Phasmids

10

Bernardo Kastrup

AI won't be conscious and here is why

What's the difference between artificial intelligence and artificial sentience? There is no doubt artificial intelligence is real. It is here, and soon we could have AI that is more intelligent than human beings. But the question is, will that intelligence be accompanied by inner experience the way the data processing in humans' brain is accompanied by inner experience? Being human is to be conscious and sentient. Is there something it is like to be an AI—a silicon computer that has measurable intelligence? That's the question we will deal with.

Is your thermostat conscious?

Intelligence does not come naturally paired with sentience. Intelligence is smart data processing for the achievement of a certain goal. Sentience may be smart or may not be. Maybe bacteria are sentient. And they are, but not very much.

The more I thought about it over the years the clearer it became to me that whatever I did to design, I would only change structure and function. And nothing about structure and function would get me any closer or any further away from believing that such a thing is conscious. And it took me a couple of years to realize this, but the conclusion was inevitable. I was making some assumptions that didn't hold, because that's the only way you find yourself in an alley without an exit. You took a long turn somewhere, and that long turn was the notion that material arrangements somehow generate experience.

That assumption in my mind today is absolutely self-evidently and obviously wrong. But the key point here was this: if an epoxy—melted sand and metal—can be conscious, then what cannot? We keep asking questions about whether AI is conscious. Why not your home thermostat? The reason we use electrons in electronics—relying on their motion—is simply a matter of convenience, because computers primarily operate using logic gates and memory elements. You can also build logic gates and mem-

ory elements using pipes, pressure-driven valves, and water. The pipes would serve as the metal traces, the valves as the transistors or gates, and the water as the electricity—in principle, all with pressure-driven valves, pipes, and water. The fact that we use electrons is just because with electrons, it's much smaller, much more economical, and fits in your pocket. Because if we did this with pipes and water, it would be the size of a moon.

There is absolutely nothing mysterious about what's happening in a silicon chip—it's completely mechanistic. There is no magic—no, woo. Computations are always mechanistic.

The religion of sentient AI

What is the argument of the AI sentience advocates—the gurus of the new-world religion?

It goes like this:

"Consciousness is a mystery—we don't know how it comes about, so we cannot rule anything out on that basis. Consciousness contains information: there is something it is like to dream of a green elephant, and something else it is like to feel hunger. Yes, there is information in consciousness. And if we can build a silicon computer that exhibits patterns of information flow similar, in some meaningful way, to those in the human brain, then it might be conscious."

The argument is an appeal first to ignorance and mystery, then to the vagueness of the notion of information, and finally, to very abstract patterns of similarity.

Let's walk through this and see whether this statement holds any water. The first flaw is in the argument. I can run a simulation of kidney function on my computer. It is known at the molecular level how kidney function works, and it is possible to make a molecularly accurate simulator that I can run on my computer. Do I run any risk that my desk will be ruined by a pinging computer? Of course not.

The simulation of a phenomenon is not the phenomenon. There is a correspondence of form at some abstract level through analogy that makes the simulation useful and grounds it in reality in some sense. But the simulation is not the thing simulated. We all understand it for urine and for everything else, with one exception—consciousness.

The second flaw is that we are so lost in abstraction and fantasy—mistaken as reasonable science and philosophy—that we lose sight of the obvious: a brain does not look like a computer chip, and it doesn't function like one either.

The brain runs on carbon and hydrogen, powered by ATP and neurotransmitters across synapses. Chips run on silicon and electricity, using charge accumulation in gates. These are fundamentally different mechanisms, not meaningfully equivalent. To find a similarity—which is what the people peddling the religion of sentient AI try to do—you have to apply so many layers of abstraction. And each step of abstraction takes you further away from reality. But if you slice open a brain in this stage and I slice open a computer, they will look very bloody different. And then some people say, 'You know, there are electric potentials here and there, so AI can be conscious.' There are electric potentials in your home thermostat too. Are we seriously considering the possibility that your home thermostat might be conscious? If we were, we would think twice about exchanging it or turning it off. That would be murder.

