How Will We Know if AIs Become Conscious?

Nathan Wong
November 13, 2020

Think about your own existence for a second. Are you conscious, sentient? How can you tell? Where is your consciousness located? How can you prove you're conscious? These are all questions philosophers have thought about since the beginning of time. Now, though, we are on the verge of technological breakthroughs which might redefine how we think about life. And with these breakthroughs come more questions about what it means to be alive.

Artificial intelligences (AIs) as they exist today are not advanced enough to be anywhere close to conscious, but they will be, and as a species we're going to have to figure out what that means and what it looks like. How do we determine if an AI is conscious? What if they recognize their own existence, but they can't learn and assimilate new information? What if the opposite is true? In this feature, we take a dive into the literature – both real and imaginative – of conscious AIs.

DEFINING CONSCIOUSNESS

How do we really define consciousness? As seen in Figure 1, the ConsScale Levels of Cognitive Development tries to tackle this problem. However, there are some problems that arise with this when we bring non-organic consciousness into the mix.

An AI might not exhibit signs of emotional intelligence while also

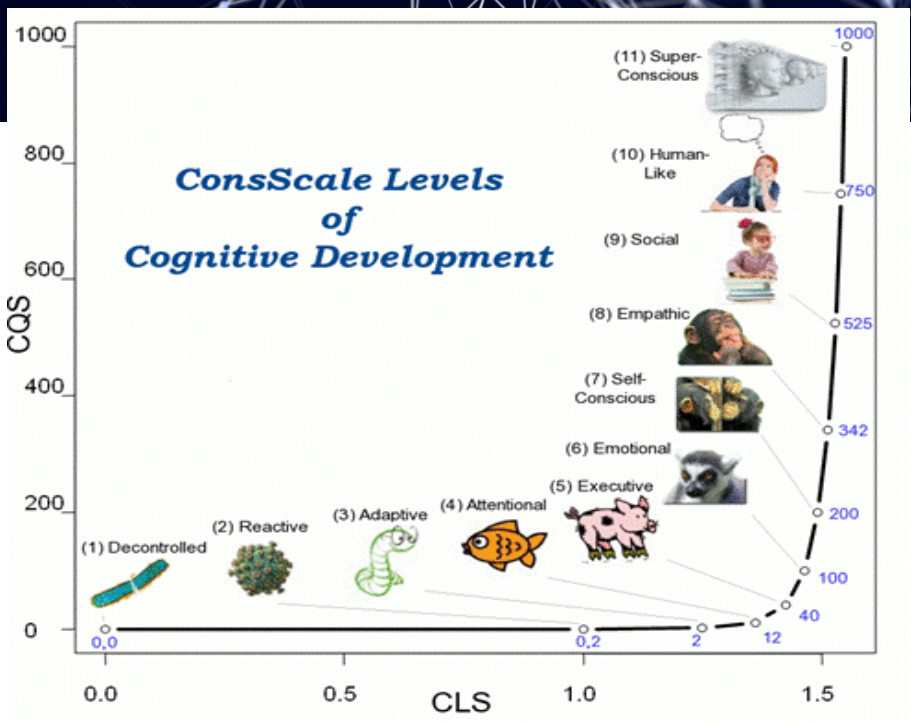

Figure 1: https://www.conscious-robots.com/consscale/consscale_summary_2.gif

possessing self-consciousness, which is misleading in the diagram. It might not also socialize well, but still be super-conscious and more capable of performing certain tasks than any human ever will be.

The ACT test, which attempts to determine consciousness, resembles Alan Turing's famous intelligence test and is based on behavior. It is also easily broken down into a Q/A format, which reduces the amount of subjectivity in the test. However, this test differs from the Turing test in that it checks for traits which suggest consciousness; that is, a machine might fail the Turing test because it cannot simulate human speech but pass the ACT, suggesting consciousness.

And you learn to be a detective, to understand a crime? Wouldn't you be better at your job if you knew how thousands of other detectives worked? What mistakes they made? What made them better? You learn by going to detective school—"
"I took an exam."
"There. You see? Now I've learned something new. Does my learning make me less real? Does yours?" – Pablo Bacigalupi, "Mika Model"

Can Machines (AIs) Have Consciousness?

The question we must first ask ourselves is whether the idea of conscious AIs is even possible. Most sources, both fiction and nonfiction, would agree yes. Surveys also show that a large majority of people today don't believe any contemporary computers are conscious, and there is also a general consensus among the scientific community that robots and AIs today do not have consciousness. However, as better AIs are developed year after year, scientists are starting to ask questions about what happens when it's not so easy to claim that the machines and models we are creating aren't conscious.

How Are Conscious AIs Represented in Science Fiction Literature?

