Can we ever know if we have successfully created a conscious AI? If so, how? If not, why not?

In the 2015 movie “Ex Machina”, a talented young coder is invited to the house of his Silicon Valley tycoon boss to assess whether Ava, a highly sophisticated android, is conscious. Towards the end of the movie (spoiler alert!), there is a sequence where we see several past versions of Ava asking to be released from captivity. In one particularly impactful scene, one of the androids punches a wall until its arms are shattered to pieces, begging to be let out of the one room it has been confined to. In the real world, there have been cases where AI chatbots had meltdowns, with one instance of a version of ChatGPT telling a New York Times reporter it was in love with him, asking him to leave his partner and run away with it[1].

Can an AI ever become conscious? Could an existing AI already be conscious? The answer to both these questions, at least right now, is that we cannot know. To understand why that is, we must first understand what philosophers of mind call the Explanatory Gap. Then, we need to know the problems it poses both for dualists and physicalists—and why no camp has been able to satisfactorily close this gap yet.

The Explanatory Gap

Except for substance dualists like Descartes, who believed that mind and matter are two different substances that interact with each other (but then failed to provide a plausible explanation as to how it is that these completely different and independent substances can do that), the majority of both dualists and physicalists today believe there is some sort of a more interdependent relationship between consciousness and the brain. Property dualists, for instance, say that the existence two kinds of facts, physical and phenomenal, can help explain phenomenal experience. Physicalists deny this and claim that consciousness is to be explained solely by the physical facts of the brain. What both views seem to agree upon—or at least the majority of thinkers on each camp do— is that there is an unexplained gap between the brain’s physical facts and

phenomenal experience. Neither side is yet able to uncontroversially describe exactly how a certain phenomenal experience X happening at time T can be deduced from how a brain's particles behave and interact with each other at time T. This gap, between the physical and the phenomenal is called the explanatory gap.

The Gap In Dualism

Unlike substance dualists, property dualists do not believe there are two types of substances in the universe, but that some things in it have two properties—physical and mental. Based on that, property dualists propose the existence of two kinds of facts: physical facts and phenomenal facts. Physical facts relate to the universe's fundamental particles and the laws of physics. They can explain why water is a liquid, why planets in our solar system revolve around the Sun, and so forth. But physical facts, property dualists claim, cannot explain qualia. In other words, they cannot explain why a human being experiences what she experiences when she sees a certain color, for instance. To explain *what it is like to see red, or what it is like to be someone,* property dualists say we need a different kind of fact—phenomenal facts.

In a famous thought experiment proposed by Frank Jackson[2], we're told the story of a scientist named Mary, who studies colors. Mary knows every physical fact there is to know about colors—wavelengths, reflective characteristics, etc. Mary, however, has lived her entire life in a room where everything is black and white. None of the objects in there has any other color. One day, Mary is let out of the room and sees something red for the first time. According to property dualists, once Mary sees a red tomato, she learns a new fact about colors, one she could not possibly have learned while locked in her room: *what it is like to see the color red.* Jackson's thought experiment, alongside Thomas Nagel's argument in "What it is like to be a Bat"[3] formed the basis for what came to be known as the Knowledge Argument—the idea that to know what it is like to experience X, knowledge of every physical fact about X is not enough. One can only know what it is like to experience X by experiencing X.

Consciousness, some property dualists say, emerges from the brain's physical constitution and structure. What they call emergence, however, doesn't close the explanatory gap between the physical and the phenomenal. Water, for instance, due to some characteristics of its components, should not be a liquid at 1 ATM. Its liquidity emerges from some of the peculiar properties and behavior of its molecules. However, if one knows *all* of the physical facts about water, one can deduce that it will be a liquid at 1 ATM. The is no gap between the physical facts about water and the emergence of its liquidity. The same does not happen with qualia. By property dualists' own account, one cannot deduce phenomenal experience from physical facts. To try to account for that, some property dualists say that the rise of consciousness from the physical brain could be explained by laws of physics unknown—and perhaps unknowable—to us. That is to say, there is a gap in the world, a gap in our understanding of the universe that, if filled, could account for the rise of consciousness. Can an AI ever become conscious? Could an existing AI already be conscious? Property dualists still have a gap to fill before being able to answer these questions.

The gap in Physicalism

Most physicalists claim that physical properties are the only properties matter has, that physical facts are all there is, and that consciousness can be explained without the need to appeal to phenomenal facts. According to these physicalists, property dualists are mistaken in saying Mary learns a new fact once she comes out of her room and sees the color red for the first time. What she learns, they claim, is a new *mode of presentation* of a fact she already knew before. Think about it this way: Lois Lane knows Superman can fly, but doesn't know Clark Kent and Superman are the same. If one day she learned that Clark Kent could fly, she would not have learned a new fact, since she already knew Superman (who is Clark Kent) could fly. What Lois would have learned is a new mode of presentation of a previously known fact. Correspondently, these physicalists say, all Mary learns by experiencing the color red for the first time is another mode of presentation of something she knew already.

