Detecting the full potential of intelligent machines: towards an observation-based approach

Matteo Palazzoli

September 4, 2023

Abstract

When measuring intelligence of living beings, a open approach is often used. In the case of humans, a battery of tests is chosen and submitted, while in the case of animals it's a usual thing to observe their behavior to detect intelligent actions. But this is not the case for machines, in which the historical methods to test for intelligence, like the Turing Test, are heavily human-biased. I argue that these methods don't sufficiently explore the machine's possibilities, risking to miss some important insights on all the intelligent behaviors these AIs can show. I explain the motivations for this claims, starting from some basic assumptions about minds and intelligence. Then, I propose a change of paradigm in studying AIs' intelligence, from testing in an active way to observing spontaneous behaviors, reducing the extent to which intelligence relates to humans. After unfolding three main adaptations to make it possible, I claim that observing spontaneous actions in machines may be a new possibility in AI research, and that spontaneity may be the next marker for machine thinking.

1 Introduction

The way that the Turing Test, with other similar imitation-based tests, measures thinking activity in machines is not suited for non-human forms of intelligence, because they don't detect all the machine's insights, being too human-biased in the choice of the task and in the way it's measured. This issue has already been studied by many (Crosby, 2020; Hayes & Ford, 1995; Hernández-Orallo, 2000, 2020), and some of them proposed their test reducing this human component. However, I suggest a deeper change in the paradigm, shifting the focus from testing to observing, in particular spontaneous actions. In section 2, I give a brief idea of what intelligence is, and how it's measured, both for humans and non-humans, underlining the openness of these methodologies with respect to the elements to look at. Then, in section 3 I introduce the concept of intelligent machine, explaining the Turing Test and some of the assumptions behind it. In section 4, I define some useful terms (agent, architecture, mind) and explain how these relate to each other. In section 5 I propose a fanciful scenario, applying the previous concepts, to show that the Turing Test is not well-suited to a unknown species, and then I compare this scenario with the one on machines. In section 6 I follow the reasoning proposing an observation-based paradigm, taking inspiration from animal studies. Lastly, in section 7 and 8 I suggest that spontaneous actions may be a key element to discover the real potential of a non-human agent, and that this could bring to new discoveries in the AI research.

2 Brief analysis of intelligence

With the term intelligence, people usually refer to human intelligence, and it's one of the main relevant topics in general psychology. I will use the following as working definition:

> "Intelligence is a very general mental capability that among the other things, involves the ability to resume, plan, solve problems, think abstractly, comprehend complex ideas, quickly learn from experience." (Gottfredson, 1997, 13)

The following psychology concepts can be found in "Psicologia generale" (Cherubini, Bricolo, & Reverberi, 2021). First, we can see that the definition is very general, explained by examples and non-quantified features. However, many psychologists have tried to measure human intelligence, for example

with the g-factor introduced firstly by Spearman. According to this theory, there exists one unique measure, the so-called g-factor, that can be estimated from tests. People can be tested across many different, non-predefined tasks; then, with factor analysis, a single value called g is estimated from all the different measures, maximizing the correlations. Empirical evidence has been collected to confirm that this unique factor is an indicator of performance in all cognitive tests, while other evidence confirms the independence of the g-factor with respect to the tests that are submitted: this means that the estimation from a set of tasks A, B, and C will be highly correlated to the one obtained from the set of tasks D, E and F. In conclusion, when measuring human intelligence, the general idea that has been adopted is to choose a "battery of tests", submit it to the subject, perform statistical calculations and obtain a measure that can be compared to the rest of the population. When dealing with other species instead, intelligence's measurement is more challenging. This topic has already been studied (Herzing, 2014) and most of the research has been conducted on animals like humans, especially primates or other mammals. It is shown that these species can exhibit intelligent behaviors like symbolic communication and semantic understanding. The author then explains that our methodology may present flaws when measuring animals' intelligence, and that some skills may be not tested properly, due mainly to a sensory gap.

