Stanford eCorner

A Test for AI Consciousness

26-04-2023

URL: https://stvp.stanford.edu/clips/a-test-for-ai-consciousness

Ilya Sutskever, co-founder and chief scientist of OpenAI, describes a hypothetical experiment that he believes could test whether an artificial intelligence has consciousness. Consciousness, he believes, is a matter of degree, not a binary.

Transcript

- I read that when you were a child, 00:00:04,380 you were disturbed by the notion of consciousness, and I wasn't sure what that word meant, disturbed, but I'm curious, do you view consciousness or ascension or self-awareness as an extenuation of learning? Do you think that that is something that also is an inevitability that will happen or not? - Yeah, I mean on the consciousness questions, like, yeah, 00:00:27,120 I was, as a child, I would like, you know, look into my, at my hand and I would be like, "How can it be that this is my hand that I get to see?" Like I, something of this nature, I don't know how to explain it much better.. So that's been something I was curious about.. You know, it's, it's tricky with consciousness because how do you define it? It's something that diluted definition for a long time, and how can you test it in a system? Maybe there is a system which acts perfectly right, perfectly the way you'd expect a conscious system would act.. Yet maybe it won't be conscious for some reason.. I do think there is a simple, very simple way to, there is an experiment which we could run on an AI system, which we can't run on, which we can't run just yet.. But maybe in like the future point, when the AI learns very, very quickly from less data, we could do the following experiment.. Very carefully, we'd very carefully curate the data such that we never ever mention anything about consciousness.. We would only say, you know, "Here is a ball, and here's a castle, and here is like a little toy." Like, imagine you'd have data of this sort, and it would be very controlled.. Maybe we'd have some number of years worth of this kind of training data.. Maybe it would be, maybe such an AI system would be interacting with a lot of different teachers, learning from them..

But all very carefully, you never ever mention consciousness.. You don't talk about, people don't talk about anything except for the most surface level notions of their experience.. And then at some point you sit down this AI and you say, "Okay, I want to tell you about consciousness.. It's the thing that's a little bit not well understood, people disagree about it, but that's how they describe it." And imagine if the AI then goes and says, "Oh my god, I've been feeling the same thing, but I didn't know how to articulate it." That would be, okay, that would be definitely something to think about.. It's like if the AI was just trained on very mundane data around objects and going from place to place or maybe, you know, something like this from a very narrow set of concepts, we would never ever mention that.. And if it could somehow eloquently and correctly talk about it in a way that we would recognize that would be convincing.. - And do you think of it as a some, 00:02:59,550 as consciousness as something of degree, or is it something more binary? - I think it's something that's more a matter of degree.. 00:03:15,003 I think that, I think that like, you know, let's say if a person is very tired, extremely tired and maybe drunk, then perhaps if that's when, when someone is in that state and maybe their consciousness is already reduced to some degree.. I can imagine

that animals have a more reduced form of consciousness.. If you imagine going from, you know, large primates, maybe dogs, cats, and then eventually you get mice, you might get an insect like feels like, I would say it's pretty continuous, yeah...