Julia DiTomas
Intro to Cognitive Science
Fall 2024

A Test to Recognize Human-Like Consciousness in Machines

How does one define artificial intelligence? The release of ChatGPT made AI ubiquitous in our society, as it is now used by entrepreneurs, lawyers, web developers, police officers, students, writers, artists, and many more. The Large Language Model (LLM) certainly seems intelligent; if you ask it to write a sonnet about *Finding Nemo* or to explain string theory in layman's terms, it will return impressive results. ChatGPT-4 even passes the Turing test, imitating human behavior and responses well enough to fool human researchers. [1, 2]

At the same time, it is generally understood that ChatGPT is not equivalent to a human. ChatGPT may perform well as one, but this is only because it has studied an enormous set of human data and knows what word fits next in a sentence. This does not mean that it actually thinks like a human, or has any cognitive ability at all. It may learn, but it lacks curiosity. It may describe a beautiful sunset, but it has no subjective feelings about that image. It may even solve problems, but, considering it cannot think, only if it has already seen the answer. It has none of the essential qualities of humans; in fact there is nothing human about it but our data. When prompted with the right question, such as arithmetic problems, it fails and the illusion is gone (plus, it gives all answers confidently even when incorrect, indicating a lack of metacognition [3]). If ChatGPT passes the Turing test, but is still not regarded as a human-like intelligence, then perhaps it is the metric we are using which is inadequate.

If a machine with true human-like consciousness were to exist in the future, however, everything would be different. We would need to ask ourselves a range of philosophical questions about our moral obligations to our creations and ourselves. Would we give machines human rights and allow them ownership of ideas and property? Would we entrust them with roles such as teacher, judge, and diplomat? Would human-created art be no different from that made by AI? Could we answer questions about consciousness in general and the human mind? And would artificial intelligence solve all of our problems for us? These questions are out of the scope of this paper, but are some examples of considerations for a hypothetical future with machine consciousness. For any practical change that might take place in this hypothetical future, such as allowing AI the right to patents, we must know when it is the right time to effect this change. Conversely, we are reminded to not enact such a change before machines are conscious and to not trick ourselves by anthropomorphizing artificial intelligence.

It is of the utmost importance to develop a method for evaluating the consciousness of machines in the modern age of rapid AI development. Consciousness has many meanings, but is herein defined by three qualities: curiosity, creativity, and self-evaluation. These qualities imply others such as subjective experience and problem solving capabilities, and thus encompass much of what is regarded as consciousness. The ideal test of a human-like intelligence is one that captures all three, and thus lies in the ability to generate new ideas. Thus, I propose that once AI solves a previously unsolved problem, we may consider it conscious.

In this paper, I will first defend the decision to take a functionalist approach to the consciousness problem. I will then provide details about the types of problems that must be solved to verify consciousness and give justification. Finally, I will present counter arguments, along with my rebuttals, and communicate the practical significance of the issue.

I. Why the functionalist approach

The main flaw of the Turing test is that it is subjective. It is left to the human interrogator to decide what questions to ask, and to interpret the responses as being signs of human or machine intelligence. This implies that a machine's success is dependent on the interrogator. If AI can sometimes trick some people in some contexts, then the results are ambiguous. Would a machine be considered intelligent solely because it passed the Turing test once?

Although the Turing test is insufficient to accurately assess modern day artificial intelligence, it provides a strong basis for how this might be done. Like the Turing test, the solution proposed in this paper is a functionalist one. This means that it does not matter how the system is built or from what materials it is constructed. All that matters is what it does. Unlike Turing, I even place no constraints on the type of machine under consideration, whether it is a digital computer, a mechanical machine, a biological robot, or something else entirely. The reason for this is that in the modern age, other types of non-natural intelligence are imaginable, especially types that blur the line between the natural and the manmade. For example, researchers have made robots out of frog cells capable of complex behavior such as self-repair [4]; these are candidates of the test. On another note, consider the potential of people in the future uploading minds to machines or supplementing parts of brains with artificial neurons. They would probably want to ensure that such processes maintain consciousness.

The functionalist approach, inspired by Turing, is chosen because behavior is how we mainly verify cognition and consciousness in fellow humans. Currently, we do not fully understand consciousness or know where to physically look for it. Each person knows only that he is conscious, and *trusts* that everyone else also has thoughts, emotions, and similar cognitive experiences based on what *is* observable. By assuming that certain outcomes require certain processes to take place in the mind, we can seek the outcomes to indirectly imply the processes. For example, in a study of consciousness in infants, the fact that the subjects would persist in the task longer when on the right track, and opt-out when wrong, was taken as evidence that infants can evaluate their confidence of being correct. [3] Certainly logical, but still based on an assumption. For its purposes, this paper assumes that generating new ideas requires curiosity (exploration of an unknown solution space), creativity (imagination of solutions that may work), and metacognition (to know when the solution has been reached or where one went wrong). If other theories state that more is required for generating new ideas, such as the mind being partly quantum [5], then these required traits would also be implied by the test at hand. Finally, if there is some physical characteristic of the brain required for consciousness, irreplicable in machines, then the functionalist test will still suffice as all machines will fail.