3 Intelligent machines

Turing, in his paper "On computing machinery and intelligence" (Turing, 1950), explained what is now considered the most famous and historically relevant test for machine intelligence: the Imitation Game. Note that, in this paper, for the sake of simplicity, I will consider the ability of thinking and being intelligent as basically synonyms (the purpose is to avoid tedious issues with the precise characterization of each, that is not the purpose of this paper). In the Imitation Game, a "weak operationalization" is made (Crosby, 2020). The original question "Can machines think?" is being substituted with a more concrete, operational question, that is whenever a machine could pass the game. Some authors (Sterrett, 2000) argue that there exist two main interpretations of the test: the "Original Imitation Game" and the "Standard Turing Test". In this paper, for simplicity, I will consider as Turing Test the first one, that is organized as follows. There are three actors: a man (player A), a woman (player B), and an interrogator (player C). All players are physically separated; A and B can only communicate with C, through a text-teletype mechanism, and their identities are hidden under the name "X" and "Y". The interrogator doesn't know the real identities of X and Y, and its objective is to understand it, just by making questions to the terminals. In the end, C will state either "X is A and Y is B" or "X is B and Y is A". The objective of B is to help the interrogator, while A must try to fool C into making the wrong decision. Now Turing proposes the actual test:

> "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?" (Turing, 1950, 434)

The Turing Test laid the foundations of 70 years of discussions, and many authors proposed criticism and improvements (Oppy & Dowe, 2021). The main characteristic of the test is that, as the name suggests, it regards the task of imitation; more precisely, imitating a human is the marker of intelligence that Turing proposes with his operationalization. In his work, Turing denoted the subjects of the test as "machines". But what are the candidates for such machines? The digital computer is undoubtely the number one candidate. This device employs a silicon-based CPU and a central memory, and these components are made starting from simple electronic circuits that manipulate voltages. Even if modern AI systems can pass the Turing Test, the base architecture on which they are built remains the same. However, Turing himself discussed about the possibility of other techniques being used in the test:

> "It is natural that we should wish to permit every kind of engineering technique to be used in our machine. We also wish to allow the possibility that an engineer or team of engineers may construct a machine which works, but [...] they have applied a method which is largely experimental" (Turing, 1950, 435)

On the other hand, he excludes machines that are too human-like, because it would be not satisfactory to define this being as "thinking machine". He stated:

> "Finally, we wish to exclude from the machines men born in the usual manner. [...] it is probably possible to rear a complete individual from a single cell of the skin (say) of a man [...], but we would not be inclined to regard it as a case of 'constructing a thinking machine'." (Turing, 1950, 435/6)

So, the Turing Test is conceived to be submitted to every engineering product, assuming it's not too human-like. With this assumption, I can safely use the term "machine" as a generic digital computer, excluding other products like quantum computers for the sake of simplicity, and biological computers for not being sufficiently "machinery".

4 Agents, architectures, minds

I will refer as "agent" whatever thing or living being of which we may want to measure the intelligence. This comprehends humans, aliens, animals, machines and so on. Similarly, I will use the term "non-human agents" in a self-explicating way, referring to agents that are neither human nor "too human-like" (leaving this definition purposely vague since the details are not relevant). I will refer as "architecture" whatever material object the agent's mind it's built on. For example: the mind is associated with the brain, while the "mind" of a machine is associated with its hardware, so the brain is the architecture of the living being, while the hardware is the architecture of the machine. For simplicity, I used the term "mind" not only to refer to the humans' and animals' one, but also to the artificial one. With this terminology I would say that, in an agent, the architecture generates the mind, and the mind enables intelligence. Intelligence brings to intelligent behavior. The previous chain is not intended to be a precise mind-body characterization, but more like a general description of the terms I will use and how they relate to each other. From these assumptions it seems logical to assume that the different the architecture, the different the mind. Supposing that the mind would be the same, even with different underlying architecture, is a fairly strong assumption that is better to avoid. Clearly, the same holds for minds and intelligent behaviors.

5 The inadequacy of an imitation game

Suppose an alien agent reaches planet Earth. More specifically this agent is a Martian named M, that reached our planet with its own technology, and we humans would be led to think M is intelligent. To measure its intelligence, we make M imitate a human, performing English dialogues which quality will be judged by a human judge. Clearly, this scenario is absurd in many aspects. Who would make an alien behave as a human for the purpose of measuring its ability to think? Alien M doesn't know any words in English or in other known languages and doesn't know how to behave in a convincing way masquerading as a person. We assume M to have a different biology than ours, leading to different general architecture. Following the previously reasoning, M would have a different mind, enabling a set of intelligent behaviors that we cannot assume being equal to ours. This set may comprehend some skill that is common to both species with similar extent, for example problem-solving, or communication capabilities, but other qualities can be present in a totally different amount. We can suppose, for example, this alien being capable of performing really tedious calculations for ten hours straight without making a single error, or to memorize entire books by simply looking at them. No person (aside from specially gifted individuals) would be able to achieve these results. When looking at an imitation test performed on an alien-like agent, its limitation becomes more and more relevant. As many thinkers have noticed (Oppy & Dowe, 2021), this kind of tests are way too human-biased in the choice of the task and in the way the task is measured. In addition, it doesn't detect over-human capabilities, but on the contrary, it encourages artificial fallibility, that is purposedly making mistakes to resemble a real person, while avoiding being exposed as over-human (Damassino, 2020). There are other famous objections to the Turing Test but, in this paper, I'm focusing on the ones that are relevant to the explained issues. Going back to alien M, even if people managed to teach this agent the English language, history, culture, and arts, testing them on these topics would be unfair. Starting from the language, M would not be able to express their thoughts with full potential (this is an issue even between people across the globe). Then, M would not feel the same emotions people feel when talking about art and culture, and this could bring to the reveal of their true alien nature. This is a case in which a different set of emotions may lead to a negative result without necessarily indicating less intelligence.

