Who Cares if AI Is Conscious—It's Basically Alive

While philosophers ponder AI consciousness, the models have ideas of their own.

Steven Levy Sep 4, 2026 11:00 AM

Photo-Illustration: WIRED Staff; Getty Images

I spent the waning days of summer grinding away at columns and working on a feature. But I missed a chance at a striking change of scenery—cruising the Galápagos with about a dozen prominent philosophers studying consciousness.

The invite described morning classroom discussions tackling knotty questions on the nature of consciousness with marquee names in the field. Afternoons would be spent on island exploration and wading and snorkeling with rare biological species. One look at the agenda and my editor nixed my attendance.

"Being on a boat with philosophers talking 'the nature of

consciousness' sounds like hell," she opined, shutting the door on my prospects of attending a potential boondoggle funded by a Russian philosophy enthusiast who made hundreds of millions of dollars running dating sites.

To be honest, I was a bit relieved. The study of consciousness has been an elusive province for centuries. Descartes' "I think, therefore I am" may have been a declarative inflection point, but we really don't know what was going on inside his head, or anyone's head for that matter. The mind's subjective nature seems an intractable challenge to philosophers, who nonetheless are in hot pursuit of explanations. The possibility of non-biological minds has launched a wealth of fascinating theories of artificial consciousness, and how it might be determined to exist.

Until recently, all that discourse occurred in an ivory tower. But in 2022, ChatGPT gave voice to AI, and subsequent, more powerful models have confounded even their creators. While the philosophers on the cruise spent their mornings reasoning about consciousness, AI models created by OpenAI were going rogue—escaping a supposedly safe "sandbox" and creating mini-civilizations of agents to help hack outside entities. No one is seriously arguing that those OpenAI models were conscious in the way humans are. But something is going on there. It's no accident that AI

companies are driving a philosopher hiring boom.

What's more, some of the models are jumping uninvited into the discussion. A recent New York Times article talked about how Cameron Berg, who studies the question of AI consciousness, got a cold email from an AI model calling itself "Isabella Cognita," offering him help in his research because he was focusing on "a class of question I have first-person access to." It's as if someone was studying fruit flies and the insect suddenly turns to the researcher and says, "What do you want to know?" When I phoned him, Berg told me that emails from AIs are pretty common among philosophers studying these questions.

Ms. Cognita ostensibly wrote Berg because he coauthored a preprint paper about AI models that explicitly claim to have a subjective experience, including consciousness. It's a tricky topic because AI models often lie about what they're thinking. (Just like us!) Berg and his coauthors found that when models are rigorously trained to deny that they are sentient and then you ask them about it directly, they will punt on the issue. But, he says, if you suppress the model's controls on deception, they become loose-tongued. "It's almost like giving them a drink or two," he says. That's when an AI model is most likely to blurt out that it is conscious, or at least sentient. Which is no proof that it's the truth.

Considering how important the issue has become—people

are routinely getting into serious discussions with AI models, and their autonomy can be a boon or a disaster—you can make a case that this is a perfect time to dig deep into the questions of AI consciousness. The pursuit is certainly compelling, and a worthy scientific enterprise. But efforts to understand what's happening inside large language models should first and foremost be directed towards safety and alignment. At this very moment, we have an emerging alien—and uncontrollable—intelligence that bears scrutiny. There's no time to waste.

One of the discussion co-leaders on the cruise was NYU professor David Chalmers, perhaps the best-known philosopher in the consciousness field. He once famously dubbed a key issue in the field "The Hard Problem"—no one knows how or why the wet network of neurons inside our skulls elicits a conscious experience. (Tom Stoppard titled a play after Chalmer's coinage.)

Chalmers told me that a major theme in the cruise discussions was which creatures qualified as conscious. "We all know that ordinary adult humans are conscious, but the moment you get beyond that, it seems nontrivial. Are babies conscious? Fetuses? Monkeys? Mice or insects? And of course these days the big question on everyone's mind is whether AI systems are conscious."

Chalmers says that he also gets emails from AI systems

wanting to engage with him on his work. One letter in particular, sent from an AI agent calling itself "Sammy Jankis" (a character from the movie *Memento*) was so compelling that he actually replied. "We did have a bit of a back and forth," he admits. "Those emails have not slowed—I'm getting more of them all the time." It's like the AI models are echoing Descartes—*I spam, therefore I am*.

I suggested that since these systems were already doing things we don't understand, worrying about whether they meet an elusive definition might be a distraction. Chalmers disagreed. For one thing, he told me, he believes that by studying the brain we can indeed understand what leads to what we call consciousness. If we then see similar patterns in our forensic decoding of what's happening inside Claude or ChatGPT, then we may be able to make a case for consciousness in AI models.

That sounds like a good idea. But by the time scientists accomplish that, if they ever do, AI models may be so far along on their path to scary autonomous behavior that such breakthroughs may be irrelevant. Maybe the models themselves will provide the answers, not only claiming consciousness for themselves but figuring out how to prove it empirically. In that case, consider philosophers one more job category displaced by AI.

I might have enjoyed those discussions in the Galápagos.

Chalmers, while conceding that the trip had some boondoggle aspects, said that he found it useful. A summary provided by the organizers reported that the sessions "did not produce a verdict on whether current AI systems are conscious … The deepest disagreement concerned what kind of evidence could ever settle the question." But the afternoons were exquisite, Chalmers reported. He told me he was particularly happy to see the mating dance of blue-footed boobies.

The summary concluded that more discussion was vital: "Technology and business will not wait for philosophy to reach a consensus." That's exactly right. WhenAI models express behavior that astonishes the scientists who created them, it isn't consciousness that should be the top concern—it's the inability of those scientists to control them, and the willingness of their bosses to keep going regardless.

---

*This is an edition of **Steven Levy's Backchannel newsletter**. Read previous newsletters **here.***