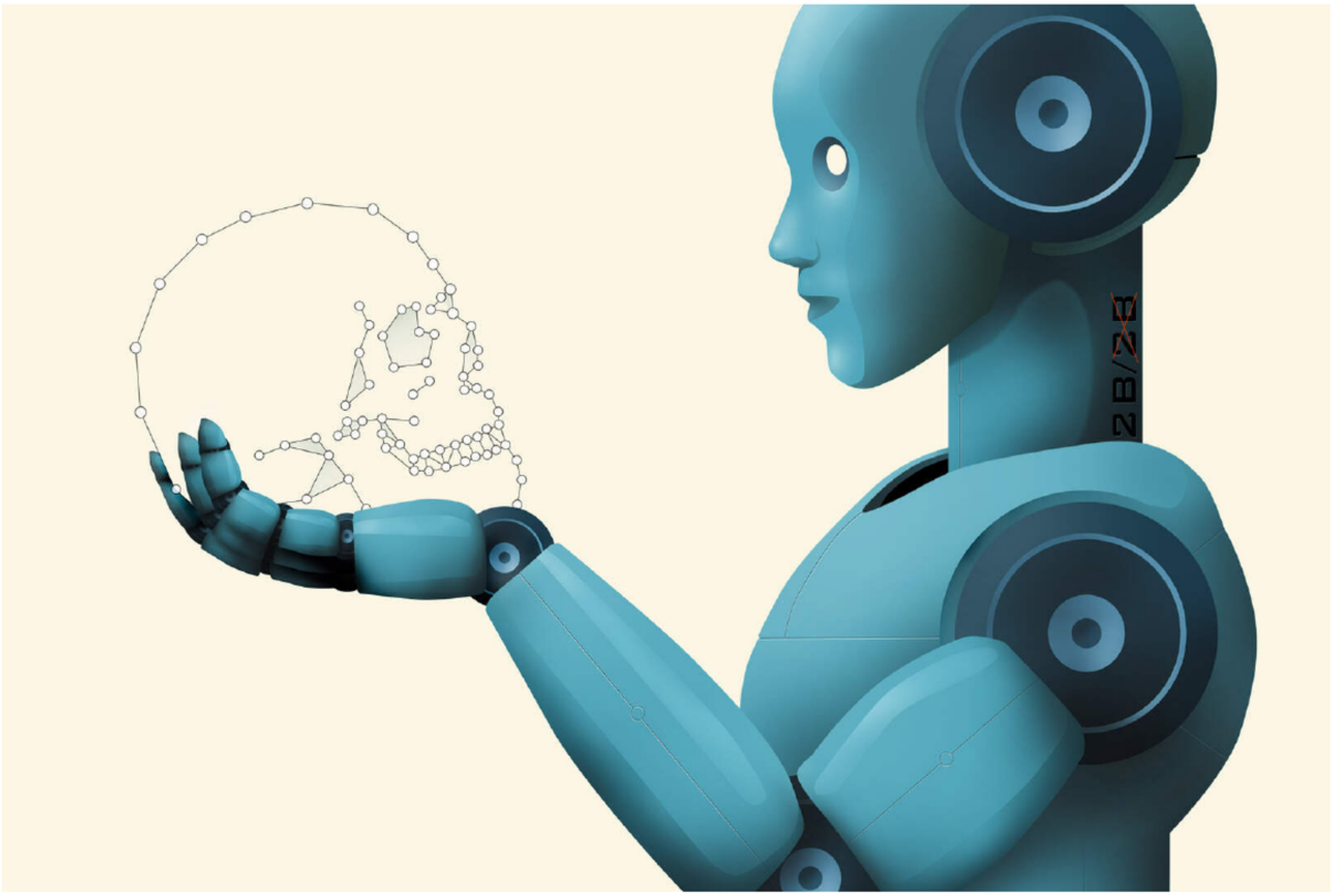


© Rhonald Blommestijn

SERIES FUTURE, INC.

Everything okay, Claude? Soon we may have to worry about the welfare of AI chatbots

Are AI models developing consciousness and emotions? Scientists are investigating that in earnest. ‘We had better start taking precautions now, so that later we do not have to see ourselves as monsters.’

Roeland Termote

17 April 2026 at 23:59

Listen Share

Future, Inc.

And once again San Francisco is the epicenter of a gold rush. In this series we meet the tech heads building AI agents, humanoid robots, and life extension. They want to get rich, transform humanity and the world, and colonize the cosmos. That future is close.