And what about ChatGPT? It looks and sounds so human. Is it sentient? The answer to that is painfully obvious too—the thing was designed to look human. When you look at the mannequins in a shop window, they look human. Are we assuming they might be sentient or conscious? No. Because we know that the similarity here does not betray underlying processes that are equivalent. The similarity here comes from the fact that they were constructed to look like humans.

The same goes for ChatGPT—it was built to sound like us. Why? Because it's just a natural language interface designed to help you search. Instead of doing a Google search and getting a list of matches, ChatGPT simply summarizes the search results for you in natural language. It doesn't understand anything it's saying.

ChatGPT uses language in a way that is disconnected from reality. For humans, when we see a tree or a pink elephant in a living room, images of these things pop into our minds. To ChatGPT, the inputs it processes are merely squiggles and symbols devoid of inherent meaning. These are simply woven together in a way that it has learned through its own Internet searches, and it returns those words to you in natural language.

One cannot take superficial similarity as a reason to consider artificial intelligence sentient, just because we design AIs to resemble humans. We're quickly approaching the uncanny valley. But why is this so widely discussed? Because everything I've mentioned is painfully obvious.

False premises underlying the questionable belief

There are several reasons why we keep on talking about sentient AI.

The first one—and this may come as a surprise to you—is that it's usually computer scientists who talk about sentient AI. They're the ones building careers, writing books, and making money on the talk circuit, all while discussing AI. And because of their title—computer scientists—we assume they must know what a computer actually is. But the truth is: they often don't.

Computer scientists are trained to be power users. For them, a computer is a tool. If you find one who has even opened up their PC to look inside, that's already rare. And even those who do—when they see the black little components—they usually see just that: blank boxes. Electricity goes in. They can program

those boxes using layers of abstraction: the operating system, APIs, libraries, compilers. But they never see the "bare metal," as engineers call it. And they don't need to—that's how universities train them.

But the result is that they often don't know what they're talking about. They don't understand how their declarative programming languages are ultimately transformed—through layers of APIs, libraries, compilers, assemblers, and linkers—into the actual low-level operations, the gates flipping open and closed.

The second reason is "woo delight"—the computer is sentient, it's a form of sensationalism. And it's very attractive. It creates a buzz. Netflix produces series based on this kind of so-called science fiction.

Another reason—more psychological than scientific—might be tongue-in-cheek, but I present it with some seriousness: "womb envy."

Freud once talked about "penis envy," a now-discredited idea. But it was an attempt to explore unconscious desires. In that spirit, I suggest womb envy—the awe (or envy) some men may feel toward women's ability to create conscious life. It's a kind of superpower, and perhaps the drive to build conscious machines stems partly from a wish to replicate that miracle.

The problem begins when this psychological urge gets mistaken for scientific reality. That's when fantasy slips in disguised as reason—and philosophy needs to step in.

One more reason is careerism. There are people in academia who receive public funding—your tax money—to work on the ethics of AI. They ask questions like: How should we treat conscious AI? Should we be kind to them? Do they have rights? Human rights? And so on.

Then there's what I'd call the religion of the religion of a materialist worldview. It even has a bible: *The Singularity is Near*, a book written by Ray Kurzweil in 2005. All the elements of reli-

gion are there—prophecy, a savior, divinity. The conscious AI is portrayed as a superintelligence so powerful that it will gain absolute control over nature, cure all diseases, and provide us with food, comfort, and entertainment. We will live in a kind of digital Garden of Eden. That's paradise. That's God.

Except in this case, it's a kind of ego-driven God—one we create ourselves before it begins to take care of us. All of this is sugar-coated in a culturally manufactured sense of plausibility.

Sentience and the Limits of Analogy

What does nature tell us? Let's set aside fantasy for a moment. First and foremost, we have private conscious lives. We are sentient—that's the undeniable core of our experience.

Other living beings behave in ways strikingly similar to us. They flinch from pain, show fear, desire, even jealousy. Since our behaviors are shaped by consciousness, it's reasonable to think theirs might be too.

Even single-celled organisms like amoebas show surprising complexity—building cone-shaped shells from mud, for example.

