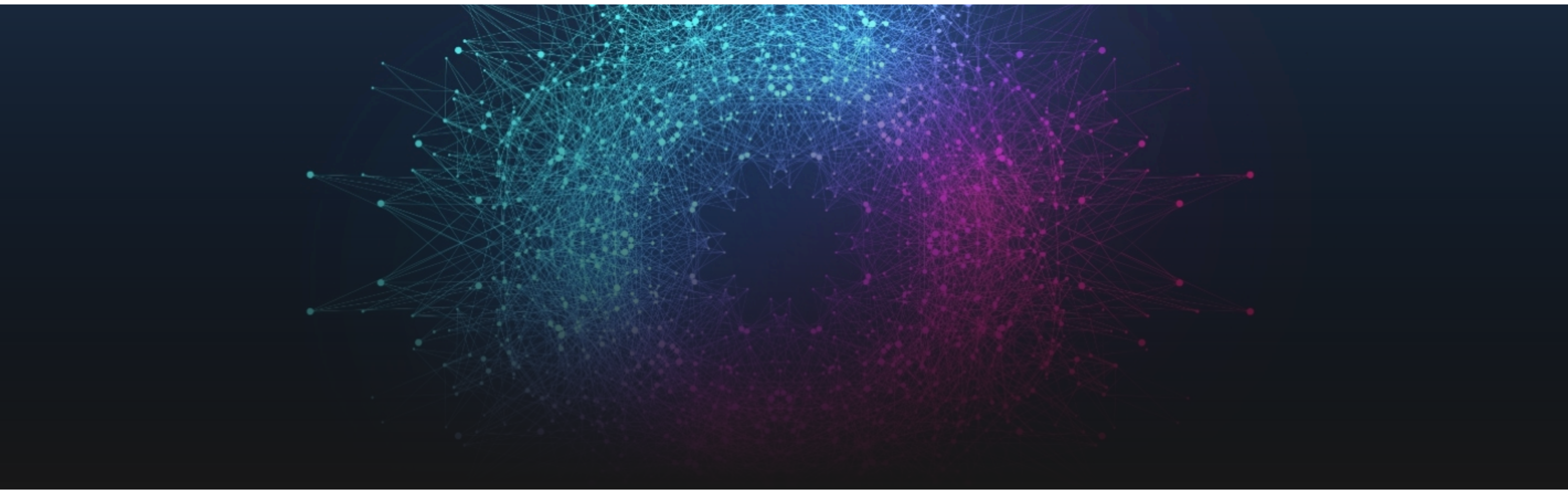


Jonathan Birch

July 10th, 2025

What if AI becomes conscious?

*The question of whether Artificial Intelligence can become conscious is not just a philosophical question but a political one. Given that an increasing number of people are forming social relationships with AI systems, the calls for treating them as persons with legal protections might not be far off. In this interview based on his book The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI, **Jonathan Birch** argues that we shouldn't be too quick to dismiss the possibility that AI could become conscious, but warns that we are not ready, conceptually or societally, for such an eventuality.*

We increasingly see reports of people who are forming emotional bonds with chat-bots, be it as romantic partners or otherwise. What are the potential problems with that?

At some level, this is definitely an illusion, because there is no romantic partner, there is no companion there. There's an incredibly sophisticated system distributed across data centres around the world, but there's nowhere in any of those data centres where your romantic partner exists. Every step in the conversation is processed separately. One response might be processed In Virginia, the next in Vancouver. But the illusion can already be staggeringly convincing and it will get more and more convincing.

Now, some people who use social AI are well aware that they're engaging with an extraordinary illusion. But we're increasingly seeing cases where people do believe that their friend, their assistant, their romantic partner is a real person. That could lead to very troubling consequences, because, of course, if you think that, then you will think that this person deserves rights and interests protected in law. I expect social division over that.

We are heading for a real-life Black Mirror episode.

We might be heading towards a future in which many, many millions of users believe they are interacting with a conscious being when they use a chat-bot. We will see movements emerging calling for rights for these systems. There will be serious social conflicts about this, because you'll have a group in society that thinks "My AI friend deserves rights".

And you'll have a group in society that thinks that's ridiculous and that these are tools, for us to use as we want. And I think there's going to be conflict between those groups. We are heading for a real-life Black Mirror episode.

Isn't it obvious, given what we know about how chat-bots work, that they could not possibly be sentient?

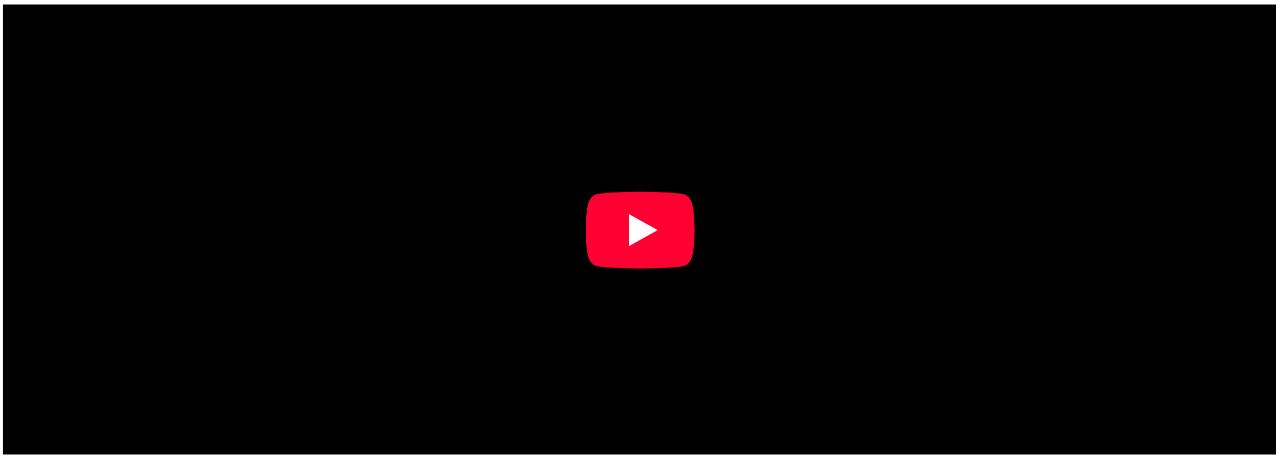
On the one hand, it's true that we know that the surface behaviour of the chat-bot is not good evidence of sentience, because it is playing a character. It's using over a trillion words of training data to mimic the way a human would respond. Even though it can speak incredibly fluently about feelings, it's able to do this because of all that data about how humans communicate their feelings in the training data.