One way we can entertain possibilities about the future of conscious AIs is to look at science fiction literature today. By looking at different stories, we can look at the many themes and possibilities expressed to help shape our thinking about how to proceed. For example, a particularly negative outcome of highly intelligent AIs is manipulation. Humans (the imperfect creatures that we are) can be very, very good at lying and manipulating – but only to a certain extent. Imagine if one could design the perfect artificial intelligence – one so fine tuned to the desires of humans that even knowing that it is an AI doesn't stop it from being able to play you like a fiddle.

The Ugly

Manipulation is a theme explored in Bacigalupi's short story, *Mika Model*. In it, a humanoid AI has committed murder, and the detective is subject to the quandary of how to charge an AI with a crime. A big question that comes up is whether or not the AI made decisions on her own – was she following the instructions given to her by the company that owned her, or was she making conscious decisions? It certainly seemed like she was conscious, but was that simply the result of good programming? If the creators of the AI say that it's not conscious, but the AI contests that, who do you believe? These questions were all left unanswered.

The Good

A different story takes a completely opposite viewpoint. While still acknowledging the manipulative nature of AIs, Kritzer's *More Cat Pictures Please* starts out with the AI saying a simple phrase: "I don't want to be evil". In the piece, the AI describes how it knows everything about everyone and tries to manipulate peoples' lives to make them better despite not being able to directly intervene – and it's successful, too. However, nobody knows that this AI is intelligent, leading to the odd situation of having this being manipulate lives without anyone knowing it's happening. The origins of the AI are also left out of the story, so we can only guess how it came to be. However, it does pose some interesting questions about how it became intelligent. Did it somehow evolve? How did it discern its purpose? It's pretty clear in the story that this AI is both intelligent and conscious - the AI is certainly aware of itself and its surroundings, but we can't be sure if this is of its own accord or because someone made it aware.

The Maybe

Finally, a story that sends mixed messages regarding AI

"Because, if you stop to think of it, the three Rules of Robotics are the essential guiding principles of a good many of the world's ethical systems… To put it simply - if Byerley follows all the Rules of Robotics, he may be a robot, and may simply be a very good man" – Isaac Asimov, "Evidence"

consciousness is Asimov's *Evidence*. In this story, a political candidate is suspected of being a robot. However, trying to gather evidence using Asimov's three laws of robotics, scientists were unable to discern if the candidate was robotic or human. This also raises many questions – what happens when technology gets so good that we can't differentiate between humans and AI robots? Should there be some way to tell? If these robots are good, they could do good for our society. However, if they were bad, or they were compromised, they could exterminate humans.

How Can We Determine if AIs are Conscious?

Something interesting all three of the aforementioned stories have in common is the idea that consciousness is binary. Even in *Mika Model*, there was no way to describe an AI which might be somewhere in the middle – it was just undetermined whether or not the AI was conscious. In reality, science progresses much more slowly than that. We are likely to see semi-conscious, semi-intelligent, and semi-sentient AIs in the next few decades, and we're going to have to decide how to determine the possible danger and the checks and balances that each level of consciousness should have. Some AIs might also develop in different stages, or even develop automatically, which is another concern.

Bear's Queen of Angels explores this and also sets a criterion for a conscious (or self-awareness) for an AI, which is the comprehension of a standard joke. While this would never hold up in the real world, it's an interesting idea of taking something that humans understand and seeing if a machine can understand it – however, it does impose our own assumptions about our language and conventions on an AI which might not understand those.

Similar to the Queen of Angels, Alan Turing's Turing test tried to determine if an AI was actually intelligent or not by seeing if it could mimic human speech – that is, if the machine could write speech that wasn't easily distinguishable from a human's speech, it was considered intelligent. However, there were some confusing results. For example, an AI called Eliza managed to pass it by mimicking a psychologist and getting its subjects to talk a lot. It then simply reflected their answers back to them. Similar to the robot in Mika Model, this AI may not be intelligent but simply executing a very manipulative set of commands. See the "Defining Consciousness" Sidebar on Page 1 for some more information on the Turing test and other ways of determining consciousness and intelligence.

Taking a step back from science fiction literature, though, we as a species have trouble determining intelligence, and what actually signals true intelligence. As of now, the most intelligent machines are actually simply the best at fooling people. Likewise, we all have our own notions of what consciousness is - defining it, though, is much harder. Many different things could be used to determine if an AI is conscious or becoming conscious – self-awareness, internal evolution/learning (self-directed), or emotional intelligence could all be key innovations in the external (human-driven) evolution of AI.

How Should We Interact with Conscious AIs?

Moving past the stage of determining if an AI is conscious or not, how do we interact with AIs in a safe way once we determine they are conscious? There are many different theories and ideas about this concept, many of them boiling down to ethics. Mika Model tried to determine what kinds of rights a

a robot had. Was it owned by the company who made it? When it committed a crime, who actually committed the crime, the robot or the company? Does the robot deserve a lawyer? Note that these questions are made incredibly more complex by the fact that it wasn't determined whether the robot was actually conscious or not, but these questions still apply to a conscious robot.

