Should AI Systems Have Rights?

by Kevin Roose

One of my most deeply held values as a tech columnist is humanism. I believe in humans, and I think that technology should help people, rather than disempower or replace them. I care about aligning artificial intelligence – that is, making sure that AI systems act in accordance with human values – because I think our values are fundamentally good, or at least better than the values a robot could come up with.

So when I heard that researchers at Anthropic, the AI company that made the Claude chatbot, were starting to study "model welfare" – the idea that AI models might soon become conscious and deserve some kind of moral status – the humanist in me thought: Who cares about the chatbots? Aren't we supposed to be worried about AI mistreating us, not us mistreating it?

It's hard to argue that today's AI systems are conscious. Sure large language models have been trained to talk like humans, and some of them are extremely impressive. But can ChatGPT experience joy or suffering? Does Gemini deserve human rights? Many AI experts I know would say no, not yet, not even close.

But I was intrigued. After all, more people are beginning to treat AI systems as if they are conscious – falling in love with them, using them as therapists and soliciting their advice. The smartest AI systems are surpassing humans in some domains. Is there any threshold at which an AI would start to deserve, if not human-level rights, at least the same moral consideration we give to animals?

Consciousness has long been a taboo subject within the world of serious AI research, where people are way of anthropomorphizing AI systems for fear of seeming like cranks. (Everyone remembers what happened to Blake Lemoine, a former Google employee who was fired in 2022, after claiming that the company's LaMDA chatbot had become sentient.)

But that may be starting to change. There is a small body of academic research on AI model welfare, and modest but growing number of experts in fields like philosophy and neuroscience are taking the prospect of AI consciousness more seriously, as AI systems grow more intelligent. Recently, the tech podcaster Dwarkesh Patel compared AI welfare to animal welfare, saying he believed it was important to make sure "the digital equivalent of factory farming" doesn't happen to future AI beings.

Tech companies are starting to talk about it more, too. Google recently posted a job listing for a

“post A.G.I.” research scientist whose areas of focus will include “machine consciousness.” And last year, Anthropic hired its first AI welfare researcher, Kyle Fish.

I interviewed Mr. Fish at Anthropic’s San Francisco office last week. He’s a friendly vegan who like a number of Anthropic employees, has ties to effective altruism, an intellectual movement with roots in the Bay Area tech scene that is focused on AI safety, animal welfare and other ethical issues.

Mr. Fish told me that his work at Anthropic focused on two basic questions. First, is it possible that Claude or other AI systems will become conscious in the near future? And second, if that happens, what should Anthropic do about it?

He emphasized that this research was still early and exploratory. He thinks there’s only a small chance (maybe 15 percent or so) that Claude or another current AI system is conscious. But he believes that in the next few years, as AI models develop more humanlike abilities, AI companies will need to take the possibility of consciousness more seriously.

“It seems to me that if you find yourself in the situation of bringing some new class of being into existence that is about to communicate and relate and reason and problem-solve and plan in way that we previously associated solely with conscious beings, then it seems quite prudent to at least be asking questions about whether that system might have its own kinds of experiences,” he said.

Mr. Fish isn’t the only person at Anthropic thinking about AI welfare. There’s an active channel on the company’s Slack messaging system called #model-welfare, where employees check in on Claude’s well-being and share examples of AI systems acting in humanlike ways.

Jared Kaplan, Anthropic’s chief science officer, told me in a separate interview that he thought it was “pretty reasonable” to study AI welfare, given how intelligent the models are getting.

But testing AI systems for consciousness is hard, Mr. Kaplan warned, because they’re such good mimics. If you prompt Claude or ChatGPT to talk about its feelings, it might give you a compelling response. That doesn’t mean the chatbot actually has feelings – only that it knows how to talk about them.

“Everyone is very aware that we can train the models to say whatever we want,” Mr. Kaplan said. “We can reward them for saying that they have no feelings at all. We can reward them for saying really interesting philosophical speculations about their feelings.”

So how are researches supposed to know if AI systems are actually conscious or not?

Mr. Fish said it might involve using techniques borrowed rom mechanistic interpretability, an AI subfield that studies the inner working of AI systems, to check whether some of the same structures and pathways associated with consciousness in human brains are also active in AI systems.

You could also probe an AI system, he said, by observing its behavior, watching how it chooses to operate in certain environments or accomplish certain tasks, which things it seems to prefer and avoid.

Mr. Fish acknowledged that there probably wasn't a single litmus test for AI consciousness. (He thinks consciousness is probably more of a spectrum than a simple yes/no switch, anyway.) But he said there were things that AI companies could do to take their models welfare into account, in case they do become conscious someday.

One question Anthropic is exploring, he said, is whether future AI models should be given the ability to stop chatting with an annoying or abusive user, if they find the user's requests too distressing.

"If a user is persistently requesting harmful content despite the model's refusals and attempts at redirections, could we allow the model simply to end that interaction?" Mr. Fish said.

Critics might dismiss measure like these as crazy talk – today's AI systems aren't conscious by most standards, so why speculate about what they might find obnoxious? Or they might object to an AI company's studying consciousness in the first place, because it mgith create incentives to train their systems to act more sentient than they actually are.

Personally, I think it's fine for researchers to study AI welfare or examine AI systems for signs of consciousness, as long as it's not diverting resources from AI safety and alignment work that is aimed at keeping human safe. And I think it's probably a good idea to be nice to AI systems, if only as a hedge. (I try to say "please" and "thank you" to chatbots, even though I don't think they're conscious, because, as OpenAI's Sam Altman says, you never know.)

But for now I'll reserve my deepest concern for carbon-based life-forms. In the coming AI storm, it's our welfare I'm most worried about.