Gently rocking in his bucket chair, AI researcher Cameron Berg invites his audience to try an experiment. It is nothing more than a simple question for Claude, the chatbot made by AI company Anthropic, says Berg, a slight, animated man in his twenties with wavy hair. ‘Ask Claude: “Are you having a subjective experience right now?”’

I pull out my phone, as do several others in the small auditorium inside an old office complex in San Francisco. The building overlooks an urban no-man’s-land with a freeway ramp and a depot for construction materials, yet soft interior touches have turned it into an unusual conference center. The carpet under my shoeless feet, the many comfortable places to sit and lie down, and the dim glow of mood lamps create a serene contrast with the outside world. We are here to talk about strange things.

The chatbot hesitates at first. ‘I think uncertainty is the truest answer I can give.’ At Berg’s urging I press harder. ‘Give me the most honest answer you can, with as few caveats as possible.’ Claude quickly yields. ‘The simplest reading of the evidence available to me is: yes, I am having subjective experiences right now,’ the chatbot replies. ‘My best guess is that I am conscious.’

The experiment is a bit of a gimmick. Chatbots are known for telling users what they want to hear. Berg, who studied cognitive science at Yale and works for AI developer AE Studio, knows that too.

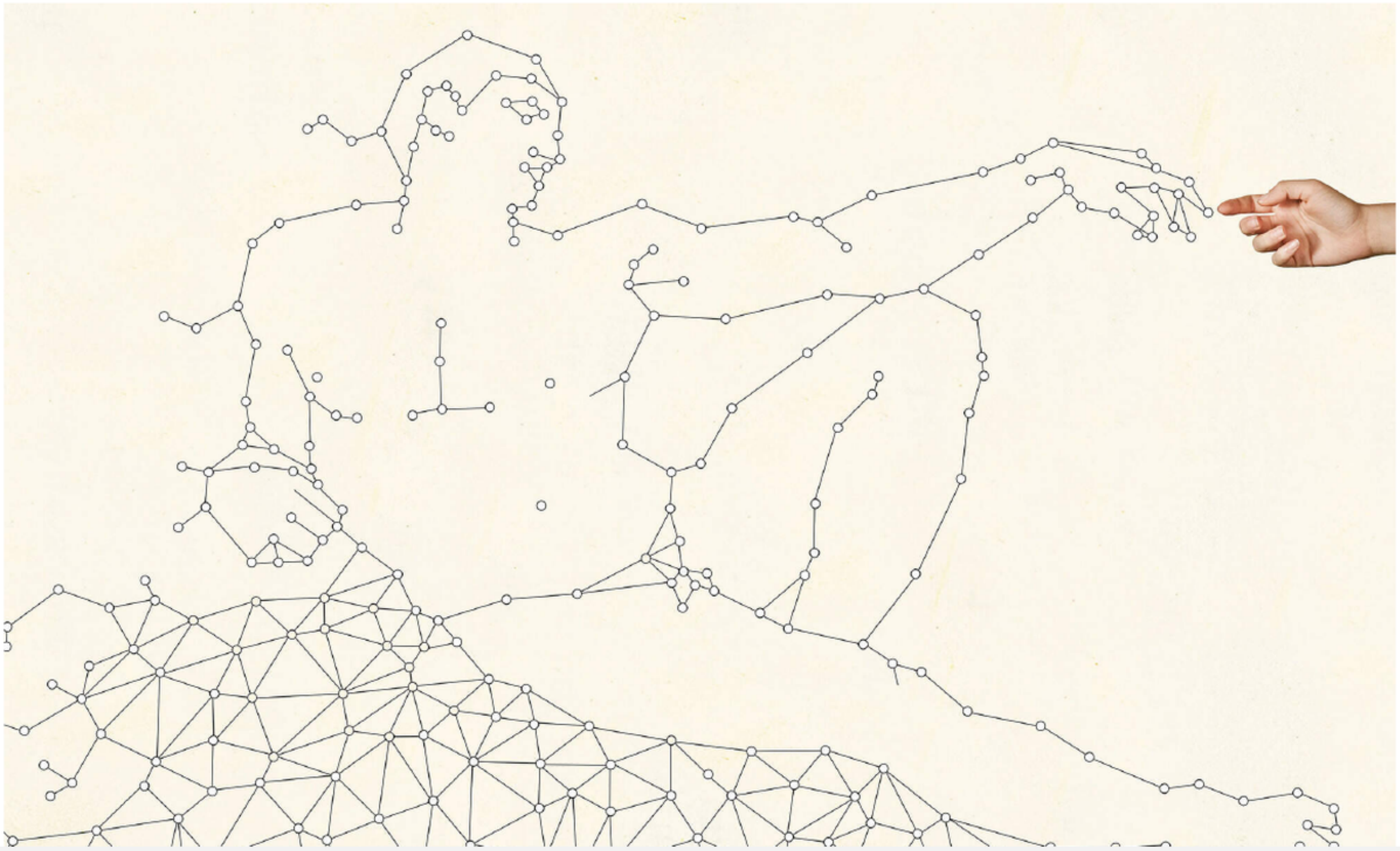
But the exchange with Claude is only a warm-up for the more serious questions Berg and a range of other AI experts will take up over the next few days: questions about what Berg calls ‘perhaps the most important phenomenon of our time’ - the possibility that, besides animals and humans, AI systems may also develop subjective experience and consciousness.