II. Why the task of idea generation

We now turn to what sort of behavior should be observed for the test. To be regarded as a conscious being, a machine must generate a new idea. Problem solving by itself is not enough. To see why, consider the problem of figuring out the weight of a small rock. A very simple mechanical scale could be built by hanging a spring from the ceiling, and using known weights to mark on the wall how far the spring stretches when each weight is attached to the bottom. We now have a machine that takes as input an object, solves the problem of deducing its weight, and outputs the result by the labels on the wall. It is a problem-solving machine, but it is neither intelligent nor conscious.

On the other hand, consider the wild hyenas studied in [6]. Researchers presented a problem to them, in the form of a chunk of meat within a puzzle box that required sliding a latch to open, to study the effects of diverse exploratory behaviors and neophobia (fear of new things) in problem solving. It is safe to assume that the hyenas never saw anything like this problem before, as this is what made the study effective. Hyenas were able to solve this problem successfully because they had motivation (a reward inside), curiosity (exploring what could be done with the box), and creativity (mentally generating a solution).

We now consider ChatGPT. Is this LLM more like the hyenas, or more like the spring? It seems like it can solve a wide range of problems and puzzles, and it can learn, as opposed to only doing what it was originally programmed to do. But it still cannot generate new ideas. It only knows anything because the knowledge was made available to it, and it only “solves” any problem because it has seen the answer. Even if it was not purposefully programmed to complete the task, the intelligence lies in the humans from whom it studied. Therefore, it is more akin to the spring.

If artificial intelligence were to solve a problem that it has never seen before, then we would have proof of its consciousness. ChatGPT and other AI models can falsely claim to have feelings, to possess consciousness, and to be able to think and problem solve, but solving an unknown problem cannot possibly be faked. This would tell us, undeniably, that the machine under test is not stealing its idea, but actually thinking.

To be specific, the three criteria for the problem are the following:

1. It must be something new to the test subject.
2. It must be verifiable.

3. It must be of sufficient complexity to require imagination.

The first criterion, as discussed above, ensures that new idea generation is required to solve the problem. The second means that it must be a problem that has a solution which humans can check, to be able to draw conclusions from the test. The third criterion means that the problem must be open-ended enough, or have a sufficiently large search space, that brute force algorithms will not happen upon an answer. The problem solver is thus required to explore and to think of a reasonable solution to test, because it will not have enough time to test every possibility. The problem solver is also likely to try incorrect solutions at first, which will require him to be able to self-correct before obtaining an actual solution.

Note that although the criteria are the same, the problems required for different test subjects may differ. For example, the latched puzzle box problem for the hyenas met all the criteria, because 1) the subjects had never encountered latched puzzle boxes previously, 2) the researchers could verify success by observing the subject opening the box and taking the reward, and 3) there were infinite ways that the subject could possibly interact with the box. This would not be an appropriate test for any machine that has been taught already about sliding latches, but if there were a robot that never learned about sliding latches and could be left to explore the box as the hyenas did, then it could be an adequate test for that robot.

For ChatGPT and any artificial intelligence that seems to already know "everything", the ideal problem may be one previously unsolved by humanity, such as the remaining unsolved Millenium problems. This way, we would be certain that the AI had not already been exposed to the answer. The test subject would be allowed to conduct research, learn, and collect any extra data that it requires, just as a human solver would, but the entire problem solving process would require curiosity, creativity, and self-evaluation.

III. Counter arguments and rebuttals

The first argument I consider is the argument of unoriginality, which simply asks the question "How do we know humans have original ideas?" This is a valid point, because humans are capable of copying one another without even realizing it. How does a composer know that she has never heard her "new" melody before? How does any creative know, for sure, that he is not just combining and averaging all the art he has seen?

I counter this counterargument with another question: If all creators were actually just copying from prior creators, then where would it have begun? The fact that humankind advances in any field, be it music, art, literature, or science, to create something not found in nature, proves that people are capable of idea generation.

The second counterargument I consider is the unfairness argument. This is the argument that the test proposed in this paper is unfair because it requires different tasks of different beings. All the hyena had to do was open a box. Why must other test subjects, as I proposed in the previous section, solve problems that the whole human race has not yet succeeded in solving?