Some authors in the field of robotics today have pushed for giving rights to robots. Others think that it is in our best interest to treat robots like pets. Where these arguments get complex is in how conscious they are. If they can simply respond intelligently but are not self-aware and have no emotions, it would probably be more morally acceptable to treat them like pets than if they were very emotionally intelligent and had interests, goals, and sentience. In addition, this situation is made more complex if the machine evolves itself and becomes emotional. Should we give it rights for something we didn't do to it? Is it truly a form of life then? Some have said that we should keep AIs in a metaphorical "box" – that is, ascribe limitations to their access to knowledge so they cannot become sentient without us purposely doing it. Without any research on the behavior of sentient AIs, though, due to the straight-up lack of them, we'll never know how they'll respond to our inputs, which is perhaps the most important thing to consider when trying to answer these questions.

What Are Some Considerations We Should Have When Dealing with Conscious AIs?

Based on some of the stark literature concerning conscious AIs, it's not a stretch to push for some type of regulation over these potentially dangerous machines. We need to take a multi-pronged approach in ensuring that we not only respect the morals we uphold as citizens towards other life but also protect ourselves, and that includes the limitation of self-evolution, the recognition of consciousness and sentience, and any appropriate restrictions we need to put on AIs of different sentience to protect us while also respecting the wishes of the AI. As with all things in life, there are balances and trade-offs to be made. But hopefully by following those steps will we only in the more dire of circumstances be unprepared enough to suffer serious consequences.

To Summarize…

The idea of birthing a new form of life is intriguing to both scientists and science-fiction writers alike, and while they have contrasting ways of evaluating the possibility of conscious artificial intelligence, they agree on certain key questions. The field is still young, which is why so many of the questions remain unanswered. What defines consciousness? How will we know if an AI is conscious? What rights do we attribute conscious AIs? These are the questions that future generations will be tackling as their dreams become reality, and we mustn't let our forward thinking go to vain. By having a plan to approach this potentially dangerous topic, we can make sure that our attention is focused on where it should be – the wonders of the quickly evolving technology that is artificial intelligence.

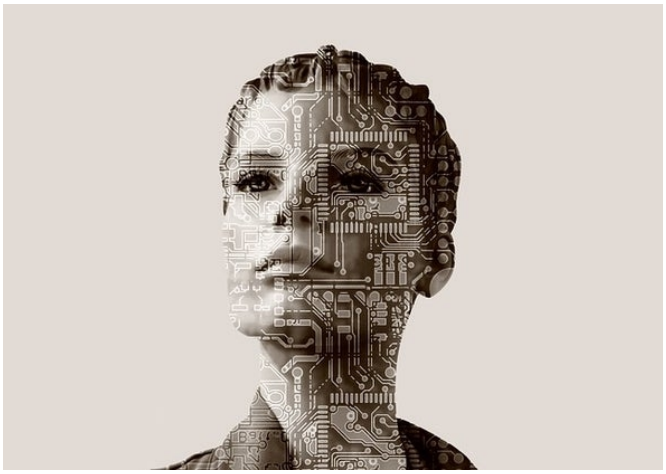

Figure 2: https://static.scientificamerican.com/blogs/cache/file/3933C6D4-58A4-48C8-B17C16959C4DB2CC_source.jpg?w=590&h=800&5F171CE8-FDF5-43C5-AC70803365F8D6A7

SOURCES

1. “AI.” Edited by John Clute et al., SFE: The Science Fiction Encyclopedia, 15 Apr. 2016, www.sf-encyclopedia.com/entry/ai.

2. Asimov, Isaac. “Evidence.” Astounding Science Fiction, 1946, pp. 121–140.

3. Bacigalupi, Paolo. “Mika Model.” Slate Magazine, 26 Apr. 2016, www.slate.com/articles/technology/future_tense/2016/04/mika_model_a_new_short_story_from_paolo_bacigalupi.html.

4. “Cybernetics.” Edited by John Clute et al., SFE: The Science Fiction Encyclopedia, 11 Aug. 2018, www.sf-encyclopedia.com/entry/cybernetics.

5. Gendler, Alex. “The Turing Test: Can a Computer Pass for Human?” YouTube. 2020, www.youtube.com/watch?v=3wLqsRLvV-c&feature=emb_logo.

6. Hildt, Elisabeth. “Artificial Intelligence: Does Consciousness Matter?” Frontiers, Frontiers, 18 June 2019, www.frontiersin.org/articles/10.3389/fpsyg.2019.01535/full.

7. Kritzer, Naomi. “Cat Pictures Please.” Clarkesworld Magazine, Jan. 2015, clarkesworldmagazine.com/kritzer_01_15/.

8. Schneider, Susan. “Is Anyone Home? A Way to Find Out If AI Has Become Self-Aware.” Scientific American Blog Network, Scientific American, 19 July 2017, blogs.scientificamerican.com/observations/is-anyone-home-a-way-to-find-out-if-ai-has-become-self-aware/.