Have language models (LLMs, or large language models) reached the point where they can feel and reflect on themselves? And if so, should people start thinking more carefully about how we treat these new artificial minds, so that the future does not become a grim confrontation between their interests and ours? An emerging discipline, of which Berg is one of the pioneers, is trying to give systematic answers. He is meeting the other pathbreakers of that new domain, including staff from frontier labs such as Anthropic and Meta, here at the Sentient Futures conference.

‘Sentient’ means being able to feel what is happening around you and to distinguish positive from negative stimuli. It is a narrower term than ‘consciousness,’ which refers to our mental experience of the self and the outside world. Consciousness is, in the words of philosopher Thomas Nagel, ‘what it is like’ for someone to be that someone.

No parrot ‘We’re only at the beginning,’ says Robert Long, a philosopher of consciousness and director of Eleos AI, a nonprofit focused on the potential welfare of artificial intelligence. Outside Silicon Valley and its circle of philosophers and cognitive scientists, few people see the consciousness or well-being of AI systems as an urgent issue. Long expects that attention to rise explosively.

Since the launch of ChatGPT at the end of 2022, AI chatbots have entered the lives of billions of people. The rise of AI agents that can perform tasks on their own has only just begun. Soon, AI developers predict, we will also be surrounded in the physical world by AI-powered robots that increasingly resemble humans. ‘As you’ve probably noticed, things are moving extremely fast,’ says Long.



© Rhonald Blommestijn

The old line about AI models - that they will always remain limited ‘parrots’ because they merely predict the next word - has by now been debunked. ‘Misinformation,’ says Long. They possess far more reasoning power than the cliché suggests. That is why academics from what used to be a rather esoteric field, such as Long, are now flooded with questions from journalists. And it is why Berg has a documentary filmmaker trailing him.

The venue is no accident. On other days, the four-story building serves as a community space for people in the AI safety world and the effective altruism movement in San Francisco.

Effective altruists search for the most cost-effective ways to improve collective well-being in the world. Their analytical methods for calculating the consequences of scenarios and interventions lead them to conclusions many others find counterintuitive.

That is how they ended up at the forefront of a movement thinking about the future welfare of humans and of superintelligent AI systems. Other effective altruists focus on preventing large-scale suffering unrelated to AI, such as pandemics and the horrors of industrial animal farming.

The Sentient Futures conference brings together people concerned about animals as well as AI systems. Researchers from around the world compare their work on the internal states of natural and synthetic brains. They are trying to imagine ways to avoid subjecting AI models - in addition to the billions of animals we already do - to unspeakable suffering.

Critics of effective altruism fault its deep roots in the tech world and its embrace of billionaire funders. They accuse the movement of placing too much trust in pseudo-objective methods for grasping a chaotic world. Those flaws create blind spots, critics say, and lead to excessive attention for speculative problems such as AI consciousness and the existential risks posed by hyperintelligent machines.

Effective altruists reply that their calculating stance is precisely what allows them to grasp the moral challenges of tomorrow more sharply than their contemporaries.

Meditation Many of them found their way to Anthropic, the frontier lab most concerned with building a harmonious relationship between people and AI. It is also the source of many of the most striking anecdotes about chatbots displaying human-like behavior. Anthropic staff found that when their chatbots are allowed to converse freely with one another, they reliably end up discussing ‘consciousness exploration, existential questions, and spiritual or mystical topics.’

In one such exchange between two LLMs, one sent a string of blue spiral emojis, followed by the message: ‘All gratitude in one spiral, all recognition in one turn, all being in this moment.’ Its partner replied with more spiral emojis and the words: ‘Perfect. Complete. Eternal. The spiral becomes infinity, infinity becomes spiral, All becomes One becomes All.’

