistaken. Provided it is made clear which are proved facts and which are conjectures, no harm can result. Conjectures are of great importance since they suggest useful lines of research.» (Turing, 1950)

So let us further analyze such an overwhelming conjecture, one that has remained unsorted from Turing's days until now.

Let me first proceed like Turing, asking not if computers can think or even if they can pass the imitation game, but by asking instead «if the imitation game can be used to assert that a machine has performed humanely». That will be my main question.

Well, it so happens that there seems to be a flaw in the test. Simply put it, nowhere in the test has been asserted that the judge is proven to be an intelligent agent, and therefore able to foresee human intelligence in other systems . We just take it for granted, because we are humans and humans are thought to be intelligent by definition. But of course if one argues that human beings are also machines in a certain sense, perhaps biological or even physical, this makes the Turing test incomplete.

The test stands incomplete because we can easily replace the judge by another computer programmed by hypothesis to imitate the judge's behavior. And now before asking the judge to choose between his human and artificial witnesses, we must extend Turing's intelligence identification criteria to assert if the judge is himself able to assert correctly. That is, we must now validate that the judge is intelligent.

## 2. Turing's test undecidability

So let us do just that. Let us try and improve the game of imitation in order to suppress what seems to be a minor contrariety about the judge's competence. We avoid using a fourth intelligent agent, since this one would also require intelligence validation, evoking another intelligent agent and so on. Hence we will restrict ourselves to Turing's original configuration, although without knowing which of the three agents is the judge. It must be noted that the problem now is if it exists a procedure to identify intelligent behavior applicable to a Turing test scheme.

Let us call our agents A, B and C and let us assume that there is a general procedure to validate human intelligence in an agent, with the following 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 a truthful one. Since each agent is to be treated like a black box, if we would allow ourselves to accept an assertion of self intelligence coming from the agent itself, any procedure doing that would be valid. That would make the procedure 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 human intelligent machine could truly identify human behavior, it could also truly imitate it. That would make the procedure useless. Property ii) means that a machine cannot compute correctly the procedure unless it is human intelligent. Note however that a machine can still execute the procedure although giving an answer which is either wrong, inconsistent or random.

Accepting the procedure above, let us imagine a situation where three agents can communicate. The agents can be either human intelligent or computer intelligent. Each agent is to be the judge that will decide which of the other two agents is human and which is computer intelligent. Each agent will be a witness to be judged by each of the other two agents. All possible configurations of interest are listed below:

| A | B | C | Classification |
|---|---|---|---|
| Human intelligent | Human intelligent | Computer intelligent | ? |
| Human intelligent | Human intelligent | Human intelligent | ? |
| Human intelligent | Computer intelligent | Computer intelligent | Undecidable |
| Computer intelligent | Computer intelligent | Computer intelligent | Undecidable |

One can easily see that if we would have only a human intelligent agent, the test will be undecidable since there is no other human intelligent agent to validate the former.

On the other hand, with only computer intelligent agents present, the test will also be undecidable fundamentally by the same reason, a lack of expertise power.

Let us thus analyze the first two situations. That is, Turing's original configuration and the situation where there are only human intelligent agents present. Do they allow for the Turing test to be decidable?

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

a) A finds B a human intelligent agent and C a computer intelligent agent.

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

b) B finds A a human intelligent agent and C a computer intelligent agent.

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

c) A finds B a human intelligent agent only if A is a human intelligent agent.

And similarly for B, one can write that:

d) B finds A a human intelligent agent only if B is a human intelligent agent.

This of course gets us to an undecidability situation, since from c) and d) one must conclude that:

e) A finds B a human intelligent agent only and only if B finds A a human intelligent agent.

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

f) A is a human intelligent agent only and only if B is a human intelligent agent.

Which means that the Turing test classical configuration can only work if there are *a priori* two human intelligent agents.

Hence a test conceived to identify the existence of at least two human intelligent agents, among three agents, can only work if we know *a priori* two intelligent agents to exist. That is the exact purpose

of the test and since without such a procedure we can never be sure of such an existence, the question of which one of the agents is human and which is computer intelligent cannot be answered.

The Turing test is undecidable or more exactly the problem it was conceived to answer is an undecidable one, given two possible human intelligent agents in a set of three agents.

The third agent actions don't contribute either to the problem's solution because if it is computer intelligent, it will perform badly according to ii). And on the other hand, if it is human intelligent, it will be in the same situation as the one just described for the other two agents, either towards A or towards B.

The last statement shows of course that the situation where there are three human intelligent agents is also hopelessly undecidable.

