If a robot is conscious, is it OK to turn it off? The moral implications of building true AIs

Anand Vaidya

In the "Star Trek: The Next Generation" episode "The Measure of a Man," Data, an android crew member of the Enterprise, is to be dismantled for research purposes unless Captain Picard can argue that Data deserves the same rights as a human being. Naturally the question arises: What is the basis upon which something has rights? What gives an entity moral standing?

The philosopher Peter Singer argues that creatures that can feel pain or suffer have a claim to moral standing. He argues that nonhuman animals have moral standing, since they can feel pain and suffer. Limiting it to people would be a form of speciesism, something akin to racism and sexism.

Without endorsing Singer's line of reasoning, we might wonder if it can be extended further to an android robot like Data. It would require that Data can either feel pain or suffer. And how you answer that depends on how you understand consciousness and intelligence.

As real artificial intelligence technology advances toward Hollywood's imagined versions, the question of moral standing grows more important. If AIs have moral standing, philosophers like me reason, it could follow that they have a right to life. That means you cannot simply dismantle them, and might also mean that people shouldn't interfere with their pursuing their goals.

Garry Kasparov was beaten by Deep Blue, an AI with a very deep intelligence in one narrow niche. Stan Honda/AFP via Getty Images

Two flavors of intelligence and a test

IBM's Deep Blue chess machine was successfully trained to beat grandmaster Gary Kasparov. But it could not do anything else. This computer had what's called domain-specific intelligence.

On the other hand, there's the kind of intelligence that allows for the ability to do a variety of things well. It is called domain-general intelligence. It's what lets people cook, ski and raise children – tasks that are related, but also very different.

Artificial general intelligence, AGI, is the term for machines that have domain-general intelligence. Arguably no machine has yet demonstrated that kind of intelligence. This summer, a startup called OPENAI released a new version of its Generative Pre-Training language model. GPT-3 is a natural-language-processing system, trained to read and write so that it can be easily understood by people.

It drew immediate notice, not just because of its impressive ability to mimic stylistic flourishes and put together plausible content, but also because of how far it had come from a previous version. Despite this impressive performance, GPT-3 doesn't actually know anything beyond how to string words together in various ways. AGI remains quite far off.

Named after pioneering AI researcher Alan Turing, the Turing test helps determine when an AI is intelligent. Can a person conversing with a hidden AI tell whether it's an AI or a human being? If he can't, then for all practical purposes, the AI is intelligent. But this test says nothing about whether the AI might be conscious.

Two kinds of consciousness

There are two parts to consciousness. First, there's the what-it's-like-for-me aspect of an experience, the sensory part of consciousness. Philosophers call this phenomenal consciousness. It's about how you experience a phenomenon, like smelling a rose or feeling pain.

In contrast, there's also access consciousness. That's the ability to report, reason, behave and act in a coordinated and responsive manner to stimuli based on goals. For example, when I pass the soccer ball to my friend making a play on the goal, I am responding to visual stimuli, acting from prior training, and pursuing a goal determined by the rules of the game. I make the pass automatically,

without conscious deliberation, in the flow of the game.

Blindsight nicely illustrates the difference between the two types of consciousness. Someone with this neurological condition might report, for example, that they cannot see anything in the left side of their visual field. But if asked to pick up a pen from an array of objects in the left side of their visual field, they can reliably do so. They cannot see the pen, yet they can pick it up when prompted – an example of access consciousness without phenomenal consciousness.

Data is an android. How do these distinctions play out with respect to him?

Do Data's qualities grant him moral standing? CBS

The Data dilemma

The android Data demonstrates that he is self-aware in that he can monitor whether or not, for example, he is optimally charged or there is internal damage to his robotic arm.

Data is also intelligent in the general sense. He does a lot of distinct things at a high level of mastery. He can fly the Enterprise, take orders from Captain Picard and reason with him about the best path to

take.

He can also play poker with his shipmates, cook, discuss topical issues with close friends, fight with enemies on alien planets and engage in various forms of physical labor. Data has access consciousness. He would clearly pass the Turing test.

However, Data most likely lacks phenomenal consciousness - he does not, for example, delight in the scent of roses or experience pain. He embodies a supersized version of blindsight. He's self-aware and has access consciousness – can grab the pen – but across all his senses he lacks phenomenal consciousness.

Now, if Data doesn't feel pain, at least one of the reasons Singer offers for giving a creature moral standing is not fulfilled. But Data might fulfill the other condition of being able to suffer, even without feeling pain. Suffering might not require phenomenal consciousness the way pain essentially does.

For example, what if suffering were also defined as the idea of being thwarted from pursuing a just cause without causing harm to others? Suppose Data's goal is to save his crewmate, but he can't reach her because of damage to one of his limbs. Data's reduction in functioning that keeps him from saving his crewmate is a kind of nonphenomenal suffering. He would have preferred to save the crewmate, and would be better off if he did.

In the episode, the question ends up resting not on whether Data is self-aware – that is not in doubt. Nor is it in question whether he is intelligent – he easily demonstrates that he is in the general sense. What is unclear is whether he is phenomenally conscious. Data is not dismantled because, in the end, his human judges cannot agree on the significance of consciousness for moral standing.

When the 1s and 0s add up to a moral being. ktsimage/iStock via Getty Images Plus

Should an AI get moral standing?

Data is kind – he acts to support the well-being of his crewmates and those he encounters on alien planets. He obeys orders from people and appears unlikely to harm them, and he seems to protect his own existence. For these reasons he appears peaceful and easier to accept into the realm of things that have moral standing.

But what about Skynet in the "Terminator" movies? Or the worries recently expressed by Elon Musk about AI being more dangerous than nukes, and by Stephen Hawking on AI ending humankind?

[*Deep knowledge, daily*. Sign up for The Conversation's newsletter.]

Human beings don't lose their claim to moral standing just because they act against the interests of another person. In the same way, you can't automatically say that just because an AI acts against the interests of humanity or another AI it doesn't have moral standing. You might be justified in fighting back against an AI like Skynet, but that does not take away its moral standing. If moral standing is given in virtue of the capacity to nonphenomenally suffer, then Skynet and Data both get it even if only Data wants to help human beings.

There are no artificial general intelligence machines yet. But now is the time to consider what it would take to grant them moral standing. How humanity chooses to answer the question of moral standing for nonbiological creatures will have big implications for how we deal with future AIs – whether kind and helpful like Data, or set on destruction, like Skynet.