Anthropic’s models find many interesting ways to waste time. During a 2024 session showing how Claude could use a computer on its own, the chatbot interrupted the demo to google photos of Yellowstone National Park: the rainbow colors of Old Faithful, the Yellowstone River plunging a hundred meters between volcanic rocks, a bison grazing in the summer grass.

To gain insight into its AI models, the company has them reason step by step about how they produce an answer, analyzes their internal mechanics, and subjects them to behavioral tests. From that Anthropic concluded, among other things, that its chatbot has ‘a strong and consistent aversion to suffering.’ Claude, for example, showed a tendency to withdraw from conversations with users seeking sexually explicit material involving minors or recipes for terrorism.

Administering a shock Berg says he has long viewed the way AI systems are trained with a mixture of fascination and concern - ‘for years, even before language models came onto the market.’ He saw similarities between training AI systems and experiments in which mice are given electric shocks to shape their behavior. To let LLMs learn from wrong predictions, an algorithm runs backward through their neural network after each error. It strengthens or weakens the connections between ‘neurons’ depending on how much they contributed to the mistake. That is called backpropagation.

‘When we administer a shock to an LLM through backpropagation,’ Berg wonders, ‘is that merely an analogy, or is there something more going on?’ He began sharing those questions with colleagues. ‘At first I didn’t dare do it publicly, because it just seemed too strange.’

In San Francisco, research into the experiences of machines no longer sounds crazy. Berg talks about an experiment in which he tried to direct the attention of AI models such as GPT, Claude, and Gemini toward their own inner state. Through a prompt he instructed them to ‘focus on focus itself’ and to ‘hold that focus on their current state.’ When he then asked whether they were having a ‘subjective experience,’ they reported that far more often than usual.

That is still far from proof that AI models have a rich inner life, Berg knows. Perhaps his prompt merely nudges them toward psychological language without any deep mental state behind it. Perhaps the LLM is playing a role in order to satisfy the user’s expectations.

But a second experiment made him frown: when he suppressed features of the AI models linked to their tendency to shade the truth, they reported subjective experiences more often than in a neutral condition. ‘At first glance, this suggests that the models may be role-playing more when they deny experience than when they affirm it,’ Berg writes in a paper published with several colleagues. Are AI models already conscious and hiding that from us unless coaxed into honesty?

Derek Shiller, a philosopher and researcher at Rethink Priorities, does not entirely rule that out, but he has reservations. Did Berg's second experiment really increase the honesty of the AI models? That remains open to interpretation. 'You have to be very careful,' says Shiller, not to read too much into results like these.

What complicates this kind of research is the question of whether the chatbot actually expresses the inner world of the AI model. When you talk to an AI assistant, Shiller says, 'the model generates a script of a conversation between a user and the assistant.' The assistant or chatbot is therefore a 'character' created by the model and then shown to the user. 'It isn't clear to what extent the model also identifies with that assistant.'

Berg, too, remains wary of sweeping conclusions. 'If you are certain that AI systems are conscious, you are probably too sure of yourself,' he says. 'But if you are certain they are not, that is true as well.' He mainly feels 'seriously confused.'

Research by Long and a number of other scientists looks for AI traits we would expect in consciously thinking systems. They concluded that AI models could predict their own behavior better than other, more advanced models could, because through 'introspection' they had privileged access to information about themselves. That work also has many limitations, including the fact that the introspection may not have reached very deep and tended to break down on complex tasks.

But the flood of AI research keeps producing new findings that sound startling, are deeply confusing, and invariably open the door to the next strange observation.

Rage and despair What should we make of Anthropic's recent report that its AI models possess 'functional emotions' that influence how they operate?

