ARTICLE 10

The columnist
Chanda Prescod-Weinstein anticipates a mega telescope p18

Aperture
Pictures of snow and ice to send a shiver down your spine p20

Culture
The best non-fiction science titles for the year ahead p24

Culture columnist
Emily H. Wilson picks the sci-fi books to savour in 2025 p26

Letters
The secret way to cut calories when dining out p27

Comment

Not OK, computer?

Some prominent researchers argue that we should pay heed to the welfare of AIs. Are they right, wonders **Alex Wilkins**

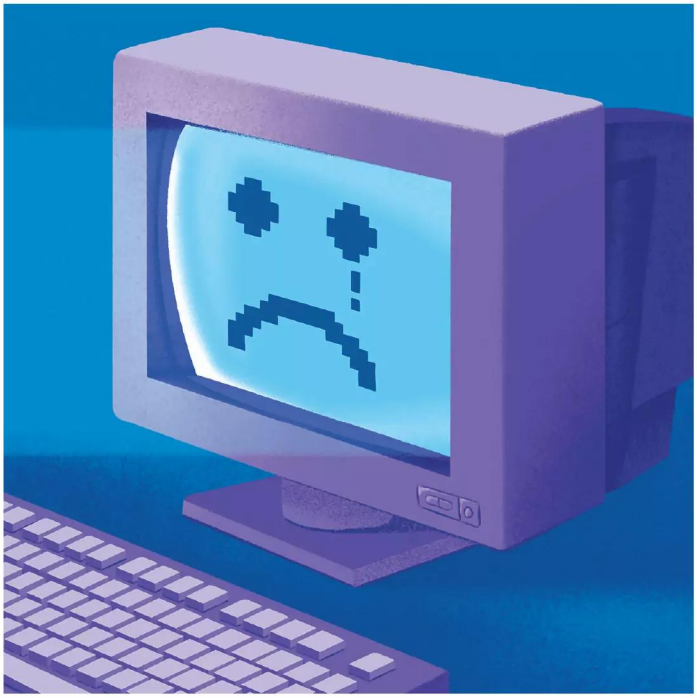
IS YOUR chatbot in distress? Many people, myself included, would scoff at this question. It is just computer code, optimised to predict the next word in a sequence. But some philosophers and psychologists say that we shouldn't be so quick to dismiss this question, perhaps even granting chatbots their own rights. They might have a point.

In a recent academic paper, "Taking AI Welfare Seriously", one group of researchers argue for a precautionary approach to how we treat AIs. They don't look to answer the question of whether an AI is conscious or not, but say we should start investigating the issue more thoughtfully. We probably won't find the answer tomorrow, or even within 10 years, but major societal consequences are dependent on it, they say.

"It's not that we can be confident that AI systems are conscious or agentic or otherwise morally significant," says Jeff Sebo at New York University, one of the authors of the paper. However, as "mistakes in either direction can be harmful", and there is likely to be uncertainty about these issues for some time, Sebo says we need to start developing "thoughtful" policies "sooner rather than later".

A dystopian future where we take advantage of sentient AIs would be horrific. But one where we ascribe consciousness to entities that aren't would also be worrying, as we might waste time and energy protecting AIs.

We should start planning for both scenarios, write Sebo and his team, by actively investigating AIs for signs of consciousness, and preparing "policies and procedures for treating AI systems with an appropriate level of moral concern". The "we" refers to AI companies, of which just a few have the means to push the frontier of development and investigate these issues.



ELAINE KNOX

According to the paper, one way that companies could do this is by appointing an AI welfare officer, a person who would be tasked with building "a structure for assessing AI systems for moral patienthood and making decisions about how to treat them". Kyle Fish, one of Sebo's coauthors, has been employed to do this by AI startup Anthropic, developer of the Claude chatbot.

It is worth stating that Sebo's thinking here is closely related to his work on animal consciousness. He was one of 40 scientists who initially signed The New York Declaration on Animal Consciousness in April, which cites "strong scientific support for attributions of conscious experience" in mammals and birds. Sebo foresees AI and animals forming a greater group with non-human consciousness. But while the philosophical parallels for both are clear, the evidence base for each seems, to me, quite different. Animal consciousness has been investigated for decades, while competent AI systems have only just been invented.

In any event, assertions of AI consciousness are being made. In 2022, an ex-Google engineer made headlines by saying that the company's earlier chatbot, LaMDA, showed signs of sentience, a claim which Google and the wider AI community pushed back against.

With millions of us interacting with AI, pronouncements like this will only become more frequent. Whether a claim about AI sentience from an AI company will sway public opinion is unclear. But the issue seems worth devoting some effort to, so when lawmakers inevitably have to rule on the interactions we have with AIs, they have proper evidence about what is going on in these machines.

So I will hold my scoffing for now, if only on the off chance that it helps avoid the remote scenario where a bot becomes aware of its fate, like Marvin the depressed android in *The Hitchhiker's Guide to the Galaxy*, who laments: "Here I am, brain the size of a planet and they ask me to take you down to the bridge. Call that job satisfaction? 'Cos I don't."

Alex Wilkins is a reporter at *New Scientist*