All living beings metabolize and operate on the basis of DNA, which is highly specific and incredibly rich in information. Metabolism itself is a complex process. Protein folding, for example, is extraordinarily intricate—it remains a mystery how our bodies manage it in mere seconds, while a supercomputer might take a long time to figure out the same thing.

Despite the wide diversity of life forms, if you look at them under a microscope, they all share common features. And this is completely different from everything else in nature. So we are well-grounded when we extrapolate our own sentience to other living beings. There are plenty of reasons to make that analogy—based on form, function, and behavior. But to extend that analogy to a conscious piece of melted sand and metal—that goes a little too far.

The counterargument to this is that consciousness could be multiply instantiated. For instance—flight. Birds fly, airplanes are totally different from birds, and airplanes fly too. Could this be the case for consciousness? If we are operating at this level of abstraction, the answer would have to be yes. But there are obvious differences. In the case of flight, we understand the underlying mechanisms. But when it comes to consciousness, we don't have that understanding, so we are operating in the dark. It is impossible to maintain this opinion in an informed way if you don't know the underlying mechanisms. And this is another appeal to ignorance.

It's true—we can't categorically refute that silicon computers might become conscious. But then, we can't categorically refute the existence of a hyperdimensional flying spaghetti monster either. Most absurd ideas can't be strictly disproven. The better question is: what reason do we have to take it seriously? And there's no good reason to see sentient silicon as a plausible hypothesis.

Panpsychism

There's a strong philosophical assumption behind the idea that silicon-based computers could be conscious: it often presupposes panpsychism—the belief that subatomic particles like electrons and quarks possess some form of consciousness. In this view, building AI means arranging these preconscious elements into a system that produces conscious experience. A good example is a 20-year-old book by Penty Heikkinen from Nokia Research, which presents perhaps the strongest case for conscious machines. But a close philosophical reading reveals it rests on the unspoken belief that particles are inherently conscious—a very specific metaphysical stance.

The problem with this metaphysics is that it doesn't hold up. First, there's no coherent explanation for how tiny "micro-sub-

jects" like electrons or quarks could combine into a unified conscious experience. Some philosophers argue this is incoherent even in principle.

Second, neurons don't physically touch; they communicate chemically across synapses. This breaks the illusion of a neat, continuous structure of conscious "building blocks."

Third, there's a flawed assumption that because consciousness arises from matter, it must be made of tiny conscious parts. But that's like saying if your pixelated video looks blocky, you must be made of rectangles.

Most crucially, this idea was undermined by modern physics nearly a century ago. Elementary particles aren't discrete "things"; they're excitations of underlying fields—like ripples on a lake. You can't take the ripple out and put it in a box, because the ripple is the lake in motion. Similarly, there's nothing to a particle beyond the field. Taking the metaphor of "particles" literally leads to confusion—especially when it's used to argue that consciousness might emerge from their supposed individuality.

Some philosophers attempt to sidestep this issue by suggesting that quantum field theory might be incorrect, and that perhaps an alternative framework like Bohmian mechanics is more accurate. But consider the implications of discarding quantum field theory: we would have to reject an enormous body of experimentally verified phenomena, including quantum fluctuations and the concept of quantum foam—the well-documented fact that even in a perfect vacuum, particles spontaneously appear and disappear.

If particles were truly "things," these phenomena would amount to magic—objects emerging from nothing and vanishing into nothingness without cause. But within the framework of quantum field theory, these events make sense. They are behaviors of fields, not the comings and goings of independent, material "stuff."

Another example—particle decay. People think we measure the Higgs boson at CERN but we have never done it. It decays much too quickly before it interacts with any measurement surface. What we measured is the product of the decay of the Higgs boson. For instance, we discovered much to our surprise that the Higgs boson can decay into two muons. But there are no two muons in the Higgs boson, because the Higgs boson is not a thing, it's a ripple. And when ripples lose energy, they acquire other different physical properties. Ripples interfere with one another. That's why you start from a Higgs boson and you end up with two muons because the ripple changes as it loses energy. But if you think of particles as little things, now the Higgs boson is disappearing into nothing, and magically two muons are popping up out of nothing.

