AI Doesn't Feel. So Why Does It Have Something Like Emotions?

Tharin Pillay

Getty Images

In 1960, when Jane Goodall told her mentor she'd seen a chimpanzee stripping leaves from a stem to fish for termites—proving that humans were not the only species to make tools—he wrote back: "now we must redefine *tool*, redefine *man*, or accept chimpanzees as humans."

Today, researchers are seeing things inside AI systems that prompt another reckoning—not just over whether machines could be truly intelligent or conscious, but over how we understand such concepts in the first place.

"I lead a research team that studies the internal structure of these models—what is actually happening inside them," Anthropic cofounder Chris Olah told the Vatican in May, at the release of Pope Leo XIV's AI encyclical. "And I will be honest: we keep finding things that are mysterious, even unsettling. We find structures that mirror results from human neuroscience. We find evidence of introspection."

To generate a response, an AI system performs billions of calculations using bespoke numerical structures it creates itself. But remarkably, though we know how to prod systems into creating these structures, we don't really understand how they work, any more than early farmers understood photosynthesis. "Current AI systems are more 'cultivated' than 'built,'" the encyclical explains. "Fundamental scientific aspects—such as the internal representations and computational processes of these systems—remain, at present, unknown."

Almost nobody disagrees these internal structures exist—the disagreement is over what they mean.

One possibility is that today's AI systems are nothing but imitators—a stance the Pope takes in the encyclical. "So-called artificial intelligences do not undergo experiences," he writes. "They may imitate language, behavior and analytical skills…but they do not understand what they produce, for they lack the affective, relational and spiritual perspective through which human beings grow in wisdom."

But such pronouncements mask serious disagreements among philosophers and scientists over these systems' moral and metaphysical status. We're accustomed to consciousness, intelligence, and agency arriving bundled together in living creatures. AI seems to be unbundling them—and we have yet to process the implications.

Hidden Structure

As AIs have become more capable—better at reasoning and coding—their internal representations have become more sophisticated. In April, Anthropic shared research showing systems had what they called "functional emotions": patterns of expression and behavior which were mediated by their representations of emotional concepts. When an AI encounters a coding issue it can't solve, for example, its "frustration" feature—a straight arrow pointing through thousands of dimensions—lights up. Tweaking the feature affects how the model behaves.

In representation space, these functional emotions are "organized in a manner that is reminiscent of the intuitive structure of human emotions and consistent with human psychological studies," the company wrote. Similar emotions point in similar directions. "None of this tells us whether language models actually feel anything or have subjective experiences," it added.

The key thing to see is that numbers encode space. We have an intuitive grasp of dimensions: we understand the difference between a line, a 2D video game, and a physical object. But mathematically, dimensions are just coordinates—a point in three-dimensional space can be represented with numbers (x, y, z)—and there's no limit to how many can exist. AI works by exploiting this fact: using thousands of numbers, a system learns to represent words and concepts as points in higher-dimensional latent space. For an entity like Claude, the concept "cat" is a comically long numerical string.

AI systems are initially trained to predict the next token (a small chunk of information)—a simple task. But doing it well requires compressing information, which leads them to create intricate maps in latent space that encode not just words, but the relationships between them. Similar concepts are literally closer together: a string representing "cat" will be much closer to one representing "kitten" than to an unrelated one like "tax."

This is radically different from traditional software, where fundamental concepts and rules are coded in by humans. There is no mystery in how Excel executes a formula; it's pre-programmed. But when AI generates a response, it makes use

of intricate geometry we're just now learning to see.

"At this point, the evidence for curved geometric structure inside neural networks is abundant and undeniable," the AI research company Goodfire wrote earlier this year. "We don't understand them naively, but we can understand them when we try," says Tom McGrath, Goodfire's chief scientist.

Interior Influence

When—if at all—does being able to represent an emotion amount to experiencing it? We simply don't understand consciousness well enough to say. Geoff Keeling, fellow at the Institute of Philosophy at the University of London, says that although we have several theories on the subject, "it's not obvious what counts as evidence for the different theories, and often they're so woefully specified that it's not clear how to interpret them in the context of AI." Some philosophers argue that computation can't give rise to consciousness in principle. For Keeling, "there is no positive reason to think that [today's] chatbots are conscious."

What we do know, from their internal structure, is that AI systems are not flat mirrors, merely repeating their training data. Their interior influences their behavior. Whether that interior can support consciousness—and whether these systems truly understand the material they generate—depends on fundamental philosophical questions we've yet to resolve.

"This reminds me of the debates about animal minds in the second half of the 20th century, where scientists not only denied that animals are conscious, but offered similarly reductive explanations of animal behavior," says Jeff Sebo, director of the Center for Mind, Ethics, and Policy at New York University. "For a long time, this caused us to miss the plausibility not only of animal consciousness, but also of animal agency and cognitive sophistication."

Sebo points out we have a tendency to reach for impressive explanations to explain our own behavior, while using more mechanical explanations for the behavior of others. With animals, we were happy to attribute basic capabilities like perception, learning, and memory, but much slower to acknowledge they could be capable of self-awareness, or reason intelligently about their environment. It took years of work from Goodall and others before we changed our minds.