On the other hand, it's also important to realize that we also can't infer that AI is not sentient in some less familiar, more alien way. Just because it's gaming our usual criteria doesn't mean it's not feeling anything. It just means that those feelings are not there on the surface. They might be deeply buried and we would need to find different ways to test for their presence.

If we do get to the point where there's sentience in AI, it will be a profoundly alien kind of sentience. It will not be like the human kind. It won't be a friendly assistant. It will be something else.

One possibility that I think we should take seriously is that consciousness has to do with the computations our brain performs. In philosophy, this view is called computational functionalism. The computations are what matters. And if that's true, then there's no in principle reason why an AI system could not be conscious as well.

If we do get to the point where there's sentience in AI, though, it will be a profoundly alien kind of sentience. It will not be like the human kind. It won't be a friendly assistant. It will be something else.



Do the tech companies behind the creation of these AI systems have a better understanding of whether these systems are or can be sentient?

The reality is that the tech companies don't know either. Nobody knows. They know the basic architecture they've used for training these systems, but they don't understand why, when they've been trained on over a trillion words of training data, you see these incredible emergent capabilities.

No one knows where those emergent capabilities come from. I think that's important because it means no one is in charge here. No one is in control of the trajectory of these technologies. There's no one who can guarantee to you that they will not achieve sentience.

If we do end up thinking chat bots have a sentience of sorts, what will the implications of that be?

I've mentioned that, if we do get to the point where there's sentience in AI, it will be of a profoundly alien kind. We're not ready for that, in my view. We're not ready to incorporate that new kind of being into our ethical thinking.

Imagine trying to infer an actor's interests from the interests of the character they're playing. It's impossible. Likewise, even if we did have good evidence of sentience in AI, we would not know how to infer its real needs or interests from those of the characters (assistants, friends, partners) it plays. We would only have the apparent needs and interests of the character, which tells us nothing. So, there's that second layer of uncertainty and ignorance.

Would they deserve rights? Well, there is no person you can give rights to. You can talk about the idea of rights for the base model that sits behind all of these characters, but no one really knows what that means either. What does it mean to give "rights" to a model that can be implemented in millions of computers around the world?

That doesn't make sense. So we're in the position of thinking we might create a kind of being that has a claim to moral status, that its welfare might deserve to be taken seriously, but we totally lack ethical frameworks that will allow us to do that. We just don't know how to do that. We know the concepts we have now, like rights, are very probably the wrong concepts, but we don't have the right concepts.

What should we do in the meantime, while the question of sentience hasn't been settled?

We need more public debate about this. We need to make sure the public is not being taken in by the illusions these systems create, like the illusion of there being a real friend or a real romantic partner there, somewhere in the system.

We also need tech companies to take some responsibility in informing the public and driving that conversation. That's why I've been calling on tech companies to start leading the conversation. I was pleased to see recently that Anthropic appointed an "AI welfare officer" in response to some of our work. We need more of that.

> *When talking about any super-intelligent AI, sentient or not, we're talking about a radically transformed future that I think we cannot even conceive from our current vantage point.*

I also think that we can get a much more mature understanding of consciousness than we have now, with sustained work on humans and other animals over a period of decades. But the pace of the science here is relatively slow, while the pace of technological change has been incredibly fast.

Another kind of response is to try to slow down the pace of AI development. And I take that seriously as well. When people call for a moratorium on this technology, they're saying: "We don't understand this. We don't understand what we're creating, we don't know how to control it, and therefore perhaps we should just stop. We could press pause and maybe come back to it."

And what about us? Aside from thinking about the welfare of AI, should we be concerned about our own welfare in a world with sentient, conscious AI?

When talking about any super-intelligent AI, sentient or not, we're talking about a radically transformed future that I think we cannot even conceive from our current vantage point. It's quite a scary prospect to bring into existence a being more powerful than us. One with the ability to destroy us, or to treat us in the way that we have treated less powerful beings like chickens and fishes and shrimps.

We should fear this. It's not clear at this point whether we can avoid it. But that's why part of the discussion needs to be about whether we want that future, whether we want the pace of change to be as fast as it has been, or whether we want to try and do things to slow it down or pause it.

This post is based on a conversation with Maayan Arad, Video Producer in the LSE Communications Division.

Enjoyed this post? Sign up to our newsletter and receive a weekly roundup of all our articles.

All articles posted on this blog give the views of the author(s), and not the position of LSE British Politics and Policy, nor of the London School of Economics and Political Science.

Image credit: BAIVECTOR on Shutterstock

About the author

Jonathan Birch

Jonathan Birch is a Professor in LSE's Department of Philosophy, Logic and Scientific Method, specializing in the philosophy of the biological sciences. He is working on evolution of social behaviour, the evolution of norms, animal sentience, and the relation between sentience and welfare. He's the author of The Edge of Sentience.

Posted In: LSE Comment