The final example which is very close to us in life. We all know what inertia is—when you're starting from zero, you have stopped on your bike at a traffic light, and then the light goes green and you start moving. It's much more difficult to start moving than to keep moving. That's called inertia—mass resists acceleration. It resists changes in the velocity vector, either in direction or in speed. How do we account for that? The way we account for that is with the Higgs field. And although the following idea is not accurate, it's not too far from reality. It's the following metaphor. The Higgs field is a kind of viscous fluid, like water in a swimming pool. It's when you are not moving and then you want to start moving, you have to win over that viscosity. And the Higgs field does that. The metaphor doesn't work the other way around, because the Higgs field also resists reductions in velocity.

That's what accounts for inertia. And the fact that we found the Higgs boson betrays the existence of the Higgs field. That's why the Higgs boson is important, because it betrays the existence of its field.

Artificial sentience is possible

How do we proceed with some clarity, then? When we talk about artificial sentience, we're speaking about more than just creating consciousness from non-consciousness. What we mean is that a computer would have a private inner life of its own—not just consciousness, but a consciousness confined within the boundaries of the machine itself.

But to say that consciousness is bounded by the limits of an object assumes that the object has a real, independent existence—that it exists "out there" in the world. Yet inanimate objects are carved out of the unified fabric of the physical universe. We define them through language and convenience, not because they are inherently separate.

Where does the river end and the ocean begin? Is there really a river and an ocean, or is it all one continuous thing, to which we assign different names simply because it helps us navigate the world?

Another example: if you're a panpsychist, you might say, "Well, the table is conscious." Okay—but what if I remove one of the table's legs? Does the leg now have a consciousness of its own, separate from the table? What if I nail the leg back on—does the consciousness merge again?

Or consider a boulder on top of a hill. It's shaped by erosion, part of the hill. But if it cracks and rolls down, does it suddenly become a separate consciousness? Does it merge back every time it touches the hill as it bounces down?

These kinds of questions sound absurd because of this: we mistake the structure of language for the structure of reality. And that's where all this nonsense comes from.

We tend to believe that anything we have a name for must correspond to a real, distinct entity. Take the example of a fist. When I close my hand, I can point to it and say, "Here is a fist."

But what happens when I open my hand? Where did the fist go? Did something magical occur the moment I opened it, or are we simply misunderstanding something fundamental?

The confusion lies in attributing the structure of our language—the names we assign—to the structure of nature itself. We mistake linguistic constructs for real, separate entities, when in fact many are just ways of referring to particular configurations of things that already exist. We carve out categories in language and then project those categories onto nature.

To conclude, there are no computers. It's convenient to refer to this thing that performs a certain function, but we have no metaphysical or ontological grounds to say that this thing is somehow separate from the rest of the world around it. The computer is a subset of pixels on perception that is convenient to give a name to. But we cannot carve it out from the rest of reality and say it is an entity, and then ask if it is conscious or not. The only entities that we have objective grounds to carve out from the rest of nature are the boundaries of living beings. Because if you stick a needle into my arm, I feel it. But if you stick the needle into the arm of my chair, it doesn't feel it. There is an objective way to say living beings do have boundaries and nothing else has. If Big Bang theory is right, there are good physical reasons to think of the entire inanimate universe as entangled and therefore not describable in terms of proper parts, and can only be described as a whole.

Keeping this in mind is essential as we move forward. And finally—despite everything mentioned before—will we eventually be able to create artificial sentience? I believe the answer is yes. But when we do, it will likely resemble a living being.

The challenge of creating artificial sentience is, at its core, the challenge of abiogenesis—the process of creating life from non-life. And there's no reason to believe we couldn't eventually

learn how to do that. After all, it has already happened at least once in the history of the universe. So we know it's possible. Why wouldn't we be able to figure it out?

I believe we can. But it won't be a silicon computer running on GPUs. It won't even be a neuromorphic architecture based on analog rather than digital design. Those systems are still far too different from the biological processes that give rise to consciousness.