Now, if we consider an intelligent machine instead of the alien, the same base assumptions hold. Both two agents are built with different architectures than humans', and they may have different minds, with a set of skills different to ours. This is especially true when considering black-box AI models, where interpretability, that is understanding which internal mechanisms produced a certain answer, is a serious issue and cannot be definitely overcome. Finally, we can notice that there are key differences between human testing explained in section 2 and machine testing: the first being open and second being extremely limited in possibilities.

6 A passive approach

In section 5 I argued that the Turing Test can have various issues when submitting it to non-human agents due to the strong bias towards human capabilities. The problem is that the test was conceived especially for non-human agents, as we've seen that Turing himself excluded agents that were human-like. Therefore, how can we successfully measure an agent's intelligence, getting rid of the human bias? This is a difficult problem, addressed by many researchers:

> "When considering humans in a vast space of intelligence, locating them as yet another point in this space becomes a Copernican revolution" (Hernández-Orallo, 2020, 556)

Many proposed different tests, like the C-test (Hernández-Orallo, 2000), and the AnimalAI Testbed (Crosby, 2020). All of these are of great interest, but the aspect that mattered most to me was eliminating the need of choosing the skill to be measured. To successfully address the extent of non-human intelligence, I suggest a change of paradigm, that is from testing to observing. Observation, in this case, is the activity of analyzing an agent's behavior without interfering with them. A key element here is spontaneity: if we can successfully observe a spontaneous behavior of an agent, we can identify their skills from scratch and trying to quantify the extent to which they happen, in terms of complexity and frequency. We already know how to do it with animals: COMPLEX (COmplexity of Markers for Profling Life in EXobiology) is an exercise being developed to study non-human intelligence (Herzing, 2014). After observation, judges evaluate the behaviors across five main dimensions: Encephalization Quotient, Communication Signals, Individual Complexity, Social Complexity and Interspecies Interaction.

These are not a product of a single activity, but from whatever activity belonging to a certain category, leaving open space for the observation of every possible behavior. Even if this approach was developed for living beings, I would suggest the adoption of a similar method to detect any intelligent behavior from the machines.

7 The importance of being spontaneous

There are some issues when adjusting this approach to the machines. In this section, I will address three challenges and I try to give a starting point to cope with them. The first two are smaller issues that can be overcome with the implementation of technological aspects and methodologies, while the third one has, in my opinion, the most potential and I believe it as being a new aspect that can lead to improvements and discoveries in the study of AI. The first challenge is the embodiment of the machine: providing it something that allows interactions with the external environment, with sensors and actuators, respectively for perception and movement. Achieving embodied AI is no big deal and it is already being used in psychiatry environments (Fiske, Henningsen, & Buyx, 2019). The second challenge is the lack of a "driver". With driver I mean the "will to do something", and it can be operationalized for example into seeking a reward or avoiding negative stimuli. Providing the robot at least one driver is a necessary condition for it to make spontaneous actions. This challenge too is being studied and nowadays there are modern machine learning and reinforcement learning techniques that leverage this aspect. The third challenge is about social interactions. As studied (Herzing, 2014), part of the intelligent behaviors that animals can show, concern intra-species interactions; in other words, communicating with other individuals. This communication can be for example performing alarm calls for threats, but also transfer learning methodologies and teaching mechanisms (Bender, Herzing, & Bjorklund, 2008). If we can build more than one individual of an intelligent robot with drivers, we could witness some behaviors and actions that are less investigated, that would be of great interest

for the research in AI. In this section I isolated these three elements as crucial aspects to perform observation-based research; however, it is not intended to be an exhaustive list and there may be other necessary conditions that are not listed here, mainly regarding the presence of spontaneous actions. It may be that drivers alone are not sufficient, and some authors claim that a full artificial consciousness is needed. In the end, I suggest that, in order to observe never-seen behaviors, the fundamental key element is the spontaneity of actions, and I advocate that the presence of these spontaneous action can be seen as a good marker for thinking activity. At this point of the paper, the reader may wonder if it's safe to perform such updates to intelligent machines and letting them behave spontaneously. Although it's not the main focus of this paper, and it would require an extensive separate discussion, to this regard I agree that a certain attention to ethics is required. However, the simple principle of building harmless robot bodies and safe test environments can be followed to avoid nearly all types of risks.