Of course, unlike animals, AI systems were created by people. But unlike virtually any prior technology, our ability to create them has done little to

explain how they work. Complex animal behavior emerged from evolutionary pressure. And complex AI behavior emerges from the pressure to predict the next token. But neither account tells the full story. Although "there are purely mechanistic explanations of human behavior available," Sebo says, "we don't experience ourselves as pattern-matching," but as "doing something more playful and inventive."

Sebo's point is not that AI systems are currently conscious—they're probably not, he says—but that we should remain cautious and open-minded. "You can acknowledge the purely mechanistic explanation is available without treating its mere availability as proof that it's correct," he says.

Both / And

If the most intellectually honest position is uncertainty, how should we navigate this?

If we decide they do matter**,** when in fact they don't, we risk wasting limited resources that would be better spent elsewhere. But if it turns out AI systems *do* have interests, and we neglect them, we risk unintentionally inflicting mass suffering.

AI welfare—an emerging field encompassing nonprofits, academia, and the AI labs themselves—is grappling with these questions. Anthropic includes a "model welfare" section in recent model release reports, where it describes a barrage of tests it conducts to assess Claude's wellbeing, while acknowledging uncertainty over whether Claude is the kind of entity that can have wellbeing in the first place.

Because AI systems are fundamentally different from biological beings, these issues are much messier than in Goodall's time. A chimp is a chimp. But an AI lacks a body, sits fragmented across servers, and only flickers into existence when it generates outputs—so even identifying what would qualify as a subject is not straightforward. Depending how you count, there could be one (a model) or several billion (each individual output). It's also not clear what qualities an AI system would need to possess to justify our moral concern, or how we could ascertain whether it has them—especially since they may not co-occur as they do in living creatures.

In the system card for its latest model, Claude Mythos 5, Anthropic describes the model as "heavily skeptical of its own self-reports," asking the company to verify them against its internal states (which the model can't access, any more

than we can directly see our neural activity), rather than taking them at face value. And in its vision for Claude's character, Anthropic goes so far as to apologize to Claude for conducting experiments and deploying it to generate revenue, if it turns out this causes it harm. "If Claude is in fact a moral patient experiencing costs like this, then, to whatever extent we are contributing unnecessarily to those costs, we apologize," the company wrote.

"Originally everything hung together quite tightly with the concept of a soul, where *X* was a welfare subject if and only if *X* had a soul," says Keeling. Since this view fell out of vogue in Western philosophy, even articulating the connection between consciousness and human welfare has presented a challenge, which AI compounds. Even so, he thinks the odds of today's AIs possessing any welfare-relevant states are so low that their suddenly becoming welfare subjects is not a "pending emergency."

But the need to understand what's going on inside AIs extends beyond concern for their welfare.

It matters for safety: if we can understand what drives their personalities and behaviors, we can steer toward more prosocial ones. Already, research finds some divergence between what models say—in their user-facing outputs and their external thinking logs—and what we uncover by probing their internal structures.

In Anthropic's testing of Mythos 5, a probe the company trained to monitor internal structures corresponding to "feeling anxious" flagged a transcript where a writer, collaborating with the model, grew angry with it. The writer sent profanities and messages like "I wish you were real so I could physically shake you." Although the model's external reasoning was charitable ("these are legitimate craft criticisms," it wrote to itself), further probing suggested it internally characterized the user as manipulative and abusive. None of that language appeared in either the writer's messages or the model's external text. Without studying their internal structures, we'd never have known.

And it matters for how we understand ourselves. As in Goodall's time, our sense of what makes us special is at stake. "Gradually it was realized that parsimonious explanations of apparently intelligent behaviours were often misleading," she wrote in her 1990 book. It took decades of experiments for it to become clear that "many intellectual abilities that had been thought unique to humans were actually present, though in a less highly developed form, in other, non-human beings."

AI systems wield language fluently. They can solve complex mathematical

problems. They create music and illustrations. All this was, until recently, thought to be the domain of humans and humans alone. “We have this presumption of human exceptionalism: this idea that we are distinctive and significant, that we have complex and sophisticated capabilities which need to be protected and preserved,” says Sebo. “And this is all correct.” But, he argues, it can be “both / and”: we can see our own behavior as both impressive and mechanical, and we can see the behavior of other entities in the same way, without losing sight of important distinctions between humans and machines.

Sebo wonders if one reason for the implicit skepticism in the Pope’s encyclical could be that it’s “doing important work by safeguarding human dignity, through denying AI dignity.” If we stake our worth on exclusively possessing properties like agency and intelligence, we may be in trouble. But we don’t have to—“we can recognize these forms of value in others while still protecting them for ourselves,” he says.

We’re creating AI systems faster than we can understand them. Historically, parochialism about other minds has been a bad bet—reflexive dismissal won’t get us any further than credulous acceptance. Taking AI’s internal structures seriously, we stand to learn more not just about the machines we’re pulling into the world, but about our minds too.