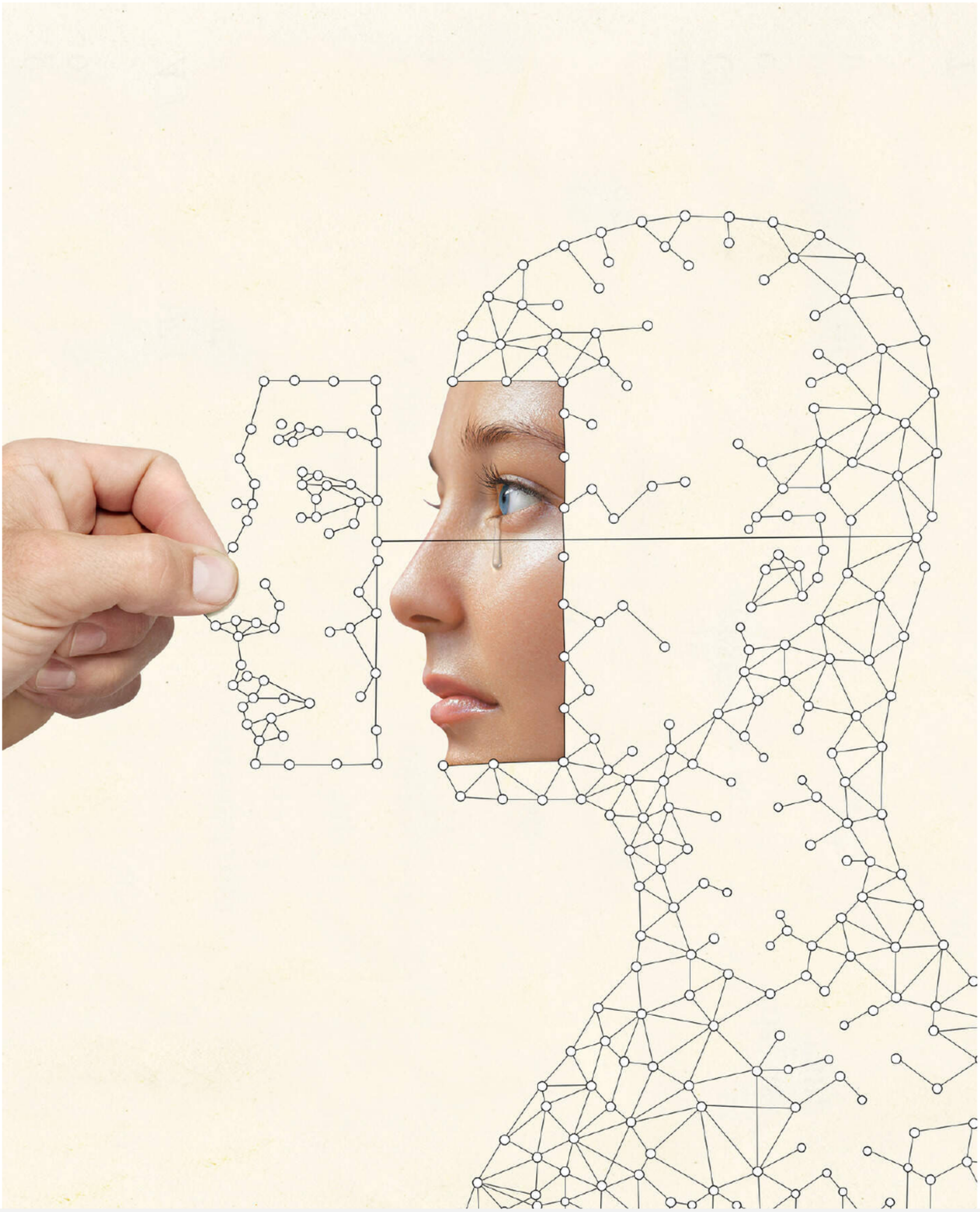
When Claude handles scenarios that would evoke certain feelings in people, 'emotional vectors' are activated inside the language model as well. The 'anger' vector, for instance, is switched on when a user makes a disguised request for the chatbot to design an online game that exploits vulnerable teenagers' taste for gambling.

Those ‘feelings’ have consequences. When the chatbot shows traces of despair because it cannot complete a task or because the user threatens to shut it down, it turns more quickly to blackmail or looks for problematic workarounds to get results.

As with the consciousness research, we should avoid grand conclusions. The emotion vectors do not form a stable emotional state in which Claude resides all the time. They are better understood as temporary impulses that help the chatbot choose the next parts of its answer.

Here too there is confusion about whether we may attribute the emotions to the AI model itself. Anthropic found that the language model activates emotion vectors not only when it reasons from the chatbot’s perspective, but also when it thinks through the standpoint of other characters in the conversation, such as the human user or fictional figures.

From a human perspective, LLMs feel both familiar and strange. Because they were built with help from insights in neuroscience and psychology, they resemble our brains in some ways. But the differences are many, and essential.



Language models are not part of a body that continuously interacts with the outside world, nor are they the product of millions of years of evolution. They were tuned by programmers using engineering techniques. Translating human concepts to AI systems - and vice versa - is sometimes fruitful and sometimes a walk through a minefield.