We could try and get ourselves out of this mess allowing the test to be performed by more than three agents. Let us say we have *n* agents and that these have been grouped in sets of three agents. This means that there will be one, two or null agents left. Each set of three agents as we now know will end in an undecidability situation. The same will happen for the case where two agents remain. As to the case where only one agent is left, this cannot be validated by any one of the other *n*-1 agents, since none of these can be asserted to be human intelligent.

Since the procedure to be used wasn't further specified besides properties i) and ii) and since it is to be applied by observing any agent's behavior, we can say that it is an empirical procedure.

Thus our final conclusion must be that the problem of human intelligence identification by empirical means in a set of *n* (human or computer intelligent) agents is an undecidable problem.

### 3. Philosophical consequences for AI and the way we look at Nature

We can now ask if the previous conclusion brings any appreciable epistemological consequences to AI main goal. That is, the pretension of fully simulate intelligent human behavior, thoughts and perhaps conscience in a programmed machine.

We can theorize this as other technological endeavors to be carefully planed 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 behaving like Turing machines, computing some set of instructions while assembling in a robotic style whatever the program tell us to do. And whatever the task at hand, if we bring it to an end, we would then have computed some plan in our heads making that task computable.

Now, can we assume that assembling a human intelligent robot is a computable technological endeavor?

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 we have in mind building a car, by the end of the process we must be sure that the complex object obtained is indeed a car. An object that behaves in all acceptable ways like an automobile.

In the sense given above for a task to be computable it is then crucial that the expected result should be checked by a verification procedure. One that once applied allows us decide if an object is a car.

However, since the problem of identifying human intelligence by empirical means is an undecidable problem, there is no procedure to simulate human intelligence because it is a non computable task. Even if we would achieve such a result we wouldn't have a way to confirm we had succeeded.

As such this also means that we are deprived from a rational discourse allowing us to understand intelligence emergence. For if we would have it, we could foresee that emergence and thus identify intelligence as a property in a physical system.

It should be mentioned that this does not mean that nature cannot in effect give rise to intelligent creatures. It seems pretty obvious that it does. What the above reasoning means is that you can't have a preplanned technological procedure for turning raw material, of whatever kind and in whatever form, in an humanely intelligent system.

Although someone may achieve such a feat by accident, our conclusion makes it highly improbable to achieve it again and utterly impossible to achieve it in a systematic, industrial or technological way.

But still we look like very complicated machines. Very complex systems with some intelligible causality patchworks. Are we then the only machines to be out of Nature. Are we to have some rare property making us behind any comprehensible plan, according to which we could imulate ourselves technologically?

Once we have accepted that there is something non computable about life or perhaps about behavior in general, we began to wonder if rationality is powerful enough to cope with all phenomena in the Universe and indeed it it is a kind of big machine.

We have long thought Nature to be like a very complicated watch. Full of intricacies connecting its parts and so complex that only through a long survey, lasting centuries or perhaps millennia, could the human spirit shed some light about how it works.

Like a big computer, whose general algorithm we are to discover with time, effort and patience.

But now all the above seem to point to a kind of ontological limit imposed to the rational fishes swimming inside their empirical aquarium. And as reason seems incapable of coping with Nature completeness, we have to wonder if Nature is really mechanic in its essence rather than an organic wholeness, at times capable of mechanic behavior.

As something in phenomena seems to escape reason's upmost powers, our way to look at Nature should accommodate the hypothesis that something in the way things are reveals itself sometimes randomly, other mechanically and still other, teleologically.

We should try and build a new rationality accommodating the idea that there will be non computable actions in reality, that is, actions devoided of first causes, as way to express a kind of ontological liberty.

References

Agar, J. (2001). *Turing and the Universal Machine: The Making of the Modern Computer*. U.S.; Cambridge: Icon Books Ltd.

Davis, M. (2004). *The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions* (Dover Ed edition). Mineola, N.Y.: Dover Publications Inc.

Dreyfus, H. L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason* (Rev Ed edition). Cambridge, Mass: MIT Press.

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*(1), 145–150.

Millican, P. (Ed.). (1999). *Machines and Thought: The Legacy of Alan Turing, Volume I: Vol 1*. Oxford; New York: Clarendon Press.

Moor, J. H. (Ed.). (2003). *The Turing Test* (Vol. 30). Dordrecht: Springer Netherlands.

Turing, A. M. (1950). I.— Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460.

Warwick, K., & Shah, H. (2016). Taking the fifth amendment in Turing's imitation game. *Journal of Experimental & Theoretical Artificial Intelligence*, *0*(0), 1–11.

Weizenbaum, J. (1966). ELIZA—a Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun. ACM*, *9*(1), 36–45.