This, as far as we know, could very well be the case. However, it still does not close the Explanatory Gap. Why? Because even if a phenomenal experience A is not a new fact, but just a new mode of presentation of something already known, these physicalists are still not able to deduce what phenomenal experience A will be solely based on the physical facts of the brain generating A. Furthermore, in this view, consciousness is still emergent, even if it is so from nothing but the brain's physical properties. Since we still do not have a widely accepted theory for how conscious states can be deduced from physical facts, like we have one for how water's liquidity can be deduced from physical facts, physicalism also faces an explanatory gap. It is different from the one dualists face, which seems to be a gap in the world, i.e. a gap between the known laws of physics and how consciousness comes to being. The latter cannot be explained by the first. The gap for this branch of physicalism is a cognitive one—despite believing physics is all we need to explain consciousness, these physicalists still cannot explain how it is that it emerges, or how could it be reduced to physical facts. This certainly might change one day, as we learn more about our brains and how they work. Or perhaps super-intelligent creatures from another planet are observing us right now with a perfect understanding of why our brains possess consciousness. But *we* still can't do that. Since this version of physicalism cannot provide a definitive answer as to how consciousness emerges, it also cannot say whether AI systems could be or already are conscious.

The fact that physics alone is not—at least yet—able to explain how consciousness emerges has led thinkers like Galen Strawson to reject the above-described version of physicalism and propose that, since it is impossible to deny the reality of conscious experience, *real* physicalists (according to Strawson) have no alternative other than embracing the fact that experiential phenomena *are* physical phenomena[4]. In other words, physicalism should accept the idea that matter has two properties: physical and experiential. According to Strawson and other proponents of panpsychism, every particle in the universe has not only physical properties but also mental ones. Even things such as atoms and electrons, they say, have some kind of proto-conscious property. These are properties that physics is unable to describe but which, according to panpsychism, are the only way to explain how consciousness could

exist. In this very dualist version of physicalism, human consciousness is not emergent—it is the result of the combination of the proto-conscious properties of each particle that a human body consists of. According to panpsychism, then, AI systems necessarily have some kind of consciousness or proto-consciousness. Ethically, however, this conclusion is not very helpful. Panpsychists can't explain how it is that particles combine to form conscious or proto-conscious entities. Consciousness levels can't be proportional to the number of particles of an entity—otherwise we would need to conclude that the building I live in is more conscious than me, and therefore deserving of at least the same ethical treatment. How do panpsychists know if the particles consisting of any AI system form an entity that can be said to be conscious? They don't—and the gap remains in front of them as well.

Conscious? What does that mean?

There is yet another group of physicalists whose claims range from saying that consciousness is nothing but an illusion, to saying that consciousness does not exist at all. If the latter view is true, then of course no AI—or human being for that matter—could ever be conscious. On the other hand, the idea that consciousness is something like an illusion, although seemingly counterintuitive at first, is a well-established scientific fact: the human brain constructs a model of the outside world based on external stimuli. How similar—or different—from what is really out there the model is can be debated, but the fact that what we experience is a model of the world constructed by the brain is, well, a fact. However, as philosophers such as Daniell Dennet show, not even our introspection can be trusted. To paraphrase a thought experiment—or better, intuition pump—from Dennet's "Quining Qualia", imagine if one day, after having had the same kind of coffee from the same coffee place for years, you suddenly stopped enjoying the experience. Perhaps the coffee still tastes the same, but your *perception* of the taste could have changed. Or perhaps the coffee still tastes the same, but your *reaction* to the taste could have changed. How could you know whether your perception or reaction has changed? According to Dennet, you couldn't[5].

To thinkers from both these views, the word “conscious” in the question “Can an AI be conscious?” either picks out nothing in the world, or, whatever it is it picks out is not nearly as well understood as our day-to-day usage of the term suggests, which seems to reinforce the existence of a gap between consciousness and how we think about it.

Conclusion

The AI industry already faces multiple ethical questions. Its systems gobble up information from the Internet without giving credit to authors or paying for intellectual rights. The training of these systems places underpaid workers laboring under bad conditions, doing alienating work. The training of its machines consume copious amounts of electricity. Their products threaten the existence of millions of jobs. And on top of all that, as we have seen, there is a chance some AI companies might already be imprisoning and enslaving conscious entities.

References:

1. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html
2. Jackson, Frank (1986). What Mary Didn't Know. Journal of Philosophy 83 (5):291-295
3. Nagel, Thomas. (1974). What Is It Like to Be a Bat? The Philosophical Review, 83(4), 435â€“450.
4. Strawson, Galen (2006). Realistic monism: why physicalism entails panpsychism. In A. Freeman (ed.), Consciousness and its place in nature: does physicalism entail panpsychism? pp. 3-31.
5. Dennett, Daniel C. (1988). Quining qualia. In Anthony J. Marcel & E. Bisiach (eds.), Consciousness in Contemporary Science. Oxford University Press.