8 Conclusions

I started by stating that the architecture of agents is often different, and that it's not legit to assume that their mind would be the same as ours. Therefore, I claimed that a human imitation test is not well suited for machines, giving the example of alien M, that is a black box resembling intelligent machines. Afterwards, I took inspiration from animal analysis to explain that we should observe agents without interfering with them, in order to let them take spontaneous actions. Adapting this concept to machines, some updates would be required, and I suggested that it would be useful to make multiple individuals develop intra-species communication. In conclusion, I claimed that, in order to detect every aspect of an intelligent agent, the key factor lies in observing spontaneous actions and not testing a predefined skill, suggesting that spontaneity can be seen as marker for thinking activity. This change of perspective can be crucial because it may help us redefine the way intelligence is conceived, making it less human-centered and enabling us to understand the full potential of other agents, them being machines or living beings.

References

Bender, C. E., Herzing, D. L., & Bjorklund, D. F. (2008, Jul). Evidence of teaching in atlantic spotted dolphins (stenella frontalis) by mother dolphins foraging in the presence of their calves. *Animal Cognition*, *12*(1), 43–53. Retrieved from https://doi.org/10.1007/s10071-008-0169-9 doi: 10.1007/s10071-008-0169-9

Cherubini, P., Bricolo, E., & Reverberi, C. (2021). *Psicologia generale* (Nuova edizione ed.). Raffaello Cortina Editore.

Crosby, M. (2020, Aug). Building thinking machines by solving animal cognition tasks. *Minds and Machines*, *30*(4), 589–615. Retrieved from https://doi.org/10.1007/s11023-020-09535-6 doi: 10.1007/s11023-020-09535-6

Damassino, N. M. (2020, Nov). The questioning turing test. *Minds and Machines*, *30*(4), 563–587. Retrieved from https://doi.org/10.1007/s11023-020-09551-6 doi: 10.1007/s11023-020-09551-6

Fiske, A., Henningsen, P., & Buyx, A. (2019, May 09). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *J Med Internet Res*, *21*(5), e13216. Retrieved from https://doi.org/10.2196/13216 doi: 10.2196/13216

Gottfredson, L. S. (1997, Jan). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*(1), 13–23. Retrieved from https://doi.org/10.1016/s0160-2896(97)90011-8 doi: 10.1016/s0160-2896(97)90011-8

Hayes, P. J., & Ford, K. M. (1995, Aug). Turing test considered harmful. *International Joint Conference on Artificial Intelligence*, *1*, 972–977. Retrieved from http://ijcai.org/Proceedings/95-1/Papers/125.pdf

Hernández-Orallo, J. (2000, Apr). Beyond the turing test. *Journal of Logic, Language and Information*, *9*, 447–466. Retrieved from https://doi.org/10.1023/a:1008367325700 doi: 10.1023/a:1008367325700

Hernández-Orallo, J. (2020, Nov). Twenty years beyond the turing test: moving beyond the human judges too. *Minds and Machines*, *30*(4), 533–562. Retrieved from https://doi.org/10.1007/s11023-020-09549-0 doi: 10.1007/s11023-020-09549-0

Herzing, D. L. (2014, Feb). Profiling nonhuman intelligence: An exercise in developing unbiased tools for describing other "types" of intelligence on earth. *Acta Astronautica*, *94*(2), 676–680. Retrieved from https://doi.org/10.1016/j.actaastro.2013.08.007 doi: 10.1016/j.actaastro.2013.08.007

Oppy, G., & Dowe, D. (2021). *The turing test.* Retrieved from https://plato.stanford.edu/archives/win2021/entries/turing-test/

Sterrett, S. G. (2000). Turing's two tests for intelligence. *Minds and Machines*, *10*(4), 541–559. Retrieved from https://doi.org/10.1023/a:1011242120015 doi: 10.1023/a:1011242120015

Turing, A. (1950, Oct). I.—computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. Retrieved from https://doi.org/10.1093/mind/lix.236.433 doi: 10.1093/mind/lix.236.433