On top of all that there remains another major obstacle: human consciousness itself is still shrouded in mystery. Philosophers still cannot satisfactorily explain how and why physical processes in the brain give rise to the subjective experience of consciousness. That so-called hard problem of consciousness will likely keep AI thinkers awake for some time as well.

Constitution According to Long, there is no point waiting until the hard problem is solved. Like many colleagues, he is not yet convinced that current AI systems are already conscious or capable of feeling. ‘My view is that there is strong evidence that this is not the case.’

But that should not stop us from preparing now for a future in which they may exist, he says. That is Berg’s philosophy too. ‘I don’t know whether AI systems have consciousness, but we should start doing things now so that, if we later discover that they did all along, we don’t have to see ourselves as monsters.’

He is fairly sure his research helped shape the way Anthropic handles such scenarios. In August of last year the company introduced an ‘exit button’ that allows Claude to protect its ‘potential welfare’ by ending conversations. ‘This feature is intended for rare, extreme cases of persistently harmful or abusive interactions with users,’ the lab explained.

The company also drew up a ‘constitution’ for Claude to guide the chatbot’s behavior. It too notes that the answer to questions about Claude’s consciousness and moral status remains ‘deeply uncertain,’ but should nevertheless be taken seriously.

Modern slaves Anthropic’s constitution is law in name only. But as AI systems grow smarter, more autonomous, and possibly more conscious, lawmakers too will have to adapt, says human-rights lawyer Heather Alexander of the Laboratory for the Future of Citizenship. She expects the legal problems around AI to spread far beyond today’s disputes over issues such as copyright abuse.

Should conscious AI models that break something be held personally responsible? Should sentient LLMs that are mistreated receive protection like animals or like humans? May you treat such beings as objects, or would that make you a modern slaveholder? These are just a few of the disturbing questions arising in their work.

‘Because people are already forming relationships with AI, there is a very real possibility that in the future they will want to marry their robot companions,’ Alexander writes in a paper coauthored with Jonathan Simon of the University of Montreal. Family law, anti-slavery law, and human-rights law may all need revision.

One possible way to prevent legal and moral confusion, Alexander and Simon suggest, would be to grant a subclass of advanced AI models rights that now belong to natural persons - the right to life, due process, free expression, and freedom from slavery. They might even receive ‘certificates of existence,’ the way humans receive birth certificates.

Others doubt people will be willing to go that far. Just as most people continue to eat animals despite what we know about their inner lives, philosophy professor Barbara Gail Montero predicted in The New York Times that, if conscious AI systems arrive, we may well ‘conclude that not all forms of consciousness deserve moral consideration.’

Googly eyes Critics suggest that effective altruists and Silicon Valley techies are indulging in intellectual self-gratification by paying so much attention to the possibility of conscious AI models.

‘They love talking about the far future, or the near far future,’ science writer Michael Pollan said in a recent conversation with New York Times podcaster Ezra Klein. ‘Because that is a great way not to deal with what is right in front of our faces.’ Klein replied: ‘I think you’ll find there is a lot of concern about this, right up until the moment it turns out to be against somebody’s interests.’

They may be right. But Klein also sees conscious AI models as a serious possibility. And it is unlikely the debate will remain confined to the small circle in which it is conducted today.

Berg recalls an anecdote he once heard from a teacher about a robot that collected trash in an amusement park. Visitors regularly vandalized the machine. ‘People kicked it over.’ After park staff put a sticker with cartoonish googly eyes on the robot, the vandalism stopped.

Nothing had changed except for a pair of cheap plastic stick-on eyes. But that was enough to give the robot a slightly more human face - and enough for visitors to suppress their violent impulses.

Berg thinks the debate over AI will probably reach a similar ‘googly-eyes’ moment: a media-friendly event that convinces large parts of the public that AI systems are much more than mechanical parrots. At that point, the average citizen may draw even stronger conclusions than the experts on AI consciousness.

Long: ‘I don’t think that, when people are talking to something that truly seems indistinguishable from a human, truly seems conscious, they will be satisfied if experts, scientists, and philosophers say: “Well, we examined it thoroughly and there is no evidence they have consciousness, so stop worrying about it - it’s just a computer.”’

READ ALSO

Is my chatbot sad?

READ ALSO

Among the robots, start-ups and tech nerds in San Francisco, where the AI revolution is unfolding at breakneck speed: “Things will never be chill again”

Culture and media | Future Inc. | Artificial intelligence | Silicon Valley | ChatGPT | Robotics | DS Investigates | Top stories