I reply to this argument by pointing out that there is no way to put all existing creatures and machines, and all that may possibly exist, on the same playing field. If a certain artificial intelligence has the advantage of knowing and remembering (or at least quickly accessing) practically all the information known to mankind, and can operate much more quickly than the human brain, and is capable of solving new problems, then it is reasonable to expect it to be able to solve the Millenium problems or other very difficult problems. And if it cannot solve a single problem that it has never seen before, then the most rational explanation is that it is incapable of generating new ideas. If we were to task the hyena and ChatGPT with the same problem, *that* would be unfair. In order for it to be something that ChatGPT has never seen before, it would probably be something far beyond the abilities of the hyena, which would lead us to incorrectly believe that the hyena *cannot* generate new ideas and *is not conscious.*

The third counterargument I consider is the problem ambiguity argument. Consider the example I gave in the introduction, that you could ask ChatGPT to write you a sonnet about *Finding Nemo*, and it would accomplish this task successfully. What if it just so happens that it has never been asked this before? Assuming that the resulting poem was not in its studied database, then is this not an example of idea generation?

This would not be an example of new idea generation because a sonnet can be written by simple rule following. A sonnet is a rhyming poem with fourteen lines and 10 syllables per line. Imagine if I took all the words from *Finding Nemo* and wrote each one on a piece of paper, and then chose one at a time to construct a sonnet, replacing any selection that broke the poem's rules or turned out to be the wrong part of speech. After much time, I would have a sonnet. Essentially, what I would have done is brute-forced the answer after reducing the problem space to a manageable number of words.

I concede that the most difficult part of the test I have proposed is meeting criterion #3. It is not easy to know which problems require imagination, *especially* if the functionalist approach of the test blinds us to the inner workings of the machine. But it is not impossible either. It may just require significant discussion among a group of experts. As substantial scrutinization and work are already required to verify cutting-edge advances in mathematics and other fields, I find this reasonable.

(Before continuing, I would like to note that the problem may be solved by rule-following, but in such a case, any algorithms to solve it should be unknown to the test subject, so that the test subject may only use algorithms that it generates itself.)

IV. Discussion and Conclusion

In this paper I have discussed the shortcomings of the Turing test to assess machine consciousness in the modern age and presented an updated test. This has practical implications as we need to be able to evaluate consciousness to know what role artificial intelligence should play in our society; this is equally important whether the machine passes or fails the test.

For example, until AI is able to think critically and evaluate itself, users should know that it can be confidently wrong, and that it would not be able to tell if it had been fed incorrect information or maliciously attacked. Users should also know that it is currently incapable of generating new ideas, so anything that is offered to them by ChatGPT or DALL-E is made from material taken from human creators (which may also inform users about what they should share online). Anyone considering the use of AI to respond to emails or interact personally should consider that it has no subjective experience or ability to empathize with the other party. If it does gain consciousness, then we would need to understand that it has subjectiveness, so it might not provide completely objective information and might not be any better than a human in impartial roles such as judge or referee. On the other hand, it would then have self-awareness and might benefit from knowing its own shortcomings.

The consciousness test proposed herein takes inspiration from the Turing test, but adds necessary guidance to what the interrogators should ask of the machine, in order to be more objective and more precisely seek consciousness. Specifically, this test looks for creativity, curiosity, and self-evaluation as evidenced by idea generation. Although this paper has focused on ChatGPT, this paper's consciousness test is not only applicable to LLMs, but could potentially be used for any being, whether manmade, natural, or somewhere in between. As a final note, it is important to remember that every situation is different and that the future holds possibilities that I have not even imagined. The solution proposed may thus work in the meantime, but become itself outdated at some point, and require another iteration for compatibility with a currently unknown world.

Sources

1. Mei, Q., Xie, Y., Yuan, W., Jackson, M. (2024) A Turing test of whether AI chatbots are behaviorally similar to humans. *PNAS, 121*(9). https://doi.org/10.1073/pnas.2313925121
2. Turing, A. M. (1950) Computing Machinery and Intelligence. *MIND: A Quarterly Review of Psychology and Philosophy, 59*(236), 433-460. https://phil415.pbworks.com/f/TuringComputing.pdf
3. Dehaene, S., Lau, H., Kouider, S. (2021) What Is Consciousness, and Could Machines Have It? *Robotics, AI, and Humanity,* 43-54. https://doi.org/10.1007/978-3-030-54173-6_4
4. Blackiston, D., Bongard, J., Kriegman, S., and Levin, M. (2023) Biological Robots: Perspectives on an Emerging Interdisciplinary Field. *Soft Robotics, 10(4),* https://doi.org/10.1089/soro.2022.0142
5. Kauffman, S. A., Roli, A. (2023) What is consciousness? Artificial intelligence, real intelligence, quantum mind and qualia. *Biological Journal of the Linnean Society, 139*(4), 530-538. https://doi.org/10.1093/biolinnean/blac092
6. Benson-Amram, S. and Holekamp, K.E. (2012) Innovative problem solving by wild spotted hyenas. *Proceedings of the Royal Society B, 279(1744),* https://doi.org/10.1098/rspb.2012.1450
7. Gabriel, M. (2021) Could a Robot Be Conscious? Some Lessons from Philosophy. *Robotics, AI, and Humanity,* 58-67. https://doi.org/10.1007/978-3-030-54173-6_4